1-1-1976

Methods for validating criterion-referenced test items.

Richard. Rovinelli
University of Massachusetts Amherst

METHODS FOR VALIDATING CRITERION-REFERENCED TEST ITEMS

A Dissertation Presented

By

RICHARD J. ROVINELLI

Submitted to the Graduate School of the
University of Massachusetts in partial
fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 1976

Psychology

METHODS FOR VALIDATING CRITERION-REFERENCED TEST ITEMS

A Dissertation

by

RICHARD J. ROVINELLI

Approved as to style and content by:

Dr. Ronald K. Hambleton (Chairman of Committee)

Dr. James M. Royer (Member)

Dr. Harry S. Schumer (Member)

Dr. Hariharan Swaminathan (Member)

Jerome L. Myers, Chairman
Psychology

February 1976

ACKNOWLEDGEMENTS

A doctoral candidate's dissertation represents the cumulative impact of not only the knowledge acquired but also the social and professional interactions experienced during his/her program. Therefore, when acknowledging individuals who were helpful to one's completion of this document, one must not look simply at the time spent on the actual preparation of the document but also at the time spent in developing the skills required for the completion of the dissertation. With this thought in mind I would like to thank my committee members, Drs. Hambleton, Swaminathan, Royer and Schumer for their contributions to the dissertation and also for the time unhesitantly spent with me in my graduate program at the University of Massachusetts.

I have reserved special thanks and acknowledgement to the chairman of my committee, Dr. Ronald K. Hambleton. His continual support and contributions to my personal as well as professional growth was greatly appreciated. Also I would like to acknowledge his contribution to this dissertation through his help in defining the problem and closely supervising the work from beginning to end. Finally, I am most grateful for the many hours spent with me in improving my writing skills and in prodding me to

produce a worthwhile piece of research.

To Dr. Swaminathan, I would like to add an additional thanks for the many hours spent helping me understand statistical methods.

To my wife, Ann, and my children, Cheryl and Becky, I give my thanks for their patience and perseverence in the seven years it has taken for me to complete my graduate studies.

While it is impossible to acknowledge all the individuals who in some manner contributed to my graduate program to this dissertation I would like to extend my gratitude to Larry Cardorette and Daniel Sheehan for allowing me to analyze their data in the study.

Finally, I should like to thank Bernie McDonald for the care and attention given to typing the many drafts of the dissertation.

Methods For Validating Criterion-Referenced Test Items

(February 1976)

Richard J. Rovinelli, B.A., Tufts University

M.Ed., University of Massachusetts

Directed by: Ronald K. Hambleton

ABSTRACT

In recent years there has been a significant movement towards the individualization of instruction in education. This movement which encompasses a number of instructional models has as a major purpose the enhancement of the learning of all students. Given this purpose it is reasonable to expect that any measurement instrument used to assess student performance will provide information which can be employed to make educational decisions on an individual basis. Further these measurement instruments should provide information that can be used to measure progress along an absolute ability continuum.

Traditional measurement devices, norm-referenced tests, are constructed specifically to facilitate making comparisons among students; hence, they are not very well suited for making most of the instructional decisions required in individualized instructional programs. The

inadequacies of norm-referenced tests for providing information related to performance on specific instructional objectives which could then be used to make educational decisions has led to the development of a more appropriate form of testing known as criterion-referenced testing. A criterion-referenced test has been defined as a test which is deliberately constructed so as to yield measurements that are directly interpretable in terms of specified performance standards.

To date, considerable research on criterion-referenced testing has been conducted to solve important problems such as those concerning the estimation of mastery scores, the estimation of reliability and validity, and the setting of cutting scores. However, one area researchers have given relatively little attention is the problem of item validation, *i.e.*, the problem concerning the extent to which items are measures of the objectives they have been designed to measure.

The problem of item validation is of particular importance with criterion-referenced tests because of the way test score information is used. Specifically, if an examinee responds correctly to several items measuring a particular objective in a criterion-referenced test it is inferred that he/she has mastered that instructional objective. In spite of the basic importance of the item validity problem, to date, there does not exist a methodology for conducting item validation studies. Clearly

then, the problem of item validity is an important one to study and consequently this particular investigation was designed. Specifically, the study was designed to achieve three goals:

1. To provide a synthesis and organization of the various item validity methods that have appeared in the literature,
2. To conduct an empirical investigation of available item validity methods to determine which ones provide the most useful information and in which situations,
3. To produce an item validation methodology that reflects the results of work in the first two areas above.

On the basis of our background research, it was decided to organize existing item validity methods around three rather different approaches: the use of item generation rules; the use of empirical analyses; and the use of the judgments of content specialists.

With regard to the use of item generation rules, the research in the study consisted of an attempt to organize the literature and discuss strengths and weaknesses of the approach. Both an organization and discussion of the available literature, as well as empirical studies, were carried out with the other two approaches.

The results of empirical studies on two sets of test data clearly suggested that empirical analyses in

and of themselves could not be used to establish item validity. Further, while the use of the judgments of content specialists appeared to be an effective means for assessing item validity, this approach was insensitive to item statistical deficiencies. Given the empirical findings of the study, it was determined that item validity must be assessed on both a content level and an empirical level.

Two models for the construction of criterion-referenced tests were presented in the study. The basis for these models was the review of the existing research and the empirical studies carried out. For model one it is assumed that extensive resources are available for developing and validating test items. For model two, it is assumed that the test constructor is limited in both the time and resources he/she has to devote to test development. The major objective of both test construction models is to increase the reliability and validity of criterion-referenced test item data through the identification, reduction, and elimination of sources of error relating to criterion-referenced test items.

In conclusion, it is noted that in this study, procedures have been set forth to ascertain whether or not an item is a measure of objective it is designed to measure (item validity). These procedures represent an important theoretical contribution to the field and provide practical means for enabling practitioners to construct valid criterion-referenced test items.

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

TABLE

CHAPTER I

INTRODUCTION

1.1 Background

While the idea of adapting instructional programs to meet the individual needs of all students is not a new theme in education (see, for example, Washburne and Marland, 1963; and Wilhelms, 1962), it has only been in the past decade that individualized instructional programs have been implemented on a large-scale basis in the schools. This trend toward individualization of instruction in education has resulted in the development of a number of different instructional models. Included among them are the *Individually Prescribed Instructional Program* (Glaser, 1968), *Project PLAN* (Flanagan, 1967, 1969), and a *Model of School Learning* (Carroll, 1963, 1970).

An individualized instructional model includes many components: instructional models that include a specification of curricula in terms of behavioral objectives, a detailed diagnosis of the entering competencies of students, individual pacing and sequencing of material, as well as the careful monitoring of student progress, will be of special interest in this study.

Since the major purpose of individualized programs is to enhance the learning of all students, it is reasonable to expect that any measurement instrument used to assess student performance will provide information which can be employed to make educational decisions on an individual basis. Since one of the major purposes of individualized programs is to maximize the opportunity for all students to learn, it follows that tests used to monitor student progress should also be keyed to the instruction. Further, these tests should provide information that can be used to measure progress along an absolute ability continuum. Norm-referenced tests are constructed specifically to facilitate making comparisons among students; hence, they are not very well suited for making most of the instructional decisions required in individualized instructional programs. Hambleton and Novick (1973) comment on the primary purpose of testing in individualized instructional programs. They note:

> It would seem that in most cases, the pertinent question is whether or not the individual has attained some prescribed degree of competence on an instructional performance task. Questions of precise achievement levels and comparisons among individuals on these levels seem to be largely irrelevant. In many of the new instructional models, tests are used to determine on which instructional objectives an examinee has met the acceptable performance level standard set by the model designer. This test information is usually used immediately to evaluate the student's mastery of the instructional objectives covered in the test, so as to appropriately locate him for his next instruction.

The inadequacies of norm-referenced tests for providing information related to performance on specific instructional objectives which could then be used to make educational decisions and to evaluate the effectiveness of the instructional programs has led to the development of a more appropriate form of testing known as *criterion-referenced testing*. Criterion-referenced tests are specifically designed to provide the data necessary to make instructional decisions and to evaluate program effectiveness.

While a number of definitions of criterion-referenced tests have been offered (Glaser and Nitko, 1971; Harris and Stewart, 1971; Ivens, 1970; Kriewall, 1969; and Livingston, 1972), Hambleton and Novick (1973) note that the definition of Glaser and Nitko (1971) is the most flexible. Glaser and Nitko define a criterion-referenced test as "a test which is deliberately constructed so as to yield measurements that are directly interpretable in terms of specified performance standards."

The amount of interest and energy that has been expanded in the area of criterion-referenced testing and measurement in the last few years has been impressive. The theoretical and practical problems that have received considerable attention from educational measurement specialists include: the estimation of mastery scores, the estimation of reliability and of validity, the setting of cutting scores and the development of systems for reporting

criterion-referenced testing information (see, for example, the work of Fremer, 1972; Hambleton and Novick, 1973; Kriewall, 1972; Livingston, 1972; Millman, 1974; and Popham and Husek, 1969). Considering its importance, researchers have given relatively little attention to the problem of item validation, *i.e.*, the problem concerning the extent to which items are measures of the objectives they are intended to measure. Important contributions to date include the work of Wells Hively and his colleagues (1968, 1973) and Millman (1974).

1.2 Importance of the Item Validity Question

The problem of item validation is of particular importance with criterion-referenced tests because of the way one uses the test score information. Specifically, if an examinee responds correctly to several items measuring a particular objective in a criterion-referenced test, it is inferred that he/she has mastered that instructional objective. In fact, the success of an individualized program depends to a considerable extent on how effectively teachers make decisions concerning the student mastery of specific instructional objectives. Thus, unless one can say with a high degree of confidence that the items in a criterion-referenced test measure the instructional objectives, any use of the test information for instructional decision-making is questionable.

The question of item validity is also central to two other important areas of instructional development. First, the introduction of objective and test item banks has been designed to reduce the work load normally associated with developing objectives-based curricula (Wood and Skurnik, 1969). The purpose of these banks is to make available to teachers and researchers a large pool of instructional objectives and test items measuring the objectives. From this pool, various combinations of items which are appropriate for the instructional units being taught or researched can be selected. These objective and test item banks will often contain thousands of items and objectives. Therefore, it is obvious that without some systematic approach to the problem of determining item-objective relationships quickly and effectively, it is unlikely that objective and test item banks can achieve their intended purpose.

Second, of course, it is also essential to determine the effectiveness of various components of an objectives-based program. This, in part, can be done with paper and pencil tests that assess the various curriculum objectives. Once again, it is essential to know that the test items measure the intended objectives. (As a side note, curriculum evaluators are presently using item-examinee sampling designs to collect more information on a program than would be available if all examinees in the population of interest took only a subset of the total set of test items.

Item-examinee sampling is particularly applicable for curriculum evaluation where information about the group is usually of more interest than information on individual examinees [Lord, 1962; Lord and Novick, 1968].)

1.3 Purposes of the Investigation

From the discussion in the first two sections, it is apparent that the problem of item validation is an important one to resolve for the effective development and implementation of objectives-based curricula. For example, as was mentioned earlier, unless one can be sure that a set of items measures a particular objective in a criterion-referenced test, it is impossible to effectively monitor student progress through an objectives-based instructional program or to effectively evaluate instruction.

In spite of the basic importance of the item validity problem to the criterion-referenced testing area, it is surprising to note that, to date, there does not exist a methodology for conducting item validation studies. What does exist in the literature is a smattering of techniques that address different aspects of the item validity problem. As recently as 1974, Popham posed two important questions for criterion-referenced test developers:

1. What techniques can be devised which will permit objective-based test developers to improve their instruments on the basis of empirical tryouts in the same ways that conventional test developers have been doing for years (*e.g.*, total test reliability, item reliability, item homogeneity, objective-item congruence)?

2. Are there technical rules which can be produced to aid reviewers in judging the congruence between test items and the objectives on which they are based?

Further, Skager (1974) noted that there are still a number of unresolved questions relating to criterion-referenced test item validity. For example:

1. How does one establish the fact that items in the pool measuring any objective are valid in the sense of being (a) congruent with the objective, *e.g.*, actually measuring the performance described in the objective and (b) comprehensive in the sense of providing adequate coverage of the domain specified by the objective?

2. How does one identify poorly written items by means of item analysis procedures when the frequency of correct response may be extremely high or low?

Clearly then, the problem of item validity is an important one to study and consequently, this particular investigation was designed. Specifically, the study was designed to achieve three goals:

1. To provide a synthesis and organization of the various item validity methods that have appeared in the literature.
2. To conduct an empirical investigation of selected item validity methods to determine which ones provide the most useful information and in which situations.
3. To produce an item validation methodology that reflects the results of the work in the first two areas above.

Relative to the third goal, the methodology is a set of guidelines or rules which take into consideration the resources available to the test developer, the data which has been previously collected on the items which will comprise the test and the purposes for which the test is being developed.

Our background research suggested to us that it would be useful to organize existing item validation methods around three rather different approaches: the first involves the use of item generation rules. The major categories of item generation procedures given by Millman (1974) are:

1. Linguistic-based Schemes: this technique derives items by using operational definitions which set forth the rules for transforming instructional material into items. This technique has been advanced by Bormuth (1970) and Anderson (1972).
2. Item Forms: this technique provides a strict framework for item generation and is reviewed extensively later in the investigation. The major developers of this procedure have been Osburn (1968) and Hively, *et al.* (1968, 1973).
3. Amplified Objectives: an amplified objective is a statement of an educational or instructional goal which contains specific information about three areas: the testing situation; response alternatives; and the criteria of correctness.

A second approach to item validity involves the use of content specialists. To date, little use has been made of this approach. Because of the intuitive appeal of the approach and the limited amount of study given the area to date, substantial amount of effort was devoted to this approach in the study.

The third approach to item validity involves the use of empirical data. Empirical analyses were conducted on two sets of data collected on a pool of items constructed to measure objectives of an individualized high school science program. The first set of data contained the results of a single test administration to a group of students. The second set of data was collected from the administration of three parallel forms of a criterion-referenced test measuring twelve objectives that was given to three groups of students. Each group was tested on all three forms of the test during either the pre-test, post-test or delayed post-test situation. For other analyses in the study, simulated data was used.

1.4 Organization of the Study

The remainder of the study is organized around four chapters. Chapter II provides a review of important background information concerning objective-based instructional programs and criterion-referenced testing and measurement. This information, while not essential to an understanding of later material in the investigation, does provide a context for the item validation problem.

Chapter III was designed to provide a review of the literature concerning various procedures relating to the item validation problem.

In Chapter IV, we have provided a comprehensive comparison of a large number of the item validity techniques using two sets of criterion-referenced test data and some computer-simulated data. The fifth and final chapter provides a summary of the results of the study, a proposed item validation methodology, and several suggestions for further research.

CHAPTER II

OBJECTIVES-BASED INSTRUCTION, TESTING, AND MEASUREMENT

2.1 Introduction

The intention in this chapter is to provide some background for the study of the problem of item validity. Specifically, the remainder of the chapter is divided into three sections: Brief Description of Objectives-Based Programs, Norm-Referenced Testing and Measurements, and Criterion-Referenced Testing and Measurements.

2.2 Brief Description of Objectives-Based Programs

An objectives-based program which includes nearly all of the basic features of individualized instructional programs was drawn up by Glaser (1970) and Glaser and Nitko (1971). The six basic components of their model are:

1. The goals of learning are stated in terms of observable student behavior.
2. When the student begins a particular instructional program, his/her initial capabilities--those relevant to the forthcoming instruction--are assessed.
3. Educational resources matched to the student's initial capabilities are presented. The student

selects or is assigned one of these alternatives.

4. The student's performance is monitored and continuously assessed.
5. Instruction proceeds as a function of the relationship between measures of student performance, available instructional resources, and criteria of competence.
6. As instruction proceeds, data are generated for monitoring and improving the instructional system.

In summary, the goals of objectives-based programs developed along the lines of the general model are, among other things, to enable students to work through the units of instruction at a pace reasonable for them, to develop self-direction and self-initiation, to encourage self-evaluation as well as motivation for learning, and to demonstrate mastery in a variety of skills.

2.3 Norm-referenced Testing and Measurements

The procedures commonly used for constructing norm-referenced tests have been extensively discussed and clearly defined. One of the most comprehensive statements on the topic has been presented by Tinkelman (1971). The following five points represent the essence of the process of test construction for norm-referenced tests set forth by Tinkelman:

1. The purpose and requirements of the tests have to be established.
2. A blueprint must be developed to establish the scope of the test--the content domain to be tested.
3. Items which measure all aspects of the content domain must be written.
4. The items are pre-tested and evaluated to determine not only their appropriateness with regards to the content domain but also to their effectiveness in meeting the stated purposes and requirements of the test.
5. The test is evaluated to determine its consistency (reliability) and its accuracy (validity) as a measurement tool.

The evaluation of a norm-referenced test begins with an examination of the items of the test. Since norm-referenced tests are designed primarily to provide information which can be used to differentiate amongst examinees, the individual items are important only in the context of their contribution to the total test score variance. Consequently as Lord and Novick (1968) have indicated, item analysis procedures used in the evaluation of items for norm-referenced tests should employ item parameters that have "a definite relationship to total test parameters such as test score, reliability and validity, in order to ensure that the test has optimal measurement properties." Three item parameters which satisfy this requirement are the

item difficulty index, the item discrimination index and item validity.

Item difficulty is "the expected relative score on an item by a population of examinees." It is calculated as the proportion of examinees responding correctly to an item. In order to maximize test score variance it is necessary to choose individual items which make maximum contributions to it. This is accomplished by selecting items with difficulty indices around .50. The .50 level can be shown to maximize item variance. A simplified interpretation of this point is that if most of the examinees respond either correctly or incorrectly to an item, item variance is minimal and accordingly the item will provide little information which could be used to rank order the examinees by ability on the trait being measured.

Item discrimination indices provide information concerning the extent to which items can be used to distinguish between high and low ability students. The biserial correlation or the point biserial correlation is frequently used to determine the relationship between item score and total test score.

In selecting items for norm-referenced tests, test constructors do not choose items with either low or negative discrimination indices. A low discrimination index means that there is little or no relationship between the item and the total test score. Thus, the item will provide little information which can be used to differentiate

between high and low ability students. A negative index is an indication that the item is being answered correctly by the low ability students and incorrectly by the high ability students. Clearly, negatively discriminating test items will contribute nothing to the overall quality of a test.

Item validity (as it is defined in classical test theory and not to be confused with the meaning of item validity in criterion-referenced testing problems) is the relationship of each item to some externally defined criterion. Lord and Novick (1968) state that a consistent estimate of this coefficient is the sample correlation between the item scores and the criterion variable.

The three parameters: item difficulty, item discrimination and item validity, are obviously important in the construction of norm-referenced tests. For example, once the test constructor knows the item difficulty and discrimination indices and the item validities, the test constructor can estimate the mean score and total test variance without actually administering the test to the group of interest. Further by choosing items which are highly correlated with an external criterion it is possible to maximize the test's validity. However, as previously stated, the total test constructs of reliability and validity are ultimately used to evaluate the test.

The construct, test reliability, refers to the consistency of measurements obtained from the test. Since

there are several different ways of viewing consistency there exist several ways of assessing reliability.

The procedures which are usually employed to obtain estimates of the coefficient of reliability are the test-retest method, internal consistency analysis and the parallel forms method. The test-retest estimate of reliability is obtained by correlating the scores of examinees on the same test administered on two different occasions. Cronbach (1970) has labelled the test-retest coefficient, the coefficient of stability, as it provides a measure of the stability of test scores for a group of examinees across some period of time. Generally, the test-retest method will give estimates of reliability that are too high because the examinees retain information from the first testing (Kuder and Richardson, 1937). A parallel-form estimate of reliability is obtained by correlating scores of examinees on parallel-forms of a test administered on the same test occasion. All of these estimates of reliability are correlational procedures and are therefore dependent on test score variance. The internal consistency estimates of reliability such as the Kuder-Richardson formulas, Cronbach's coefficient alpha and Hoyt's (1941) analysis of variance coefficient provide a measure of the extent to which the items in the test measure a single ability.

Test validity refers to the extent to which the test scores serve their intended use (Cronbach, 1971). One estimate of the coefficient of validity can be obtained by

computing the correlation between the test and a specified criterion. From the formula for correlating scores on a predicator test with a criterion test (Lord and Novick, 1968, p. 332) one can see that in order to maximize a test's validity, the test constructor must select items which are highly correlated with the criterion but are not highly correlated with each other. Other kinds of validity and methods of estimation are presented by Cronbach (1971) and Lord and Novick (1968).

To summarize, in constructing a norm-referenced test, one relies heavily on item parameters in the process of selecting items, and ultimately the credibility of information received from norm-referenced tests depends upon the validity coefficients being above some minimally acceptable levels.

2.4 Criterion-referenced Testing and Measurements

In a paper by Hambleton (1974), he notes that

> Nearly all of the testing and measurement that takes place in the context of individualized programs is criterion-referenced. Unfortunately, this particular branch of testing is not well understood by most practitioners in the field. The standard procedures for constructing, administering, analyzing tests, and interpreting scores, in the context of standard instructional models and methods are certainly well-known to educators. With these standard procedures, tests have been used primarily and most successfully to estimate each examinee's ability level and to permit comparative statements (*e.g.*, ranking) across examinees. However, with [objectives-based programs] the well-known classical mental test models for test construction and test score interpretation appear to be less useful.

Hambleton and Novick (1973) go on to say that

> The primary problem in the new instructional models, such as individually prescribed instruction, is one of determining if π_i, the student's true mastery level, is greater than a specified standard π_o. Here, π_i is the "true" score for an individual i in some particular well-specified content domain. It *may* represent the proportion of items in the domain he could answer successfully. Since we cannot administer all items in the domain, we sample some small number to obtain an estimate of π_i, represented as $\hat{\pi}_i$. The value of π_o is the somewhat arbitrary threshold score used to divide individuals into the two categories described earlier, *i.e.*, Masters and Nonmasters.

The construction of criterion-referenced tests can be seen as consisting of four steps: the establishment of a domain of behaviors which are to be measured; the development of a procedure for generating items to measure these behaviors; the development of an item sampling strategy; and the evaluation of the test.

Step one is simply the identification and categorization of all the behaviors which are to be tested (*i.e.*, a behavioral objective bank). This step is similar to the process of developing a blueprint for norm-referenced tests. The second step, the generation of items, can be accomplished in two ways. The first way, and to date the most commonly used, is the empirical approach. In this approach, items are written without regard to prescribed rules or formulas and are then refined on the basis of information received from a pre-testing. The second approach to the problem of item generation employs item formats (for example, see Hively, *et al.*, 1968). The item format approach relies on carefully defined rules and specifications to generate

items operationally from the domain of behaviors being tested. In section 3.2 we will present an extended review of the literature concerning item generation rules. Next is step three. Once the items have been written an item sampling procedure must be carefully formulated to ensure that items which are representative of all the categories of behaviors are included in the test.

Finally, the test has to be evaluated. Among the things we need to know is the extent to which the items in the test measure the appropriate objectives. To date, there is no well-accepted methodology to assist in answering the question. On the contrary, as we will see later, many of the suggested procedures for item validation are quite misleading and in some cases, incorrect.

On this last point it is important to note initially that a criterion-referenced test may be multidimensional. Hambleton and Novick (1973) note that:

> It is apparent that the [criterion-referenced] test may be multi-dimensional while made up of unidimenisional subscales. That is, the items from a criterion-referenced test are organized in distinct and different subscales of homogeneous items measuring common skills. . . . Major interest may rest on the reliability and validity of subscale scores.

Thus, when writers discuss internal consistency, reliability and validity issues they are referring to subtest scores unless the test itself is a measure of only one objective.

The use of the procedures of classical test theory to evaluate criterion-referenced tests has been examined by a

number of researchers (Popham and Husek, 1969; Livingston, 1972; Ozenne, 1971). Popham and Husek (1969) while noting that criterion-referenced tests should be internally consistent, state that the use of classical procedures for obtaining estimates of the reliability with their dependence on test score variance are probably inappropriate as criterion-referenced test scores are likely to be quite homogeneous. They further add that a straightforward application of these procedures for criterion-referenced tests may even result in negative estimates for the internal consistency of a criterion-referenced test (for example, se Sirotnik, 1970). With regards to test validation, Popham and Husek (1969) opt for content validity approaches over correlational procedures for the same reasons.

Livingston (1972) pointed out that norm-referenced definitions of reliability have the desirable properties of being interpretable as the proportion of variance which is caused by variation in the examinees' true scores rather than errors of measurement; of being estimable from either a single form or two parallel forms; of being related to test length; and for being useful in the correction of correlation coefficients attenuated by errors of measurement. Accordingly, he developed a means of applying classical test theory to criterion-referenced tests. Basically, he has just conceptualized the purpose of criterion-referenced tests as one of determining how far the examinees deviate from the specified criterion rather

than the mean of the test. Thus, in his approach, all deviations are taken about the cutoff score rather than the test score mean as is done in norm-referenced test theory. With this simple modification to the classical test theory model, Livingston (1972) is able to apply all the theorems and assumptions of classical test theory to criterion-referenced tests.

While Livingston's work is appealing because of its relation to classical test theory, it would appear that it has limited usefulness because of ways criterion-referenced tests are being used in practice (see for example, Hambleton, 1974). The information required by teachers in a criterion-referenced testing situation is not primarily how far above or below the passing line the examinee is but whether or not the student has achieved below, at, or above the criterion. Consequently, we do not envision any wide scale use of Livingston's procedures to evaluate criterion-referenced tests as the theoretical basis for his work is not in line with the purposes for much of criterion-referenced testing.

Ozenne (1971) has presented a methodology for evaluating criterion-referenced tests. He states that in a criterion-referenced testing situation, the interesting question is "How effective has the instruction been?". Thus, for a criterion-referenced test to be of any value it must be sensitive to instructional effects. Hence, he develops a response model to take into account the variability due to instruction and the interaction of examinees and levels

of occasions as well as individual differences. (There are two levels of the occasion factor, pre-test, and post-test.) The model is given as

$$Y_{jk} = \pi + \alpha_j + \beta_k + (\alpha\beta)_{jk} + e_{jk} ,$$

where Y_{jk} is the score of individual j on test occasion k,

π is the population parameter,

α_j is the effect due to individual differences,

β_k is the effect due to instruction,

$(\alpha\beta)_{jk}$ is the effect of the interaction of examinee and occasion factors; and

e_{jk} is the error of measurement.

The number of occasions is obviously a fixed factor since there are only two occasions of interest and it is assumed the examinees are sampled randomly from a population (a random factor). The various components of variance can be obtained from an analysis of variance table. The analysis of variance for the above design is shown in Table 2.4.1.

Ozenne then defines the sensitivity of a test (S) to be the ratio of the variance due to the instructional effects (the occasions effect), to the sum of variances due to instructional effects and errors of measurement.

The work of Livingston (1972) on reliability estimation and Ozenne (1971) on developing a measure of the

TABLE 2.4.1

ANALYSIS OF VARIANCE COMPONENTS FOR THE

EXAMINEE X OCCASION DESIGN

Source	df	EMS
A (occasions)	a-1	$\sigma_e^2 + bn\sigma_A^2 + n\sigma_a^2{}_b$
B (examinees)	b-1	$\sigma_e^2 + an\sigma_a^2{}_b$
AB (interaction)	(a-1 (b-1)	$\sigma_e^2 + n\sigma_a^2{}_b$
Error (within AB)	ab (n-1)	σ_e^2

sensitivity of a criterion-referenced test to detect the effect of instruction represent one of two approaches to the problem of developing a methodology for evaluating criterion-referenced tests. We are of the opinion that other approaches to this problem are perhaps more appropriate. We say this for two reasons. First, they conceptualize the problem in the norm-referenced sense of evaluating the total effectiveness of a test summing across items when in fact the crucial question for criterion-referenced tests is whether or not the individual items are measures of the behaviors being tested. Second, the procedures they use, correlation and analysis of variance, are based on a loss function not appropriate for the measurement problem associated with criterion-referenced tests (Hambleton and Novick, 1973). We will now expand on this second point because it is an important one.

The most commonly used loss function in the context of testing is the squared-error loss function. Here it is assumed that the loss associated with over or under estimation is equivalent to the square of the error of prediction (Novick and Jackson, 1973):

$$\text{loss } (e) = e^2 = (\hat{y} - y)^2.$$

Novick and Jackson (1973) state that with the squared error loss function one should choose a prediction function which minimizes the average value of the squared errors of prediction. Further, they point out that standard

regression and correlational theories are "derived from this loss function." With the squared error loss function, one tries to minimize the average loss and accordingly choose y to be equal to the group mean. The average loss is thus the group variance.

The threshold loss function that has been recommended for use with criterion-referenced tests is based on the assumption that the important information is whether or not an individual is above a specified minimal competency level or cutting score. We quote from Hambleton and Novick (1973):

> Criterion-referenced measurement involves what Cronbach and Gleser (1965) would call a "quota-free" selection problem. That is, there is no quota on the number of individuals who can exceed the *cut-off scores* or *threshold* on a criterion-referenced test. A cut-off score is set for each subscale of a criterion-referenced test to separate examinees into two mutually exclusive groups. One group is made up of examinees with high enough test scores ($\geq$ the cut-off score) to infer they have mastered the material to a desired level of proficiency. . . .At this stage of the development of a theory of criterion-referenced measurement, the establishment of proficiency levels is primarily a value judgment. . . .
>
> . . .suppose we take some criterion level π_o, and define a parameter ω such that
>
> $$\omega = 1 \text{ if } \pi \geq \pi_o$$
>
> $$\omega = 0 \text{ if } \pi < \pi_o$$
>
> Persons having ω values of one are those who have true ability levels equal to or greater than the criterion level π_o, and those having ω values of zero are those whose π_o values are below π_o. Now if we obtain an estimate of π_i, then an

estimate of ω could be obtained in the following way:

$$\hat{\omega} = 1, \text{if } \hat{\pi}_i \geq \pi_o \text{ and}$$

$$\hat{\omega} = 0, \text{if } \hat{\pi}_i < \pi_o.$$

Defining our error of estimation as $(\hat{\omega} - \omega)$, the difference between the estimated and the true value, it is clear that the error takes on one of three values, +1, -1, 0, corresponding to whether we make a false-positive error, a false-negative error, or a correct classification.

Then, from the above problem we could define the following function:

$$\ell(e) = 0, \text{ if } e = 0$$

$$= a > 0 \text{ if } e = +1$$

$$= b > 0 \text{ if } e = -1.$$

That is, if an individual is declared to be above the criterion value and he/she is not, then the loss is a; if an individual is declared to be below the criterion value, then the loss is b; and, of course, if a correct statement about the individual's performance is made, the loss is zero. It is then possible to develop a method for estimating mastery to minimize the threshold loss function (Hambleton and Novick, 1973; Hambleton, Swaminathan, Algina, 1976; Millman, 1974; Swaminathan, Hambleton, Algina, 1975).

Moving on to the reliability problem, Hambleton and Novick (1973) have presented an example of the use of the decision-theoretic approach to estimate test reliability. Given two tests which are parallel in the classical sense, criterion-referenced test reliability could be defined as the proportion of times that the same classification

decision is made for examinees on the two tests. Details of this approach have been expanded on by Swaminathan, Hambleton, and Algina (1974).

Another approach for the estimation of reliability for criterion-referenced tests requires that the test items be ordered in a manner such that when an examinee responds incorrectly, the examinee will not be able to respond correctly to any items which follow. An obvious requirement for such a test would be that the objectives being taught are also capable of being ordered in some hierarchical fashion. Jackson (1970) has noted that scalable tests can only be obtained in a limited number of situations. Cox and Graham (1966) presented the construct of reproducibility as a means of assessing the reliability of such tests. The coefficient of reproducibility represents the extent to which a test satisfies the condition of scalability in the Guttman sense.

Using the decision-theoretic approach one could define the validity of a criterion-referenced test in the same way as the reliability except, of course, that a new test serves as the criterion.

It is important to note that while the conceptualization of what the constructs, reliability and validity, represent is as crucial for criterion-referenced tests as norm-referenced tests, researchers examining the problem of assessing reliability and validity for criterion-

referenced tests must do one of two additional things:

1. Show that the items are indeed measures of the behaviors being measured.
2. Assume that the items are measures of the behaviors being measured.

To date, no procedure for evaluating whether or not an item is a measure of an objective has been presented which is acceptable to most of the researchers interested in the problem. The present study was designed to provide a more satisfactory solution to the problem of item validity than that offered in any work to date.

CHAPTER III

APPROACHES TO ITEM VALIDITY

3.1 Introduction

In this chapter, two different approaches to the problem of establishing item validity will be discussed. In the first approach, an attempt is made to establish item validity by using specific item writing guidelines. That is, a direct relationship between the items and their matched objectives is established by requiring test constructors to conform to rigorous guidelines when writing items. In the second approach, an attempt is made to ascertain whether or not a set of items has item validity through the use of information collected after the items have been constructed. Under the second approach, two types of information are possible: data from content specialists and data collected from a pilot testing of the items.

3.2 Item Writing Procedures

While a number of formal item writing procedures have been discussed in the literature, in this section only two of the more prominent ones, item forms analysis and linguistic based schemes, will be discussed. In

addition, the recent contribution to item development, the use of amplified objectives (Popham, 1974), will be discussed in some detail.

In developing a set of test items, the test constructor is rarely interested only in the specific information obtained from the student's performance on the test being constructed. Usually, the purpose of collecting information on the student's performance is to obtain estimates of the student's level of competency on the items comprising the whole content domain defined by the objectives covered by the test.

The approach to test construction, which views a single test as being composed of items sampled from "a large, well defined domain of items" has been set forth in the Cronbach, *et al.* (1963) Generalizability Theory. In this approach, "parallel test forms are obtained by repeated sampling according to a given plan, and analysis of variance techniques are used to obtain estimates of components of variance due to sampling error and other facets which may effect the reliability of the score obtained from a particular test, under particular field conditions" (Hively, Patterson, and Page, 1968). The critical assumption underlying generalizability theory, that the items are randomly sampled from a universe of items from the whole content domain, has been used as the structural basis for item generation procedures.

3.2.1 Item Forms Analysis

Osburn (1968) noting that while "few measurement specialists would quarrel with the premise that the fundamental objective of achievement testing is generalization," in fact "current procedures for the construction of achievement tests do not provide an unambiguous basis for generalization is due to the fact that current test construction practices do not state in operational terms "the method of generating items and criteria for inclusion of the items in the test" (Osburn, 1968). Norm-referenced test construction procedures obfuscate this weakness through the use of the concept of a latent variable--"an underlying continuum which represents a hypothetical dimension of knowledge or skill" (Osburn, 1968). That is, norm-referenced tests have only an illusion of generalization which is created by having a group of test items labelled as the latent variable and assigned a content specific name such as mathematics (Osburn, 1968). Osburn goes on to add:

> Statistical analysis of test data is, of course, very useful. But no amount of item analysis or factor analysis can provide a firm basis for generalization to a universe of content. The basis of generalization must be contained in the operational definition of the procedures used in generating a sampling of items that go to make up the test.

Osburn's (1968) approach to the problem of ensuring generalization is through the concept of a *universe-defined* test. The universe-defined test is "a test constructed and administered in a way that an examinee's score on the test

provides an unbiased estimate of his score on some explicitly defined universe of item content" (Osburn, 1968). Osburn's (1968) universe-defined test has two requirements which are given as follows:

1. All the items which could be written from the content domain to be tested must be written in advance of the final item selection process.
2. A random or stratified random sampling procedure must be used in the item selection process.

One way to achieve these two requirements is through the process of item forms analysis (Hively, 1962). An item form is actually a process having the following characteristics:

1. It generates items with a fixed syntactical structure.
2. It contains one or more variable elements.
3. It defines a class of item sentences by specifying the replacement sets for the variable elements.

Essentially, the development of a universe-defined test then consists of decomposing the content domain into a hierarchical arrangement of item forms and then generating the test items according to a specified sampling procedure. One of the obvious advantages of such a system is that the workload which would be required in writing the large number of items needed to satisfy the conditions for generalizability would be reduced. Further, the question of test validity is automatically answered in a

universe-defined test as what the test is measuring is implicitly established by the item generation rules.

Hively, *et al.* (1968), present what they defined as the first nontrivial application of Osburn's universe-defined testing procedure in developing a test to analyze performance in a mathematics program. In trying to use classes of item forms as a means for diagnosing learning difficulties and/or predicting responses, the authors met with only moderate success. However, their study is important in that it demonstrated that it was possible to develop and use item generation rules to construct a test. Their study also underscored one of the major weaknesses of item generation procedures--the procedures are more easily employed with highly structured subject matter areas such as mathematics.

While Hively, *et al* (1968), used Osburn's theoretical formulations, they did not use a computer to generate items from the item forms. Fremer and Anastasio (1969) were able to use a computer to generate spelling items from specified item formats.

Ferguson and Hsu (1971) also used a computer-based system for generating items. While they employed item forms as a basis for their item generation process instead of using single item forms, they developed a procedure which operates on clusters of item forms which were established by uniting item forms which share a similar content. Since their work was intended as a theoretical exposition of test

construction procedures, they did not provide any data in support of their approach.

The most extensive incorporation of item forms analysis into a testing procedure, domain-referenced achievement testing, has been presented by Hively, Maxwell, Rabehl, Sension, and Lundin (1973) in their work *Domain-Referenced Curriculum Evaluation: A Technical Handbook and a Case Study from the Minnmast Project.* A comprehensive discussion of this important work will serve to delineate the use of item generation procedures to ensure item validity.

As previously mentioned, in a criterion-referenced testing situation, the fundamental purpose is to ascertain whether or not the examinee has attained specified instructional objectives. Hively, *et al.* (1973), feel that a gap exists between the instructional objectives and the tests which purport to measure them. They note that the most common approach to ensure that the items are measures of objectives has been to construct prototypical test items that are keyed to more generally stated descriptions of the desired behavior (Hively, *et al.*, 1973; Hively, 1974). An inherent deficiency in this approach is that there is no way to ensure that the test items completely defined the spectrum of desired behaviors. In the previously discussed paper of Hively, *et al.* (1968), the authors attempted to rectify this deficiency by specifying "all the behaviors which comprise specific pieces of knowledge." While this approach would appear to solve the problems of operational

definition and of generalizability it soon became apparent that it was impossible "to exhaustively define universes of criterion behavior" (Hively, *et al.*, 1973).

The next model proposed by Hively and his associates defined the sets of test items not as universes of items but as the "nuclei of hypothetical repertoires of behavior." These nuclei are called "domains" (Hively, *et al.*, 1973). The following paragraph taken intact from their paper clearly develops the idea of a domain-referenced testing model:

> The basic notion underlying domain-referenced achievement testing is that certain important classes of behavior in the repertoires of experts (or amateurs) can be exhaustively defined in terms of structured sets or *domains* of text items. Testing systems may be *referenced* to these domains in the sense that a testing system consists of rules for sampling items from a domain and administering them to an individual (or sample of individuals from a specified population) in order to obtain estimates of the probability that the individual (or group of individuals) could answer any given item from the domain at a specified moment in time.
>
> Domains of test items are structured and built up through the specification of stimulus and response properties which are thought to be important in shaping the behavior of individuals who are in the process of learning to be experts. These properties may be thought of as stratifying large domains into smaller domains or subsets.
>
> Precise definition of a domain and its subsets makes statistical estimation possible. This provides the foundation for precise diagnosis of the performance of individuals over the domain and its subsets. In addition, clear specification of the properties used to structure the domain makes possible inductive generalization beyond the domain to situations which share those properties. That is, once we have diagnosed a student with respect to a defined domain we may be able to predict his behavior

(in a non-statistical, inductive fashion) in natural situations which have some properties in common with the test items within the domain.

Hively, *et al.* (1973), go on to list the following strategies for developing domains of test items:

1. Start with a list of prototype items taken from the instructional material and then alter these items to produce sets of equivalent items measuring the objectives supposedly measured by the prototype items. Then have content experts review the items so as to end up with a pool of items which purport to measure the instructional objectives.
2. State the instructional objectives and have the item writers develop items which supposedly measure the instructional objectives.
3. Develop hypotheses about sequences and hierarchies of instruction through a careful examination of the basic goals of the instructional unit. Then construct items in accordance with these sequences and hierarchies.

A crucial aspect of domain-referenced testing is the construct of "item form" first discussed by Hively (1962) and by Osburn (1968) as an integral part of universe-defined testing. Hively, *et al.* (1973), give two reasons for the use of item forms:

1. To obviate the necessity to store individual items by substituting a set of written rules through

which items can be generated when needed, and

2. To enable the relationships among items to be traced by giving clear specification of relevant item characteristics.

Thus, an item form is composed of two parts. The first part details how to generate the items while the second part depicts the item characteristics. See Hively, *et al.* (1973) for a detailed description of an item form.

In order to clarify domain-referenced testing, we have prepared a brief overview of this testing model. The first step consists of making the properties of the behavioral outcomes explicit and then developing the educational objectives. Second, one of a number of strategies for writing test items which have been developed (see Rabehl, 1971) is employed to construct the domains of test items. Third, item transformation rules such as those developed by Bormuth (1970) are applied to generate new items by altering the item characteristics. That is, once the item forms have been developed for an instructional unit, one can construct any number of test items by applying the item transformation rules. Such a procedure ensures a direct relationship between the test items and the instructional objectives as well as satisfying the requirement of random sampling needed for generalization beyond the subset of items in the test (Bormuth, 1970).

3.2.2 Linguistic-based Schemes

Bormuth (1970) has presented another category of item generation procedures which can be classified as a linguistic based scheme (Millman, 1974). Relying heavily on the logic and terminology of linguistic theory, Bormuth has set forth an operational approach to item generation. The basic concept underlying this linguistic-scheme for generating items is the process of transformation. This process is a method by which information from the content domain is changed into test items. Bormuth (1970) defines four types of transformations which can be used to generate items: sentence verification types; wh-transformations (*who, what, when, where*); semantic substitutes, and paraphrases. The example below shows that one transformation, the semantic substitute, forms items by replacing words in the base sentence with synonyms and then using an interrogative word such as who or what to complete the question.

Base Sentence	Derived Item
The car was green.	What color was the vehicle?

Bormuth states that the objective of such a formal theory for item generation is to "define item populations for the purpose of demonstrating the logical relevance of items to instruction and of providing operationally meaningful referents for the basic concepts used in testing, the classes of test items." That is, such an approach ensures a direct relationship between items and objectives.

3.2.3 Summary of Formal Item Generation Procedures

While the use of item generation procedures appears to be an effective way to establish a direct relationship between the items and their matched objectives during the initial stages of test construction, a few notes of caution are in order. First, applications of these procedures have been limited to content areas which are sequential and hierarchical in nature such as mathematics. Second, applications which have been reported have been conducted either in experimental situations or well-funded projects. Therefore, before these test construction procedures can be advocated for large scale implementation, the extent to which they can be used for all cognitive domains, with higher order objectives, and in practical settings such as with classroom teachers, will have to be ascertained.

3.2.4 Amplified Objectives

Popham (1974), questioning the practical feasibility of the sustained use of sophisticated item generation procedures but also recognizing the importance of establishing a direct relationship between the items and their matched objectives presented an excellent alternative approach. The alternative approach has been labeled "the use of amplified objectives." An amplified objective is "an expanded objective which contains sufficient details regarding the nature of the measurement procedure to help the item

writer produce homogeneous items" (Popham, 1974). Homogeneous items in this sense are items which measure the same variable (objective) (Davis and Diamond, 1974).

An amplified objective has two major elements: a delimitation of the stimuli which can constitute the test items, and a description of the learner response options. The first element is concerned with clearly defining the information which can be used by the item writer to construct items. The amount of information provided should be a compromise between "sufficient detail for complete homogeneity of resulting items" and "economy of resource investment" (Popham, 1974). The second element describes the required response mode and format of the items. There are two basic response modes: selection amongst alternatives and construction of an answer. If the response mode is selection then the amplified objective should include information about the distractors which will be used. If the construction mode is used then information regarding the criterion which will be used to judge the response must be provided. The following is an example of an amplified objective:

> When a student is given the atomic number and weight of an element, the student will be able to select the correct Bohr atom diagram from amongst four alternative diagrams. The elements and distractors used must be metals.

In this example, the stimulus element defines the information provided the examinee and the task required. The response mode given is selection. Further, the

distractors are defined to be Bohr diagrams of metals. Test constructors using this amplified objective should be able to produce a homogeneous set of items.

As with item generation procedures, the underlying purpose for the use of amplified objectives is the production of a set of items which can be interpreted as random subsets of items from the content domain being tested. That is, theoretically, correct responses to this subset are an indication that the examinee has mastered the content domain and would be able to respond correctly to any items measuring the same objectives. Besides avoiding the problems inherent in the use of sophisticated item generation rules, the use of amplified objectives has the advantage of including information on the response mode. While there have been no data presented to date which supports the efficacy of this approach to item construction, it appears to be a reasonable compromise between the use of item generation rules and traditional item writing approaches.

3.3 The Use of Empirical Methods to Assess Item Validity

Of the three procedures for assessing item validity which are presented in this study, the use of empirical methods is the one which has to be used with the most caution. Four statements are offered in support of this position. First, these methods are based on the performance of a specific group of examinees which limits the

generalizability of the results. Second, it is difficult to determine the impact of instruction on the item statistics. Third, these methods require sophisticated statistical techniques which are beyond the scope of most practitioners. Fourth, many of these procedures require pre-test and post-test data on the same test items. Such data is not usually collected by practitioners.

However, not withstanding these problems, the use of empirical methods to assess item validity is not only now the most frequently used of the three procedures discussed in this study will also probably continue to be so. This is due to the fact that the data for at least some of the procedures are collected as a natural part of the instructional process. Whereas, the use of item generation procedures and the judgments of content specialists requires additional work. Further, there will be many times in which the only course of action available will be the use of empirical data. Therefore, it is important to understand the strengths and limitations of the empirical approach to item validation for criterion-referenced tests.

The empirical approach for item validation developed in this chapter consists of requiring that items meet three conditions. These conditions are given as follows:

1. That the items be able to provide information which allows the test administrator to differentiate between those examinees who have received instruction and those examinees who have not received instruction.

2. That the items which measure the same objective have similar statistical indices.
3. That an item not be a measure of more than one objective.

While the checking of these conditions for each test item may result in some overlapping information, each condition has some unique aspects which would seem to be necessary to incorporate into a methodology for item validation. Further elaboration of this point will be presented in the next few sections.

3.3.1 Condition One

Dahl (1971) presented a comprehensive discussion of the problem of item validation for criterion-referenced tests. Since in objectives-based (criterion-referenced) testing situations, achievement on test items is equated to mastery of specific instructional objectives, Dahl notes that the ascertainment of a relationship between an item and the objective of which it is purported to be a measure, "is the most important consideration in the construction and use of such tests." Dalh refers to this relationship between items and objectives which they measure as item-objective congruence or simply congruence. Congruence is the "correspondence between an objective and items which are written to measure performance on the objective" (Dahl, 1971).

According to Dahl (1971) there are two ways in which one can determine whether or not an item is a measure of an objective: the use of the judgments of content specialists;

and the comparison of data from groups receiving and groups not receiving instruction. The rationale for the use of comparison data from instructional and non-instructional groups is that the performance differential will provide a measure of congruence. Although the use of such data is not without problems such as having to determine the impact of instructional effects, Dahl (1971) feels that this type of data is essential in the determination of item-objective congruence. While we agree with Dahl as to the importance of instructional/noninstructional comparison data, we do not feel that such data provides sufficient information for assessing item validity.

There are a number of indices which appear to provide information which can be used to distinguish the examinees who have received instruction and those who have not. However, it will be important to choose those indices which are relevant for criterion-referenced tests. This poses a problem since there is an extensive literature on this topic and no closure has been reached as to which indices are most appropriate for use in evaluating criterion-referenced test items. Several indices appear relevant for the assessment of whether or not Condition One has been met and therefore they will be examined next.

The first index, referred to as a discrimination index, is the difference between the proportions of correct responses (or referred to as item difficulty indices)

from the pre-test and post-test results. A number of researchers have examined this index (Cox and Vargas, 1966; Hambleton and Gorth, 1971; and Rahmelow, Matthews and Jung, 1970).

Cox and Vargas (1966) in one of the early studies concerned with the analysis of criterion-referenced test items, computed two discrimination indices in order to determine whether they yielded similar evaluative information. One of the indices was the difference in the item difficulty indices for the pre-test and post-test data. The other discrimination index was the common D statistic (Englehart, 1965) on the post-test data. Their results showed that the use of the D statistic would have resulted in the elimination of items which would have been desirable for criterion-referenced tests. In a later paper, Cox (1970) noted, "The pre-test and post-test method of item analysis produced results sufficiently different from traditional methods to warrant its consideration in those cases where score variability is not the concern, such as in criterion-referenced measures."

Rahmelow, Matthews and Jung (1970) examined the effectiveness of item analysis procedures for criterion-referenced test items to ascertain which procedures could be used to evaluate student performance and to provide information which could be used to improve the instructional units. They computed item difficulty indices and point biserial correlations and conducted an analysis of the

change in correct responses from the noninstructional to the instructional situation. They concluded that the item difficulty index is "best for telling something about the mastery level of items but is not relevant when considering whether or not the instructional unit itself was useful." They advocated the noninstruction to post-instruction gain as a useful index for evaluating instruction (Rahmelow, *et al.*, 1970).

Hambleton and Gorth (1971) replicated and extended the Cox and Vargas study. In their study, the items were administered to the examinees on three separate occasions: a pre-test; immediate post-instructional test; and delayed (one month) post-instructional test. Three item statistics were computed from the data:

1. r_g -- the biserial correlation for item g on the post-test;
2. p_{gp}-- the difference between the proportion of individuals who correctly answered item g on the pre-test and post-test;
3. p_{gd} -- the difference between the proportion of individuals who correctly answered item g on the pre-test and delayed post-test. Again, the results showed that the choice of item statistics significantly affected the final choice of test items. The authors concluded that their results "emphasize the importance of choosing the appropriate item statistics

to select items for criterion-referenced tests."

Saupe (1966) presented an item selection technique derived from the measurement of change situation which also may be useful for checking condition one. He felt that items selected using this technique "would prove optimal in the measurement of change involving two administrations of the same test."

It is important to note that the information obtained from the application of the indices discussed in this section is not sufficient to establish the validity of the test item. Such information may even provide misleading data as these indices are susceptible to poor instruction. However, these indices do provide important clues which when used with the information obtained from the other conditions may enable the test constructor to make a decision regarding the validity of the test items.

3.3.2 Condition Two

In most test construction situations, multiple items are written to measure each available objective. The second condition is concerned with the calculation, examination, and comparison of the statistical characteristics of items measuring the same objectives to determine whether or not the items intended to measure the same objectives have similar statistical characteristics. The rationale behind Condition Two was first set forth by

Brennan and Stolurow (1971). Their position is that items which measure the same objectives should be equivalent--the items should have equal means, variances and intercorrelations. If these terms are satisfied, then the items are said to be corresponding. We share the opinion of Brennan and Stolurow that the concept of corresponding items is crucial for criterion-referenced tests. In this section a number of item characteristics considered relevant for criterion-referenced test items will be examined.

Since Brennan and Stolurow (1971) first set forth the concept of corresponding items, a detailed discussion of their approach is in order. The depict corresponding items as being analogous to the total test concept of parallelism. That is, if k tests are given to N students and the k means, k variances and the k(k-1)/2 intercorrelations are equal then the tests are parallel. The Brennan and Stolurow approach replaces the k tests with k items. Thus, a comparison of items consists of administering k items to N students to determine whether the k items have the same means, variances and whether the k(k-1)/2 interitem correlations are equal.

Since one of the major purposes of criterion-referenced tests is to provide information for making instructional decisions, the concept of corresponding items should be applied to items measuring the same objectives and which have been administered to the same group of examinees after instruction. While it would be possible to

compare items administered to different groups receiving the same instruction, the assessment problem would be more complex as the group and instructional effects would have to be taken into consideration.

In order to test the equality of the item means, Brennan and Stolurow (1971) have suggested the use of Cochran's Q-test (Cochran, 1950). This procedure is a test of hypothesis of equal correlated proportions and can therefore be used to determine whether two or more difficulty indices of dichotomously scored items differ significantly amongst themselves (Siegel, 1956).

When the number of examinees is sufficiently large, Q is approximately distributed as a Chi-Square variable with k-1 degrees of freedom where k is the number of test items. To reject the null hypothesis, however, provides no guidance as to which items are significantly different. This must be accomplished by setting up confidence bands for each pair of items.

The idea of the equality of interitem correlations of the items is also extremely important in determining whether or not items are corresponding in the Brennan and Stolurow sense. One technique which can be used in making this determination has been set forth by Hotelling (1940). This technique is used to test for the equality of pairs of product moment correlation coefficients. While one assumption underlying the use of Hotelling's technique is the normality of the item scores, this technique can

be used generally if the researcher is willing to assume that a departure from normality will not significantly reduce the power of the test. If this assumption is not tenable, researchers will have to make subjective judgments as to the equality of these interitem correlation coefficients.

The concept of item correspondence can be expanded to include procedures which help to identify faulty items. Specifically, the KI coefficient (Sabers and Kania, 1972) and Popham's (1971) use of a fourfold table from the results of a pre-test and post-test on a set of items measuring the same objective. Sabers and Kania (1972) noting the inadequacies of norm-referenced item reliability or validity indices for cirterion-referenced test items, have introduced a new item index, the KI coefficient that is used to compare test items. Using Cronbach's (1971) suggestion for determining the validity of a test plan, Sabers and Kania attempted to quantify the agreement between parallel criterion-referenced tests in the form of the KI coefficient. First, they eliminated the very high (90% and over) and the very low 40% and under) scoring students reasoning that the elimination of such students would set more stringent standards for determining item precision. For each item, they developed a 2 x 2 contingency table based on pass-fail decisions:

Item i, Form 1 \ Item j, Form 2	Pass	Fail
Pass	A	B
Fail	C	D

where A = Students who passed the corresponding items on both forms,

B = Students who passed the corresponding items on Form 1 and failed the items on Form 2,

C = Students who failed the corresponding items on Form 1 and passed the items on Form 2,

D = Failure on the corresponding items on both forms.

The author notes that while it would appear that the phi coefficient could be used to measure association, the use of that index would be disadvantageous because the difficulty level of the item affects the value of the phi coefficients and the phi coefficient does not enable the test user to determine which test has the poor item. With the Sabers and Kania technique, one can identify the faulty items by comparing the B and C cells. In making this comparison one calculates the index of precision for each item in each form. Form 1, item i precision is given as:

$$Di = 1 - B/N$$

For Form 2, precision of item j is given as:

$$Dj = 1 - C/N$$

These two values are then used to compute the coefficient of item equivalence, the KI coefficient:

$$KI = .5\ (Di + Dj)\ (1-|Di - Dj|).$$

The range of KI is from zero to one. The higher the value of KI, "The greater the degree of agreement between the decisions made by the two forms" (Sabers and Kania, 1972). Since the phi coefficient is not capable of discerning this degree of agreement between two forms, it cannot be used to determine which items need to be replaced or refined, whereas the KI coefficient can be so used.

Sabers and Kania added that the average (arithmetic mean) of all the KI indices for a subtest can be considered as a measure of the precision of the subtest. If we carry this a step further and define a subtest as consisting only of items measuring one objective we will have a means of determining how well these items reflect performance on the objective. The KI coefficient correlated very highly with a version of Findley's (1956) net discrimination index extended by Whitney and Sabers (1971) to handle the situation of multiscore items with two or more criterion groups. Consequently, Sabers and Kania concluded that additional studies will be required to determine the conditions under which these notions of item precision will be useful. The Saber and Kania item precision index appears quite promising.

3.3.3 Condition Three

The third condition of the empirical approach for assessing item validity is concerned with determining whether or not items are measures of only one objective. Davis and Diamond (1974) clearly defined the importance of this condition when they noted:

> Unless all of the items in a test measure exactly the same variable or variables for which true scores are highly correlated (say, .90 or greater), it is inappropriate to use the test for diagnostic purposes; that is, to determine an examinee's level of performance on a single "pure" variable. This is because of the fact that two different examinees may obtain identical scores by marking correctly the same number of different items. . .
>
> The implication for the preparation of homogeneous items for a multi-item diagnostic test is that each item must measure only one "pure" variable plus error or the same weighted combination of two or more "pure" variables, plus error. In either of these cases, the item scores would be found to measure, at a preselected level of significance, the same dimension except for errors of measurement and for differences of origin and of units of measurement. . . .

As indicated by Davis and Diamond (1974), Kriewell (1972) has also expressed the same idea in his statement:

> The item-sampling model described here as the paradigm for CRT construction is one of the simplest models. It places no conditions on the items except, to preserve score meaning, all items must share at minimum the objective attributes which serve to characterize a learning objective. . .

While statements such as these can be found in the literature, this condition hs been the least explored of the three; hence there is little information available to compare the relative merits of different procedures. Two

techniques that in theory offer the potential of providing useful information to address Condition Three are: Guttman's Scalogram Analysis (Guttman, 1950) and Factor Analysis (Lawley and Maxwell, 1971).

Guttman's Scalogram Analysis

In attempting to solve the problem of analyzing questionnaire data, Guttman (1950) developed a scaling procedure in order to ascertain whether or not a single factor or dimension underlies the responses to a set of test items. Basically Guttman's procedure, scalogram analysis, focuses on the ranking of individuals formed by an examination of their response patterns. The following example will help to clarify this statement. If there are n dichotomous items, there are 2^n possible response patterns. Thus, with three dichotomous items, there are eight possible binary response patterns. For the three dichotomous items to form a scale, there can only be four observable response patterns (*i.e.*, 111, 110, 100, 000).

As Torgerson (1958) has pointed out, in the Guttman scale the notion of reproducibility is crucial. That is, given a perfect scale one can reproduce the responses of the examinees from knowledge of his/her rank position alone. Therefore, an error is construed as "a response by a subject which would have been predicted wrongly on the basis of his assigned rank position" (Guttman, 1950). The amount of deviation from a perfect scale is measured by a "coefficient of reproducibility" which is given as:

$$\text{Reproducibility} = \frac{\text{Total Number of Errors}}{\text{Total Number of Responses}}$$

The total number of errors is determined by calculating the number of deviations for each examinee's response pattern from the ideal response pattern determined by the individual's test score.

This procedure appears to be relevant for the analysis of criterion-referenced test data as such data is usually collected from instructional programs which have hypothesized ordering of objectives. Scalogram analysis can be used to validate the existence of any hierarchical structure for the instructional objectives. In fact, Boozer and Lindvall (1971) used multiple scalogram analysis, a procedure developed by Lingoes (1963) to analyze "the response pattern of a set of dichotomous variables for the purpose of searching out optional scalable subsets within larger sets of data" (Boozer and Lindvall, 1971). Basically, multiple scalogram analysis can be used to locate groups of variables which form a scale when a single scale cannot be obtained from the total set of variables. Boozer and Lindvall (1971) point out that if the hypothesized ordering is not verified by the scalogram analysis either one of the following conclusions can be drawn:

1. The hypothesized scale is erroneous;
2. The items are not measures of the objectives they are purported to be.

Boozer and Lindvall concluded that their results "demonstrated the usefulness of both simplex and scalogram analysis for the purpose of assessing hypothesized hierarchical relationships among specified behavioral objectives as well as curriculum units. To use Guttman's scalogram analysis as a technique in an item validation methodology, one would need to know the hierarchical structure of a set of objectives. The fit of the responses to individual items to the perfect scale model could be used as a means of examining Condition Three.

Factor Analysis

While factor analysis is a commonly employed procedure for the dimensional analysis of test data, it has rarely been used in the area of criterion-referenced test construction. Nevertheless, it seems that the methods of factor analysis can be very effectively brought to bear on the problem of determining whether or not the item in a criterion-referenced test measure the objectives set forth. (Excellent references to factor analysis include Harman, 1967; Lawley and Maxwell, 1971; and Mulaik, 1972.)

Factor analysis is a statistical procedure developed primarily as a means of representing a group of variables $y_1, y_2, \ldots, y_p$ in terms of fewer underlying factors or hypothetical constructs, $x_1, x_2, \ldots, x_k$ ($k<p$). The most well known and mathematically tractable

model for factor analysis is the linear model in which

$$y_i = \lambda_{i1}x_1 + \lambda_{i2}x_2 + \ldots + \lambda_{ik}x_k + e_i$$
$$= \sum_{j=1}^{k} \lambda_{ij}x_j + e_i, \quad i = 1, 2, \ldots, p.$$

Here λ_{ij} is the "loading" of the ith variables on the jth factor and e_i is the random error component or the unique score. In matrix notation the factor model is more conveniently wirtten as

$$\underline{y} = \Lambda\underline{x} + \underline{e}$$

where

$\underline{y}$ is the (pxl) vector of observed variables,

$\underline{x}$ is the (kxl) vector of factors,

Λ is the (pxk) matrix of factor loadings, also known as the factor pattern,

and

$\underline{e}$ is the (pxl) vector of unique scores, or error.

Without loss of generality, we may assume that

$$E(\underline{y}) = E(\underline{e}) = \underline{0},$$

and

$$E(\underline{x}) = \underline{0},$$

where $E(\cdot)$ is the usual expectation operator. Further, we assume that the factors are correlated with the correlation matrix Φ and that the unique scores are uncorrelated with each other, *i.e.*,

$$E(\underline{x}\,\underline{x}') = \Phi$$

and

$$E(\underline{e}\,\underline{e}') = \Psi, \text{ a diagonal matrix.}$$

Since $\underline{x}$ and $\underline{e}$ are unknown quantities, it is not possible to estimate the matrix of factor loadings Λ from the factor model. In order to estimate the number of factors and the factor loadings, it is necessary to reformulate the model in terms of the variances and the covariances of the variables $y_1, y_2, \ldots, y_p$. If Σ is the (pxp) population variance-covariance matrix of the vector variable $\underline{y}$, it can be shown that the corresponding structural model is

$$\Sigma = \Lambda \Phi \Lambda' + \Psi;$$

and under certain conditions, the elements of Λ, Φ and Ψ can be estimated and the number of factors determined.

The factor model and the factor analytic procedure outlined above have been extensively used in exploratory studies where the number of underlying dimensions or factors are not known. An entirely different situation prevails in the realm of criterion-referenced tests, however. In the context of criterion-referenced tests, the variables $y_1, y_2, \ldots, y_p$ are the p items in the test and the factors $x_1, x_2, \ldots, x_k$ are the k objectives, which, unlike the situation in exploratory factor analysis, are usually known. The basic problem in criterion-referenced tests which is to ascertain whether or not the items are measuring the objectives, becomes, in the language of factor analysis, the problem of determining whether or not the factor pattern matrix has a prescribed form. As an example, suppose that there are nine items and three

objectives. Suppose further that items 1, 2 and 3 measure objective I, items 4 and 5 measure objective II and that the remaining items measure objective III. If this were true, then the factor pattern matrix, Λo, would have the form shown below:

$$\Lambda o = \begin{bmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ \lambda_{31} & 0 & 0 \\ 0 & \lambda_{42} & 0 \\ 0 & \lambda_{52} & 0 \\ 0 & 0 & \lambda_{63} \\ 0 & 0 & \lambda_{73} \\ 0 & 0 & \lambda_{83} \\ 0 & 0 & \lambda_{93} \end{bmatrix} .$$

Thus, the problem in criterion-referenced test construction is not merely to estimate the number of factors and the factor loadings, but to establish that the factor pattern has the structure shown above, or equivalently, exhibits simple structure.

Several procedures are currently available to test hypotheses regarding the structure of the factor pattern. A procedure that has been frequently used is to (i) estimate the factor pattern by any one method of factor extraction, (ii) rotate or transform the factor pattern to resemble

as closely as possible the target factor matrix Λo, (iii) compute any one of the acceptable indices of factor similarity (Harman, 1967, p. 268-272; Mulaik, 1972, p. 354-356), (iv) decide whether or not the fit is sufficiently good by examining the index of similarity. Though the above procedure is extremely simple to implement, it has several drawbacks. There are several measures of factorial similarity, none of them completely acceptable. Furthermore, no statistical tests are available for testing these indices.

In view of the above limitations, the maximum likelihood procedure for estimating the parameters is particularly interesting, since the maximum likelihood estimates have, at least asymptotically, the desirable properties of estimators. Furthermore, likelihood ratio tests can be construed to test the various hypotheses of interest.

The maximum likelihood estimates of Λ, Φ, and Ψ are those values of the parameters that minimize the following variant of the likelihood function

$$L = \log |\Sigma| + \text{tr}\,(S\Sigma^{-1})$$

where S is the sample estimate of Σ. Joreskog (1967, 1969) has outlined procedures for the minimization of the likelihood function and has made available the necessary computer software. Once the parameters are estimated, it is possible to test the hypothesis

$$H_o : \Lambda = \Lambda o$$

against

$$H_1 : \Lambda \neq \Lambda o$$

where o is the prescribed or the target matrix. If L is the minimum value of

$$L = \log |\Sigma| + tr\ (S\Sigma^{-1})$$

when the alternate hypothesis is true, and L_o is the value of the same function when the null hypothesis is true, then it can be shown that the quantity

$$\chi^2 = (L_o - L)$$

has, in large samples, the χ^2 distribution with n(k-1) degrees of freedom, where k is the number of objectives and n is the number of items. If $\chi^2 > \chi^2_{\alpha:n(k-1)}$, then the hypothesis that the factor pattern has the prescribed structure can be rejected and vice versa. Thus, it is possible to establish by accepting or rejecting the hypothesis whether or not the items measure the objectives.

3.3.4 Summary of the Empirical Approach to Assessing Item Validity

The empirical approach to item validation described in the last sections require what may appear to be stringent conditions. However, it should be pointed out that while each of the conditions deals with an important aspect of the item-objective relationship, none of them in and of themselves can provide sufficient information for the assessment of item validity. This follows because while items may satisfy all three conditions it still cannot be stated

definitively that the items measure the intended objectives. What can be said is that if an item fails to meet one of the conditions, it cannot be a valid measure of the objective. Thus, the three conditions can be used to detect "bad" items. The three conditions serve as necessary but *not* sufficient criteria for establishing item validity.

3.4 The Use of Judgments of Content Specialists to Assess Item Validity

Given the limitations of item generation rules and empirical methods for assessing item validity, the use of the judgments of content specialists assumes a degree of particular importance at this stage of development of a methodology for item validation. Therefore, this section will consist not only of an examination of this procedure as a means for assessing item validity but also attempt to begin the development of a methodology for the collection and reduction of this judgmental data.

The first step in the development of a methodology for the use of the judgments of content specialists to assess item validity is the identification of the important issues. Some of the important issues presented in question form are:

1. Can the content specialists make meaningful (valid) judgments about the relevance of items to instructional content? This question corresponds to instrumental validity.

2. Is there a consistency of agreement amongst the content specialists? This question corresponds to instrumental reliability.
3. What information is one seeking to obtain from the judgmental data?
4. What are the variables which effect the judgmental techniques?
5. What techniques can be used for collecting this data?

The first question concerning the ability of content specialists to make meaningful judgments was examined by Ryan (1968). He requested four judgments for each item. These judgments were:

1. How good or poor is the item for determining knowledge and understanding of the instructional content presented in each of your classes?

   Very poor Poor Fair Good Very Good

2. What proportions of pupils in each class will answer the item correctly?

   0 .20 .40 .60 .80 1.00

3. How much better will the most proficient third of the pupils in each class do on the item compared to the least proficient third.

   Same Slightly Better Somewhat Better Much Better Very Much Better

4. How appropriate or relevant is the item for the instructional materials and content presented in each class?

   Not Relevant Somewhat Relevant Quite Relevant Very Relevant

Ryan (1968) concluded that teachers can make judgments about test items on two dimensions: (1) the relevance of the items to the instructional content; (2) the difficulty of the item. He based his conclusions on results which showed a "relatively higher frequency with which relevance as compared to judged difficulty was correlated with overall quality and the relatively higher frequency with which judged difficulty, as compared to relevance, was correlated with actual difficulty."

While Ryan's (1968) study is a step in the right direction, his conclusion on the issue of relevance is weakly supported in that one has no way of knowing whether the teachers perceive the judgment of quality the same as a judgment of relevance. On the other hand, the judgment of difficulty correlated highly with actual difficulty which gives a more conventional substantiation of judgmental validity. Since with criterion-referenced measures, the question of item relevance is of utmost concern, it is suggested that a similar form of validation be considered. For example, one could correlate the judgmental data with the objective scores (scores obtained by summing across items measuring the same objectives). Another approach might be to obtain correlations of the judgment data on each item and then use a data reduction procedure to test against an hypothesized structure. In Chapter IV, several exploratory techniques will be used in order to ascertain which ones are most appropriate for validating the judgmental procedures.

The second question concerning the consistency of agreement amongst the content specialists, or actually the reliability of the instrument, has been examined by a number of researchers (Lu, 1971; Cohen, 1960; Light, 1971; Fleiss, 1971; and Brennan and Light, 1973). It is not our intention to review this extensive literature. However, a description of two methods for assessing agreement amongst content specialists will serve to outline the direction of our research in this area.

Lu (1971) has presented a method by which one can ascertain the intensity of agreement amongst judges to an instrument requiring a classification of items into a set of ordered categories. The observed results of such a rating procedure is given as follows:

	J_1	J_2 . . . J_j . . . . J_m
S_1	X_{11}	X_{12} . . . X_{ij} . . . . X_{1m}
S_2	X_{21}	X_{22} . . . X_{2j} . . . . X_{sm}
S_i	S_{i1}	S_{i2} . . . X_{ij} X_{im}
S_n	X_{n1}	X_{n2} . . . S_{nj} . . . . S_{nm}

where X_{ij} is the judgment of the ith item by the jth judge. X_{ij} may take on the values of the t categories.

Lu derives a set of weights for each category "based on a transformation from the data's own distribution." These weights are derived from the following array:

	C_1	C_2	. . . .	C_k	. . . .	C_t	Σ
J_1	n_{11}	n_{12}		n_{1k}		n_{1t}	n
J_2	n_{21}	n_{22}		n_{2k}		n_{2t}	n
J_j	n_{j1}	n_{j2}		n_{jk}		n_{jt}	n
J_m	n_{m1}	n_{m2}		n_{mk}		n_{mt}	n
Σ	n_1	n_2		n_k		n_t	n_n

where n_{jk} is the count of items placed in the kth category by the jth judge. The weight, y for the kth category is defined as:

$$Y_k = \sum_{r=1}^{k-1} P_r + \frac{1}{2} P_k \qquad k=1,2...t$$

where $P_r = n_r/nm$.

After the transformed weights are obtained as analysis of variance is conducted. Then a ratio of the observed within subject variance (Si) over the expected within subject variance (SE) under the conditions that all the ratings were equally likely. The coefficient of agreement is defined as follows:

$$A = \frac{SE-Si}{SE} \quad .$$

The significance of A is tested indirectly. Under the hypothesis that the assignment of items are random, the following would hold

$$SE = Si$$

Thus the statistic

$$\theta = \frac{Si}{SE}$$

is χ^2/df distributed with n(m-1) degrees of freedom (Lu, 1971). If the hypothesis is rejected, one can conclude that A is significant.

Another method which can be used when the data is in a contingency table format has been presented by Light (1971). This method consists of comparing group agreement to a specified standard to obtain a 'G' statistic. This statistic is normally distributed for large samples and "tests the null hypothesis that the joint group's assignments indicate chance agreement with the standard."

Again, since there is little or no research using these methods within the context set forth in this study, a number of them will be examined in Chapter IV in order to determine those which are more appropriate.

The third question related to the information which one seeks to obtain from the judgments of content specialists with regard to determining item validity. It would seem that such judgments should provide two categories of information: (1) information which is considered essential; and (2) there are two types of information which must be collected. These types are given as follows:

1. Information relating to whether or not an item is judged to be a measure of an objective,
2. Information relating to whether or not an item is judged to be a measure of more than one objective.

The choice of the types of information which is to be collected under the second category will vary from study

to study as they are dependent on secondary goals or methodological considerations. Examples of secondary goals would be the determination of whether or not the content specialists can judge the difficulty of the items or whether the items were well written. An example of a methodological consideration would be the collection of data which would help validate the instrument.

The fourth question concerning the variables which effect the judgments of content specialists has not been extensively researched. However, the importance of this question has been demonstrated by Girard and Cliff (1973). In comparing methods for judging the similarity of personality inventory items, they found that "the criteria by which subjects were instructed to judge similarities between items in a pair made a large difference in the judgments." Three of these variables which are felt to be important are given as follows:

1. Judgmental Procedures: whenever possible, one should use the simplest of techniques available to collect data. For example, usually, categorical judgments obtained from sorting, rating and ranking procedures are less complex than comparative judgements obtained from similarity, dissimilarity or choice procedures.
2. Format of Presentation: the response task should not be tedius and time consuming. For example, while there are methods which can be used to reduce

the number of required responses (Torgeson, 1958), generally the method of paired comparisons should be avoided if the number of stimuli (items) is large, because of the great number of responses involved.

3. Definition of Task: when describing the response task, one should ensure that all the judges are operating under the same assumptions. If one merely asks the judges to rank or choose items according to personal preference, the judges could obtain significant results based not on real differences in the items but on the dimension of preference. For example, the judges could have been ranking the items on any one of the following levels of the preference dimension:
   a. simplicity/complexity of item,
   b. closeness of match
   c. response mode required,
   d. style in which the item was written.

   The directions relating to the response task must clearly define the criteria on which the choices are to be made.

4. Settings for Data Collection: in choosing an instrument for collecting the judgments of content specialist, the setting in which the data is to be collected must be taken into consideration.

That is, the practicality of its use in both research and non-research settings is a key factor in the choice of instrument.

The fifth question is concerned with the choice of instrument which will be used to collect the judgmental data. It is suggested that the researcher choose a technique which conforms as closely as possible to the guidelines set forth under the discussions on questions 1, 2 and 4 while providing the information described in question 3.

In the next section, three procedures for collecting judgmental data to assess item validity will be examined: the Hemphill-Westie (1950) index of homogeneity of placement, (a categorizing procedure); the Semantic Differential (Osgood, Suci and Tannenbaum, 1957) (a rating procedure); and a matching procedure.

3.4.1 An Index of Item Homogeneity

Hemphill and Westie (1950) developed an index of homogeneity of placement for use in constructing personality tests. This index is a numeric representation of the judgment of content specialists on the extent to which they feel that an item belongs to one and only one personality dimension. By substituting "objective" for "personality dimension," the Index of Item Homogeneity can be used in item validation work.

According to Hemphill and Westie (1950).

> This index was adopted to give a single numerical evaluation of each item with respect to its homogeneity. Agreement among judges that the item applied to a dimension and agreement that it did not apply to other dimensions in the description were given approximately equal weight in the value of this index.
>
> The index of 'homogeneity of placement' differs in two ways from certain other techniques for examining item content. First, it is based on 'expert' judgment of probable response to the items, not on actual item response data. Second, unlike indices such an 'internal consistency,' 'homogeneity,' or 'unidimensionality' all of which refer to relationship among items, the index of 'homogeneity of placement' involves both relationships among items (as reflected by judge agreement that certain items apply to the same dimension) and independence of relationship of the item to other dimensions making up the same general heuristic system.

Since this index provides both types of information needed for assessing item objective congruence, it appears to be a valid procedure for collecting the analyzing judgmental data on item validity.

The mechanics for collecting data through the use of the Hemphill-Westie consists of having the content specialists rate each item on each of the objectives by assigning a value of +1, 0 or -1 where

+1 = I definitely feel that the item is a measure of the objective

0 = I cannot decide whether the item is a measure of the objective

-1 = I definitely feel that the item is not a measure of the objective.

The formula presented by Hemphill and Westie (1950) to compute the index of homogeneity of placement is given

as follows:

$$I_{ik} = \frac{N \sum_{j=1}^{n} X_{ijk} - \sum_{i=1}^{N} \sum_{j=1}^{n} X_{ijk}}{2[2n(N-1)] = \sum_{i=1}^{N} \sum_{j=1}^{n} X_{ijk} - \sum_{j=1}^{n} X_{ijk}}$$

where

I_{ik} is the Index of Homogeneity for item k on objective i

N is the number of objectives (i=1,..,N)

n is the number of content specialists (j=1,.,n)

X_{ijk} is the value of the rating, +1, 0, -1, assigned to item k on objective i by content specialist j.

While the Hemphill-Westie procedure is conceptually appropriate for the task of collecting judgmental data from content specialists for the purpose of assessing item validity, the computational formula given above has some serious deficiencies. First, while the maximum value of this index which will occur when each content specialist assigns a +1 to the item for the appropriate objective and a -1 for all the other objectives is .67, its minimum value, which occurs when content specialists assign a -1 to the item for the appropriate objective and a +1 for all the other objectives, is -.40. Given the range of rating values, for ease of interpretation, the maximum and minimum values ideally should be +1 and -1 respectively. Second, and more seriously the value of the index will vary as a function of the number of content specialists and objectives

making the choice of a cutoff score to separate "good" from "bad" items arbitrary.

Given these deficiencies, this author has developed a new computational formula for providing a numerical representation of Hemphill-Westie data. This new formula will be called the *Index of Item-Objective Congruence.* The assumptions under which this index was developed are:

1. That perfect item objective congruence should be represented by a value of +1 and will occur when all the specialists assign a +1 to the item for the appropriate objective and a -1 to the item for all the other objectives.
2. That the worst judgment an item can receive should be represented by a value of -1 and will occur when all the specialists assign a -1 to the item for the appropriate objective and a +1 to the item for all the other objectives.
3. That the assignment of a 0 to an item is poorer than a +1 but better than a -1. This is in effect saying that it is better for a specialist to not be able to definitely decide whether an item is a measure of an appropriate objective than it is for the judge to feel that the item is definitely not a measure of the objective.
4. That this index should be invariant to the number of content specialists and the number of objectives.

The new computational formula is given as follows:

$$I'_{ik} = \frac{(N-1)\sum_{j=1}^{n} X_{ijk} - \sum_{l=1,\, l\neq i}^{N}\sum_{j=1}^{n} X_{ljk}}{2(N-1)n}$$

where I'_{ik} is the Index of Item-Objective Congruence for item k on objective i,

N is the number of objectives,

n is the number of examinees,

X_{ijk} is the value of the rating, +1, 0, -1, assigned to item k on objective i by content specialist j.

The choice of cutoff score for this index can now be based on some absolute standard relating to specific proportions of perfect ratings for the items. For example, if one-half of the content specialists judged an item to be a perfect match to an objective, while the others were not able to make a decision, the computed value of the index would be .50. Thus, test constructors obtaining I' value of .50 would know that at a minimal level at least 50 percent of the content specialists gave a perfect rating to the item.

As with the Hemphil-Westie Index there is no means for determining the statistical significance of the values for the Index of Item-Objective Congruence. However, the use of Lu's coefficient of agreement amongst the judges will give an indication of how reliable (consistent) the

judgments were. This indication of consistency of judgments along with the known values that the index would take with specific proportions of perfect ratings will give test constructors some idea as to how meaningful a particular I' value is for an item.

It is therefore suggested that the assignment of the values, 1, 0 and -1 to the items be considered the same as assigning the items to one of three categories. Then one can transform this data and calculate Lu's coefficient of agreement. The statistical significance of that coefficient can be determined. This will give some indication of whether or not the content specialists are consistent in their assignment of values.

3.4.2 Semantic Differential Technique

The second procedure employs the use of the semantic differential procedure (Osgood, Suci and Tannenbaum, 1957). The content specialists are presented with an objective and all the items on which ratings are desired. They are asked to make a judgment which consists of deciding whether the item objective relationship is best described by the adjective toward the left end or toward the right end of the scale.

The following is an example consisting of one objective, one item and two objective scales along with a set of typical directions:

Objective: Given the chemical formula for a molecule, determine the number of atoms in a molecule.

Item 1: How many atoms are there in a molecule of sulfuric acid H_2SO_4?

Scale 1:

very relevant	relevant	feeling	irrelevant	very irrelevant
1	2	3	4	5

Scale 2:

very unimportant	unimportant	feeling	important	very important

Directions

Given the objective at the top of each page, you are to make judgments on the relationship between it and all the items below it by circling the values on the adjective scales which you feel are most appropriate.

The data obtained from the use of this technique can be analyzed without employing any elaborate statistical procedures. Therefore, it can easily be used in practical settings such as in the classroom by teachers. The information which is needed is the scale mean score for each item on each objective. However, the data also lends itself to more elaborate statistical analysis if required. An examination of the standard deviations of each scale on the objectives summed over the content specialists will give an indication of the extent of agreement among the content specialists.

3.4.3 A Matching Procedure

The third procedure which is used to obtain the judgments of content specialists involves the use of a matching task. The content specialists are presented with two lists. The first list contains a series of items. The second list is composed of the objectives. The judgment involves assigning a number representing an objective to the most appropriate item. A contingency table can then be constructed from the frequencies obtained from this procedure. According to Light (1971), the Pearson Chi-Square test for independence is commonly used to analyze data which is presented in a contingency table format. One tests the hypothesis that "the responses in each column follow the same probability distribution over the rows" (Light, 1971).

However, such an analysis would not provide the information required to make meaningful statements concerning item validity. Even if one utilizes Light's (1971) G statistic discussed earlier, one does not have specific information on each item. That is, in the event the null hypothesis is accepted, deviations from the standard are not identified.

3.4.4 Summary of the Use of Content Specialists to Assess Item Validity

In this section an attempt has been made to initiate the development of a methodology for use of content

specialists in assessing item validity. Further, three techniques for the collection and analysis of the judgments of content specialists have been described. These techniques were chosen primarily to provide information on the efficacy of the use of content specialists as a means for assessing item validity not as definitive answers to the question of which techniques are most appropriate. However, an assessment of each of these techniques will be made in Chapter IV with regard to their effectiveness in assessing item validity.

3.5 Conclusion

In this chapter, three procedures relating to the problem of item validity have been examined. The first of these procedures, the use of item generation rules, attempts to ensure item validity by developing a direct relationship between an item and objective during the construction phase. As such it is an *a priori* approach as compared to the other *a posteriori* procedure which is designed to assess whether or not a direct relationship between an item and an objective exists through analyses of data conducted after the item is written. However, the use of item generation rules as currently formulated contain inherent problems which make their implementation impractical. Since, these rules appear to be an effective way of obtaining item validity, a less formal item writing procedure, the use of amplified objectives, has been proposed as a compromise which encompasses

the positive aspects but reduces the negative aspects of such procedures.

The second approach, the use of empirical procedures, has been presented with a few notes of caution. The notes of caution were based on the following issues:

1. These procedures are dependent upon group composition and instructional effects.
2. They require sophisticated statistical techniques and computer programs which are not available to the practitioner.
3. The use of these procedures to eliminate items may disrupt the item sampling plan which is essential in generalizing from the examinee's score to the whole content domain.
4. Some of the possible techniques require pre-test and post-test data which is rarely collected in classroom settings.

In situations where a large sample of examinees is available and where one is interested in identifying aberrant items not for elimination but for correction, the use of the empirical approach to item validation should provide important information with regard to the assessment of item validity.

The third procedure, the use of the judgments of content specialists, appears to offer promise as a means for assessing item validity. This procedure is not dependent on group composition or instructional effects; may not

require sophisticated statistical techniques; is not restricted to highly structured content domains; and can be implemented easily in practical settings. While there has been a great deal of discussion on the use of this technique in the literature, no systematic implementation of this procedure has been reported. Material presented in the next chapter represents an initial step to incorproate content specialist's ratings into the development of a methodology for determining item validity.

CHAPTER IV

AN EXAMINATION OF TWO APPROACHES USED IN THE ASSESSMENT OF ITEM VALIDITY

4.1 Introduction

In this chapter, a comparative study of two approaches to the problem of assessing item validity, the use of empirical analyses and the use of the judgments of content specialists, is presented. The purposes of the study were to determine the appropriateness of these two approaches to assessing item validity.

The empirical information on the items studied in this chapter was obtained from two different sets of examinees. While both sets of examinees received a similar individualized science instructional package on *the structure of matter*, the tests administered and the testing formats used were different. Therefore, it is important that in the ensuing discussions, the empirical information on the test items from these two sources be clearly differentiated. In order to accomplish this differentiation, the information obtained from examinee source one will be labeled Data Set One while the information on test items from examinee source two will be labeled Data Set Two.

The test information for Data Set One was collected through pre-instructional and post-instructional administrations of the same test form to 294 ninth grade students (Hambleton, 1971; Sheehan and Hambleton, 1972). This testing format resulted in the collection of both pre-test and post-test information for each of forty items measuring the eleven instructional objectives presented in the science unit. The expected match between the forty items and their intended objectives is presented in Table 4.1.1

The test information for Data Set Two was collected through the use of a more complex research design (Royer, Hambleton and Cadorette, 1975). First, three parallel test forms (denoted A, B, and C), each consisting of forty-eight items, were constructed to measure twelve instructional objectives. The intended match between items and objectives for Data Set Two is presented in Table 4.1.2. Second, the examinee source, consisting of 185 students, was subdivided into three groups on the basis of a random assignment. Third, each of these three groups were then administered the three parallel test forms on the pre-test, post-test and delayed post-test occasions. The first group received them in the order, A, B, C; the second, B, C, A; and the third, C, A, B. This testing design was chosen because it ensured that each of the three forms of the test was administered on each test occasion and no group of examinees saw the same form twice. However, since

TABLE 4.1.1

EXPECTED MATCH BETWEEN THE TEST ITEMS AND THE OBJECTIVES THEY ARE INTENDED TO MEASURE

(Data Set One)

Objective	Test Items
1	1, 2
2	3, 4, 7, 9
3	5, 6, 8, 10
4	11,.12, 13, 14, 15, 16, 17, 18, 19, 20, 21
5	22, 23
6	24, 25
7	26, 27, 28
8	29, 30, 31
9	32, 33, 34
10	35, 36, 37
11	38, 39, 40

TABLE 4.1.2

EXPECTED MATCH BETWEEN THE TEST ITEMS AND THE OBJECTIVES THEY ARE INTENDED TO MEASURE

(Data Set Two)

Objective	Test Items
1	1, 13, 25, 37
2	2, 14, 26, 38
3*	3, 15, 27, 39
4	4, 16, 28, 40
5	5, 17, 29, 41
6	6, 18, 30, 42
7	7, 19, 31, 43
8*	8, 20, 32, 44
9	9, 21, 33, 45
10*	10, 22, 34, 46
11	11, 23, 35, 47
12*	12, 24, 36, 48

*Objectives comprising the item forms group of data set two.

each of the examinee groups in Data Set Two was administered a different form of the test during each of the three administrations, any analyses conducted on items across testing occasions would not be based on the same examinees and therefore many of the available statistics could not be applied. The decision then was made to study only those parallel items which appeared to have been constructed through the use of an item form. Through a visual examination of the items contained in the three parallel forms, sixteen items measuring four objectives were identified as having been so constructed.

The data analyses presented in this chapter were conducted using Fortran IV computer programs written by the author. Exceptions are indicated within the test.

4.2 The Use of Empirical Analyses to Assess Item Validity

As set forth in Section 3.2, the use of empirical analyses to assess item validity, consists of determining whether the items satisfy three empirical conditions:

1. That items provide information which enables the test administrator to differentiate between those examinees who have received instruction and those who have not.

2. That items which measure the same objective have similar statistical indices.

3. That items not be a measure of more than one objective.

In this section, the items from Data Sets One and Two will be examined to determine whether or not these

items satisfy all three conditions. Summary statistics for the groups of examinees on the criterion-referenced tests are reported in Tables 4.2.1 and 4.2.2. A further goal of this section is to report on the reasonableness of the three empirical conditions as well as the indices and techniques that are available for assessing each condition.

4.2.1 Condition One

In order to ascertain whether an item provided information which could be used to distinguish between those examinees who have had instruction and those who have not, three statistical indices (the difference index, Saupe's change index and the point biserial correlation) were calculated. (Since, Iven's Iv statistic requires delayed post-test data, it was not calculated for the items in Data Set One.) The results of these calculations are presented in Tables 4.2.3 and 4.2.4. A summary of the results for Data Set One is presented below:

1. The values of Saupe's (1966) change index were so low as to not provide any reliable information as to the effectiveness of items in measuring change from the pre-test to the post-test occasion.
2. All items with difference index (DI) values lower than .40 were identified as not having satisfied condition one. The value .40 was selected as it is an approximation to the proportion of pre-test

TABLE 4.2.1

SUMMARY STATISTICS FOR DATA SET ONE

Objective	Number of Examinees	Number of Items	Pre-test Mean	Pre-test Variance	Post-test Mean	Post-test Variance
1	294	2	.74	.48	1.26	.49
2	294	4	1.99	1.39	3.19	1.03
3	294	4	1.62	1.58	2.60	1.49
4	294	11	2.84	6.96	7.96	5.80
5	294	2	.25	.34	1.61	.48
6	294	2	.30	.37	1.73	.36
7	294	3	.31	.35	1.97	1.14
8	294	3	.70	.91	2.18	.92
9	294	3	.21	.25	1.89	1.22
10	294	3	.37	.50	1.90	1.39
11	294	3	.29	.40	1.62	1.53
TOTAL	294	40	9.60	42.76	27.89	75.09

TABLE 4.2.2

SUMMARY STATISTICS FOR DATA SET TWO

Examinee Group	Number of Examinees	Objective	Number of Items	Pre-test Mean	Pre-test Variance	Post-test Mean	Post-test Variance	Delayed Post-test Mean	Delayed Post-test Variance
I	60	1	4	.98	.90	2.11	1.73	2.03	1.69
	60	2	4	1.62	1.16	3.10	.91	2.72	1.33
	60	3*	4	1.43	2.35	2.27	2.50	2.10	1.35
	60	4	4	.25	.26	1.57	1.10	1.28	1.22
	60	5	4	2.03	1.69	2.28	1.94	2.43	1.64
	60	6	4	.80	.77	2.53	1.61	2.15	1.82
	60	7	4	.32	.29	1.70	1.43	1.37	1.49
	60	8*	4	.53	.56	1.43	2.08	1.57	2.05
	60	9	4	1.35	1.38	2.40	1.53	1.88	1.70
	60	10*	4	.43	.72	1.82	2.83	1.58	2.15
	60	11	4	.25	.33	1.82	1.51	1.72	2.27
	60	12*	4	.43	.49	1.55	1.34	.92	1.37
	60	TOTAL	48	10.43	38.62	24.58	114.38	21.75	118.19

TABLE 4.2.2 (continued)

Examinee Group	Number of Examinees	Objective	Number of Items	Pre-test Mean	Pre-test Variance	Post-test Mean	Post-test Variance	Delayed Post-test Mean	Delayed Post-test Variance
II	65	1	4	1.03	1.16	1.98	1.42	1.31	1.47
	65	2	4	1.85	1.67	2.62	1.30	2.14	1.06
	65	3*	4	.86	1.21	1.94	1.00	2.83	1.89
	65	4	4	.49	.50	1.22	1.14	1.00	.94
	65	5	4	1.20	.94	1.81	1.68	2.26	1.54
	65	6	4	.71	.84	2.37	1.49	1.94	1.71
	65	7	4	.23	.31	1.43	1.09	1.43	1.50
	65	8*	4	.23	.21	1.17	1.27	1.09	1.05
	65	9	4	1.43	1.03	1.80	1.57	1.65	1.23
	65	10*	4	.31	.53	1.54	2.10	1.40	2.09
	65	11	4	.38	.52	1.80	1.73	1.70	1.80
	65	12*	4	.51	.50	.72	.95	1.22	1.27
	65								
		TOTAL	48	9.23	25.24	20.40	86.37	19.96	87.66

TABLE 4.2.2 (continued)

Examinee Group	Number of Examinees	Objective	Number of Items	Pre-test Mean	Pre-test Variance	Post-test Mean	Post-test Variance	Delayed Post-test Mean	Delayed Post-test Variance
III	60	1	4	.97	1.19	1.97	1.80	2.12	1.76
	60	2	4	1.37	1.08	2.68	1.17	2.88	1.12
	60	3*	4	.85	1.11	3.05	1.84	2.58	1.98
	60	4	4	.22	.21	1.32	1.47	1.23	1.23
	60	5	4	1.28	1.29	2.35	1.89	1.92	2.08
	60	6	4	.80	.87	2.83	1.67	2.52	1.41
	60	7	4	.23	.25	1.65	1.72	1.23	1.67
	60	8*	4	.33	.33	1.38	1.43	1.10	1.31
	60	9	4	1.01	1.20	1.90	1.35	2.28	1.46
	60	10*	4	.27	.47	1.78	2.48	1.93	2.23
	60	11	4	.13	.19	2.07	2.03	1.65	1.79
	60	12*	4	.25	.26	1.22	1.60	1.18	1.24
		TOTAL	48	7.72	29.66	24.20	107.65	22.63	91.59

*Objectives comprising the item forms subgroup.

TABLE 4.2.3

ITEM VALIDITY STATISTICS TO INVESTIGATE

EMPIRICAL CONDITION ONE ON DATA SET ONE

Test Item	Difference Index	Saupe's Change Index	Test Item	Difference Index	Saupe's Change Index
1	.50	.10	21	.59	.16
2	.02	.06	22	.69	.20
3	.30	.10	23	.67	.19
4	.30	.13	24	.70	.21
5	.40	.10	25	.72	.21
6	.07	.07	26	.56	.13
7	.11	.12	27	.61	.15
8	.40	.09	28	.49	.13
9	.49	.11	29	.34	.13
10	.10	.08	30	.59	.16
11	.25	.17	31	.55	.16
12	.56	.21	32	.52	.15
13	.56	.18	33	.54	.17
14	.32	.16	34	.62	.17
15	.55	.18	35	.46	.17
16	.68	.17	36	.59	.18
17	.58	.16	37	.47	.13
18	.58	.09	38	.47	.13
19	.05	.09	39	.38	.16
20	.39	.08	40	.48	.14

TABLE 4.2.4

ITEM STATISTICS CALCULATED ON DATA SET TWO FOR THE

ASSESSMENT OF ITEM VALIDITY

(Empirical Condition One)

Test Item	Examinee Group I				Examinee Group II				Examinee Group III			
	DI_{AB}	r	I_v	DC_{AC}	DI_{BC}	r	I_v	DI_{BA}	DI_{CA}	r	I_v	DI_{CB}
3	.13	.07	.10	.40	.64	.07	.62	.69	.42	.08	.35	.43
15	.37	.10	.36	.35	.38	.10	.37	.35	.43	.11	.38	.35
27	.17	.02	.09	-.32	-.25	.09	-.07	.46	.78	.03	.63	.30
39	.17	.07	.16	.23	.29	.09	.24	.46	.57	.07	.43	.07
8	.25	.08	.21	.12	.12	.06	.09	.35	.40	.10	.36	.30
20	.25	.06	.24	.27	.09	.03	.09	.11	.12	.07	.11	.18
32	.05	.03	.04	.25	.42	.03	.36	.28	.35	.04	.23	.02
44	.35	.08	.33	.40	.31	.12	.25	.12	.18	.07	.17	.27
10	.33	.09	.31	.40	.40	.10	.36	.31	.35	.04	.32	.45
22	.33	.09	.31	.27	.22	.12	.21	.26	.43	.10	.43	.45
34	.25	.10	.23	.15	.35	.10	.34	.32	.50	.11	.43	.37
46	.47	.06	.40	.33	.26	.14	.25	.20	.23	.14	.19	.40
12	.42	.04	.31	.15	.00	.03	.00	.08	.30	.04	.29	.27
24	.43	.00	.25	.02	.00	.06	.00	.17	.23	-.01	.19	.40
36	.18	.07	.18	.20	.18	.09	.17	.26	.25	.07	.22	.13
48	.08	.06	.08	.11	.03	.02	.03	.20	.18	.06	.17	.13

to post-test gain for the total group. The following items for Data Set One did not satisfy condition one: 2, 3, 4, 6, 7, 10, 11, 14, 19, 29.

For Data Set Two, a summary of the results is as follows:

1. Saupe's (1966) change index provided little reliable information for the items for any of the examinee groups.
2. The following items, organized by examinee group, were identified as not having satisfied condition one on the basis of having difference index (DI) values less than .25:

Examinee Group	Items
I	3, 27, 32, 36, 39, 48
II	8, 12, 20, 22, 24, 27, 36, 48
III	20, 44, 46, 48

3. The use of Iven's Iv index provided no additional information than that which was obtained from the use of the difference index (DI).

4.2.2 Condition Two

The determination of whether or not items measuring the same objectives have similar statistical indices was made by calculating and examining where appropriate the indices presented in Table 4.2.5. The results of these calculations for Data Sets One and Two are presented in Tables 4.2.6 and 4.2.7 respectively.

TABLE 4.2.5

ITEM INDICES USED FOR THE ASSESSMENT

OF ITEM VALIDITY

(Empirical Condition Two)

Index	Symbolic Representation	Formula
Difficulty level	p_i	$p_i = R_i/N$
Point Biserial (corrected)	r_i	$r_i = \frac{k}{k-1} \cdot \frac{r_{pbis}\, S - p_i(1-p_i)}{S^2 - \sum_j p_j(1-p_j)}$
D Statistic	d_i	$d_i = (U_i - L_i)/n$
Item Precision	Form 1 IP_i	$IP_i = 1 - B_1/N$
	Form 2 IP_j	$IP_j = 1 - C_2/N$
KI Coefficient	KI_{ij}	$KI_{ij} = -.5(IP_i+IP_j)\ (1-\lvert IP_i-IP_j \rvert)$

Explanation of Symbols

R_i = the number of correct answers to item i

N = the total number of examinees

S = the standard deviation of the N examinee test scores

U_i = the number of correct responses in the group comprising the upper 27 percent of the examinees

L_i = the number of correct responses in the group comprising the lower 27 percent of the examinees

n = .27 N

B_1 = the number of invalid passes on item i for Test Form 1

C_2 = the number of invalid passes on item j for Test Form 2. Test Form 1 and 2 are considered parallel and therefore items i and j are corresponding items

k = number of items

r_{pbis} = point biserial correlation

TABLE 4.2.6

ITEM STATISTICS CALCULATED ON DATA SET ONE

FOR THE ASSESSMENT OF ITEM VALIDITY

(Empirical Condition Two)

Item	p	r	d	Q/df	Comments Based on Q Statistic
1	.73	.44	.36		
2	.53	.26	.52	21.40/1*	Item 1 is more difficult than item 2.
3	.62	.46	.33		
4	.84	.50	.53		Item 3 is more difficult than items 4, 7 and 9.
7	.86	.44	.40		
9	.86	.38	.40	97.29/3*	
5	.73	.51	.27		Two clusters of items exist. Items 6 and 10 are more difficult than items 5 and 8.
6	.53	.37	.39		
8	.78	.39	.44		
10	.56	.41	.57	84.60/3*	
11	.92	.49	.59		Three clusters of items exist. Items 14, 19 and 20 are more difficult than items 15, 18, 21 and items 11, 12, 13, 16 and 17.
12	.85	.63	.71		
13	.86	.50	.72		
14	.54	.55	.53		
15	.77	.60	.43		
16	.88	.51	.47		
17	.84	.47	.66		
18	.77	.50	.40		
19	.18	.04	.30		
20	.60	.45	.26		
21	.74	.53	.38	776.42/10*	
22	.83	.62	.33		
23	.78	.62	.32	4.20/1	No differences.
24	.88	.59	.48		
25	.85	.53	.33	2.10/1	No differences.
26	.67	.49	.18		
27	.71	.47	.20		Item 28 is somewhat more difficult than items 26 and 27.
28	.59	.42	.58	16.40/2*	
29	.63	.42	.58		
30	.86	.55	.50		Items 29 and 31 are more difficult than item 30.
31	.68	.60	.29	62.05/2*	

TABLE 4.2.6 (Continued)

Item	p	r	d	Q/df	Comments Based on Q Statistic
32	.59	.57	.16		
33	.63	.57	.20		
34	.67	.59	.15	7.76/2*	Item 32 is somewhat more difficult than items 33 and 34.
35	.62	.68	.40		
36	.67	.73	.20		
37	.60	.33	.28	5.60/2	No differences.
38	.53	.54	.15		
39	.50	.62	.27		
40	.57	.51	.23	5.95/2	No differences.

*Significant at the .05 level.

TABLE 4.2.7

KEM STATISTICS CALCULATED ON DATA SET TWO

FOR THE ASSESSMENT OF ITEM VALIDITY

(Empirical Condition Two)

Test Item	Examinee Group I[1]									Examinee Group II[2]									Examinee Group III[3]								
	Post-test			Delayed Post-test			Item Precision			Post-test			Delayed Post-test			Item Precision			Post-test			Delayed Post-test			Item Precision		
	p	r	d	p	r	d	IP_B	IP_C	KI_{BC}	p	r	d	p	r	d	IP_C	IP_A	KI_{CA}	p	r	d	p	r	d	IP_A	IP_B	KI_{AB}
3	.47	.66	.88	.73	.54	.65	1.00	.68	.57	.74	.45	.58	.78	.32	.44	.95	.86	.83	.78	.40	.44	.61	.48	.70	.82	.94	.77
15	.68	.61	.81	.67	.66	.88	1.00	.88	.83	.66	.50	.58	.63	.50	.79	.86	.84	.83	.65	.55	.63	.78	.38	.38	.91	.86	.85
27	.56	.52	.63	.08	-.17	-.11	.46	.94	.34	.03	-.26	-.11	.74	.45	.60	1.00	.14	.08	.80	.44	.56	.60	.33	.40	.73	1.00	.63
39	.55	.49	.63	.51	.32	.48	1.00	.59	.47	.51	.20	.37	.68	.48	.60	.95	.68	.60	.81	.32	.44	.58	.42	.58	.73	.97	.65
8	.40	.68	.76	.27	.57	.61	.76	.94	.70	.17	.35	.26	.45	.35	.46	.92	.62	.40	.45	.50	.66	.35	.50	.71	.73	.88	.69
20	.33	.52	.59	.35	.53	.66	.84	.78	.76	.20	.40	.26	.21	.18	.10	.86	.84	.83	.22	.43	.36	.28	.37	.24	.88	.86	.85
32	.28	.46	.50	.48	.49	.60	.94	.65	.57	.45	.37	.37	.31	.15	.23	.70	.84	.63	.48	.21	.25	.15	.15	.10	.64	.97	.54
44	.42	.65	.70	.47	.58	.77	.88	.82	.80	.35	.26	.21	.17	.34	.26	.76	.95	.70	.23	.64	.51	.32	.53	.49	.91	.79	.75
10	.43	.72	.82	.50	.65	.83	.94	.78	.72	.48	.60	.74	.38	.48	.67	.70	.84	.63	.45	.46	.60	.55	.47	.64	.91	.73	.67
22	.42	.72	.83	.35	.55	.72	.82	.88	.80	.32	.59	.63	.37	.59	.67	.90	.76	.71	.47	.64	.88	.48	.67	.89	.88	.75	.75
34	.45	.59	.71	.35	.60	.72	.76	.91	.72	.43	.65	.84	.40	.48	.56	.79	.84	.78	.53	.46	.71	.40	.48	.59	.86	.86	.86
46	.52	.65	.70	.38	.28	.44	.88	.82	.80	.31	.46	.58	.25	.57	.50	.95	.76	.70	.33	.60	.74	.50	.58	.64	.75	.94	.70
12	.50	.44	.52	.23	.41	.33	.58	.91	.50	.15	.26	.11	.23	.14	.08	.84	.95	.80	.38	.36	.27	.35	.07	.08	.76	.79	.76
24	.60	.42	.51	.18	.29	.22	.52	.94	.42	.20	.45	.32	.37	.16	.15	.90	.70	.64	.28	.28	.11	.45	.19	.24	.88	.67	.62
36	.27	.42	.49	.28	.46	.50	.78	.84	.73	.26	.33	.37	.34	.34	.38	.84	.73	.70	.30	.41	.34	.18	.36	.27	.76	.91	.71
48	.18	.27	.21	.22	.34	.39	.88	.97	.84	.11	.30	.26	.28	.33	.34	.95	.81	.76	.25	.43	.40	.20	.41	.26	.86	.88	.85

[1]Forms B and C were administered on the post-test and delayed post-test administrations, respectively.

[2]Forms C and A were administered on the post-test and delayed post-test administrations, respectively.

[3]Forms A and B were administered on the post-test and delayed post-test administrations, respectively.

A summarization of these results for Data Set One is as follows:

1. The simultaneous use of all the indices in Table 4.2.5 which are appropriate for Data Set One would result in identifying all items except items 22, 23, 32, 33, 34, 38, 39, and 40 as not having satisfied condition two. A close examination of the results further revealed that the correlation indices were very unstable even for items with similar difficulty indices. Therefore, it was decided to use a somewhat less restrictive statement of condition two. The revised condition two is that items measuring the same objectives should have certain similar statistical indices and other item characteristics deemed desirable such as positive discrimination indices.

   It is suggested that item statistics such as the difficulty index, the index of item precision, the KI coefficient and Popham's (1971) chi-square statistic be used to test for similarity among items. These statistics were chosen as they are not as susceptible to the nuances of criterion-referenced test data as are conventional discrimination indices and/or they provide a measure of how consistent the item scores are for examinees across testing occasions.

   The second part of this restatement of condition two is in essence a reflection of the requirements of

conventional item analysis. That is, even items with similar statistical indices such as item difficulty levels and precision indices, are not acceptable if they have negative discrimination indices.

2. For Data Set One, the restatement of condition two was applied by requiring that the difficulty indices be similar and that the items have positive point biserial coefficients and D statistics. On the basis of these criteria, items 2, 3, 14, 19, 20, 28, 30 and 32 were identified as not having satisfied condition two.

3. Cochran's Q statistic was calculated for the difficulty indices for all items measuring the same objectives. As can be seen from the results presented in Table 4.2.6, seven out of eleven objective subgroups has significant Q statistics. A significant value for the Q statistic is an indication that the items difficulty indices for items measuring the same objective are not the same. Since the Q statistic does not identify which item or items are aberrant, a visual examination of the items was conducted. This examination revealed the existence of clusters of items with similar difficulty indices within some of the objective subgroups with significant Q statistics. The existence of these clusters of items raises the question

of whether or not condition two can ever be realistically applied. For if the same items measuring the same objectives have different statistical indices, how can the test constructor make a decision as to which items are aberrant and which items are not? The problem of clusters of items with similar difficulty indices within objective subgroups will be dealt with in more detail later.

For Data Set Two, the item indices presented in Table 4.2.7 were calculated for each examinee group on both the post-test and delayed post-test scores. Thus, post-instructional information was available on each item from two different groups of examinees. A summary of the results obtained through these calculations is given as follows:

1. The discrimination indices, the point biserial coefficient and the D statistic, for the same items were quite unstable across examinee groups. Further, there were a number of instances within examinee groups where items measuring the same objectives and having similar item difficulty indices had quite different discrimination indices.
2. For each item in each of the three parallel forms, item precision indices and KI coefficients were compiled from the post-test and delayed post-test scores of each examinee group. Items from the same

test form having two item precision indices less than .80 were identified as problem items. These problem items, organized by test form were:

Test Form	Items
A	8, 27, 46
B	12, 24
C	10, 32

3. For items measuring the same objective within an examinee group, Cochran's Q statistic was calculated to determine if the items from the same objective subgroup were similar with regard to difficulty level. These results are presented in Table 4.2.8. For the three examinee groups there were a total of twelve objective subgroups. For only three of the twelve objective subgroups were the Q statistics not significant. This finding is an indication that few of the objective subgroups had items with similar difficulty indices. However, as with items from Data Set One, there were clusters of items within the objective subgroups which had similar difficulty indices.

4. On the basis of the results described in this section, the following items organized by examinee groups were identified as not having satisfied condition two:

TABLE 4.2.8

Q STATISTICS CALCULATED ON DATA SET TWO

Objective	Test Items	Q Statistics/df Examinee Group I	II	III
3	3, 15, 27, 39	11.72/3**	78.18/3**	11.59/3**
8	8, 20, 32, 44	5.02/3	19.31/3**	19.20/3**
10	10, 22, 34, 46	3.51/3	9.21/3*	9.86/3
12	12, 24, 36, 48	33.08/3**	7.07/3*	3.83/3

* $p<.05$

** $p<.01$

Examinee Group	Items
I	3, 10, 12, 20, 24, 32, 36, 38
II	8, 10, 12, 24, 27, 29, 36, 46, 48
III	8, 12, 15, 24, 27, 46, 48

5. To reduce the importance of examinee group on the results, only those items identified as aberrant for at least two out of three examinee groups were considered to be problem items. Thus, for Data Set Two, only items 8, 10, 12, 24, 36, 46, and 48 were identified as not having satisfied condition two.

4.2.3 Condition Three

The analyses for determining whether or not items satisfied condition three were conducted in two phases. In the first phase, a Monte Carlo study was conducted in order to determine whether data reduction techniques such as factor analysis should be used to analyze dichotomized test item data with minimal item variance. In phase two, actual test data was examined to determine whether or not the items from Data Sets One and Two satisfied condition three.

For phase one, a Fortran IV program was written to generate data according to the factor model presented in Section 3.3.3. The model is as follows:

$$y_i = \lambda_{i1} x_1 + \lambda_{i2} x_2 + \ldots + \lambda_{ik} + e_i,$$

$$(i = 1, ., p \quad \text{and} \quad j = 1, \ldots k)$$

where λ_{ij} is the loading of the ith variable on the jth factor and e is the random error component for each y_i. In matrix form, this model is given as

$$\underline{y} = \Lambda\underline{x} + \underline{e},$$

where $\underline{y}$ is the (pxl) vector of observed variables,

$\underline{x}$ is the (kxl) vector of factors

Λ is the (pxk) matrix of factor loadings

and $\underline{e}$ is the (pxl) vector of unique scores or error scores.

The following parameters were set in advance:

n, the number of examinees = 200,

p, the number of items = 12,

k, the number of factors = 3.

The matrix of factor loadings was set in advance to have simple structure and of the following form:

$$\Lambda = \begin{bmatrix} .60 & .00 & .00 \\ .60 & .00 & .00 \\ .60 & .00 & .00 \\ .60 & .00 & .00 \\ .00 & .70 & .00 \\ .00 & .70 & .00 \\ .00 & .70 & .00 \\ .00 & .70 & .00 \\ .00 & .00 & .80 \\ .00 & .00 & .80 \\ .00 & .00 & .80 \\ .00 & .00 & .80 \end{bmatrix}.$$

The values for x, the factor scores and e, the error scores were generated through the use of a random number generator. The factor scores were generated to be distributed normally with a mean equal to zero and variance equal to one. The error scores were generated to be distributed normally with mean zero and with a specified variance.

Once the (nxp) matrix of observed scores (y's) were obtained, three additional data sets were obtained by comparing the y values for each examinee against three cutoff scores. If the y value was greater than or equal to a cutoff score, a new y value, y', was set to be 1 otherwise the new value was set to be 0. In this manner, three strings of 1's and 0's were generated for each examinee. The y' values were used to simulate examinee scores on a set of p items. The cutoff scores were set so as to provide data sets with different levels of variability.

A (p x p) variance-covariance matrix for the items in each of the four data sets was obtained. These four matrices were then factor analyzed using maximum likelihood procedures. A Fortran IV program, Acovs (Joreskog, 1970), was used for the analyses. The results of these four analyses are presented in Table 4.2.9.

The purpose of these factor analyses were to determine if the simple structure of the nondichotomized data set could be reproduced from the variance-covariance matrices of the dichotomized data sets. This purpose was

TABLE 4.2.9

FACTOR PATTERN MATRICES FOR FOUR SETS

OF SIMULATED TEST DATA

Test Item	Non-Dichotomized Data			Dichotomized Data Sets								
				Data Set One			Data Set Two			Data Set Three		
	Factors			Factors			Factors			Factors		
	I	II	III	I	II	III	I	II	III	I	II	III
1	.62	.00	.00	.29	.00	.00	.19	.00	.00	.09	.00	.00
2	.59	.04	-.14	.24	-.02	-.06	.11	.03	-.04	.34	-.07	.00
3	.58	-.03	-.05	.16	-.02	-.01	.19	-.03	-.01	.10	-.02	.02
4	.57	.05	-.11	.18	.01	-.03	.33	-.08	-.04	.08	.01	.02
5	.00	.69	-.05	.01	.29	.00	-.01	.32	-.04	.06	.23	-.02
6	.18	.72	-.02	.12	.32	-.03	.04	.21	.00	.06	.22	-.02
7	.01	.68	.07	.01	.26	.03	.02	.27	.05	.05	.14	.05
8	.09	.65	-.05	.02	.28	-.03	.09	.18	-.05	.01	.22	.00
9	.07	-.01	.85	-.03	.01	.38	.03	.02	.28	-.02	.01	.28
10	.01	-.03	.72	-.06	-.05	.27	.02	-.01	.21	-.04	.00	.21
11	.10	.01	.74	-.01	-.01	.33	.00	-.02	.23	.02	.01	.19
12	.00	-.02	.85	.00	.03	.30	.00	.02	.36	.00	.01	.26
	$\chi^2(33) = 39.08$			$\chi^2(33) = 39.53$			$\chi^2(33) = 39.73$			$\chi^2(33) = 38.35$		

accomplished through the use of confirmatory factor analysis in which the factor patterns are hypothesized in advance. For this study the following factor pattern was hypothesized:

$$\Lambda = \begin{bmatrix} X & 0 & 0 \\ X & 0 & 0 \\ X & 0 & 0 \\ X & 0 & 0 \\ 0 & X & 0 \\ 0 & X & 0 \\ 0 & X & 0 \\ 0 & X & 0 \\ 0 & 0 & X \\ 0 & 0 & X \\ 0 & 0 & X \\ 0 & 0 & X \end{bmatrix}$$

where "X" denotes parameters to be estimated and the "0" denotes parameters which are fixed as zero.

The null hypothesis for these analyses was that the variance-covariance matrix of the dichotomized data sets depends upon three factors with the loading matrix Λ having the above pattern. The chi-square tests for the three analyses were not significant indicating that the null hypotheses were not rejected.

While the results show that factor analysis can be used to reproduce the data structure even with dichotomized data, the size of the factor loadings became progressively smaller as the test variability became smaller.

Since for the purposes of this study, the use envisioned for factor analysis is the determination of whether or not an item is a measure of more than one objective, the results of the Monte Carlo study support such a use even for test data with minimal variance.

In the second phase, the task was to determine whether the items from Data Sets One and Two satisfied condition three. This task was accomplished through the use of factor analysis. For Data Set One, this analysis was conducted in two steps. In the first step an exploratory factor analysis was conducted on the interitem correlations of the 40 test items. The number of factors was set equal to the number of objectives in Data Set One. For the exploratory analysis, the principal factor analysis model of the SPSS computer programs package was used to obtain estimtes of the factor loadings and uniquenesses. The initial solution was rotated using the varimax rotation method. Table 4.2.10 contains the rotated matrix of factor loadings and uniqueness estimates. In the second step, a confirmatory analysis was planned using the results obtained in the exploratory analysis to establish the hypothesized structure of the factor pattern matrix and initial values for the parameters to be estimated. Since, the

TABLE 4.2.10

FACTOR PATTERN MATRIX FOR TEST

ITEMS FOR DATA SET ONE

Test Item	Factors										
	1	2	3	4	5	6	7	8	9	10	11
1	.15	.16	.58	.04	.04	.22	-.05	.05	.11	-.06	-.06
2	.06	.18	.05	-.08	.07	.27	.10	.02	-.02	-.06	-.09
3	.29	.02	.49	.17	.10	-.05	.05	-.12	.14	-.09	.06
4	.47	.10	.28	-.01	.09	.15	.05	.05	-.02	.01	.10
5	.14	.11	.66	.02	.10	.11	.12	.11	.02	.11	.06
6	.04	.18	.40	.11	.15	.04	.09	-.02	.02	-.02	-.05
7	.36	.17	.41	.07	-.00	.20	-.17	.15	-.07	-.13	.05
8	.26	.08	.28	-.01	.07	.04	.19	.05	-.07	.14	.08
9	.21	.13	.18	-.01	.09	.47	.00	.12	-.06	-.02	.16
10	.12	.02	.27	.13	.08	.34	.05	.03	.10	.05	-.06
11	.68	.05	.14	.00	.09	.05	-.02	.09	-.04	-.01	.04
12	.68	.05	.19	.08	.19	.10	.13	.03	.14	.01	-.23
13	.68	.05	.22	.02	.04	.04	.03	.03	.03	-.07	.09
14	.10	.39	.28	.43	.16	-.04	.17	.12	.06	.06	.12
15	.42	.27	.13	.18	.18	.18	-.06	.20	.11	-.02	.31
16	.53	.04	.11	.14	.08	.14	.15	.10	-.06	.25	-.11
17	.48	.09	.24	-.01	.16	-.08	.19	.01	.02	.02	.24
18	.38	.13	.16	.15	.11	.12	.04	.49	.11	-.11	.02
19	.00	.04	.00	.01	-.01	-.02	-.07	-.03	.02	.53	.01
20	.14	.19	.23	.03	.16	-.01	.22	.09	.03	.05	.09
21	.27	.16	.16	.57	.17	.02	.12	.02	-.07	.04	.02
22	.50	.11	.07	.17	.16	.31	.10	.33	.12	.10	.06
23	.47	.18	.12	.36	.20	.26	.09	.11	-.01	-.04	-.14
24	.77	.09	.02	.11	.16	.08	.10	-.01	.05	-.12	-.03
25	.54	.10	.00	.19	.11	.12	.17	.01	.14	.02	.03
26	.21	.10	.11	.18	.20	.14	.63	-.06	.08	-.07	-.05
27	.34	.27	.07	.04	.23	.39	.18	-.12	.09	.01	.01
28	.20	.29	.03	.04	.17	.06	.46	.16	.02	-.18	.06
29	.17	.17	.23	-.02	.18	-.03	.06	.17	.43	.04	-.02
30	.57	.10	.03	-.04	.22	.08	.13	-.05	.37	.10	.15
31	.26	.17	.05	.07	.50	.26	.16	-.07	.22	-.03	.08
32	.13	.40	.24	.02	.25	.16	.18	-.06	.20	.01	-.12
33	.14	.29	.23	.23	.31	.21	.12	-.13	.27	.24	.09
34	.21	.11	.21	.16	.49	.24	.18	-.15	.09	.05	.18
35	.22	.32	.14	.12	.72	.06	.13	.13	.02	.02	-.06
36	.35	.27	.16	.12	.69	.07	.12	.11	.05	-.06	.02
37	-.03	.15	-.00	.35	.34	-.05	.06	.14	.14	-.05	-.05
38	.08	.66	.18	.15	.27	.10	-.01	.02	-.12	.10	-.03
39	.14	.60	.12	.18	.24	.07	.05	.01	.28	-.03	.02
40	.07	.67	.11	.05	.11	.14	.16	.09	.07	.02	.07

computer program used for the maximum likelihood factor analysis was limited to analyzing a maximum twenty-one variables, the items from Data Set One were divided into two groups of 21 and 19 items corresponding to objectives one through four and five through eleven, respectively. However, the computer program was unable to accomplish the analysis for either group of items. This inability to reach a solution was a result of either poor initial estimates or the nature of the data being analyzed. The consequence of these findings was that the confirmatory analysis step was not completed.

Therefore, for Data Set One, condition three was assessed through the use of information obtained from the exploratory analysis. On the basis of this information, items 1, 4, 14, 21, 22, 23, 24, 25 and 30 were identified as not having satisfied condition three.

For Data Set Two, the interitem correlations were found to be extremely low. While the use of data reduction procedures such as factor analysis was shown to be useful in analyzing dichotomous data, the results of such analyses are meaningful only if there is some underlying structure for the data. The lack of consistent patterns for the interitem correlations within and across objective subgroups was taken as an indication that data did not have a meaningful underlying structure. Consequently, a decision was made to use a visual analysis of the interitem phi coefficient as a means of assessing condition three for the items of Data Set Two.

On the basis of a visual analysis of these interitem correlations, the following items organized by examinee group, were identified as not having satisfied condition three:

Examinee Group	Items
I	3, 22, 34, 44, 46
II	3, 15, 20, 32, 39, 44, 46
III	8, 15, 44, 46

As with the previous analyses for conditions one and two, only those items identified as aberrant from the data of at least two out of three examinee groups were considered to be problem items. These problem items were items, 3, 15, 44, and 46.

4.2.4 Summary of the Results of the Use of Empirical Analyses as a Means for Assessing Item Validity

The use of empirical analyses as a means for assessing item validity resulted in the identification of a number of items for Data Sets One and Two which did not satisfy one or more of the three empirical conditions. The fact that in most instances the application of each of the three empirical conditions resulted in the identification of different items as aberrant is an indication that these conditions are sensitive to different aspects of the items' statistical characteristics. Thus, the application of these conditions does not result in redundant information.

Also, while the empirical analyses performed the important function of identifying problem items, such analyses do not lead to definitive answers concerning the specific problem(s) with the items. Besides the possibility that an item is not from the domain of items measuring an objective, other possible explanations include the following:

1. Items measuring the same objective require different levels of knowledge from the examinees.
2. Items are poorly written or have a poor set of distractors.
3. Information required for a correct response to an item was presented to the examinee either poorly or not at all.

In order to identify specific examples where the empirical identification of an aberrant item might have been due to one of the causes above, a visual analysis of the items was conducted. Of particular interest were the items which formed clusters based on similar difficulty levels within their intended objective subgroups. These clusters pose a dilemma in that items within some clusters satisfied all three empirical conditions while analyses conducted on these items and the other items within the objective subgroup would result in the identification of items which failed to satisfy condition two. An excellent example of these clusters is obtained from Data Set One. For objective four there are eleven items which can be grouped into four clusters. Cluster one consists of items

11, 12, 13, 16 and 17. Cluster two consists of items 15, 18, and 21. Cluster three consists of items 14, and 20 and cluster four has only item 19. Of all the items, only item 19 (cluster four) has serious statistical deficiencies. Further, its item form, distractors and content are quite different from the other items enabling one to include that it is not a measure of the intended objective. Item 14 has the same item form and type of distractors as the items in cluster one and yet had a significantly lower difficulty index. Thus, a probable cause for the different difficulty index is that the information required for a correct response was not available to a segment of the examinee group (cause 3) not that the item is not a measure of the intended objective. The items in clusters one and two have different item forms and type of distractors. Also, the items in cluster two appear to require more general knowledge than do the items in cluster one. Since an examination of the content of items from both clusters would lead to a conclusion that these items were measures of the intended objective, the lower difficulty indices are most likely a function of the different knowledge requirement (cause 1).

Finally, while all the items selected for analysis for Data Set Two conformed to item form guidelines, eleven out of the sixteen items failed to satisfy one or more of the empirical conditions, although, no items from this set failed to satisfy all three conditions. Through the use

of two examples from Data Set Two, an attempt will be made to explain why certain items were identified as aberrant. The first example involves objective 3 and item 27 of test form C. While the item form is exactly the same as other items within the objective subtest, the item stimulus was more difficult and the distractors included "none of these." The majority of the respondents from all three examinee groups chose that incorrect response. For the same item on test form B, the stimulus was also more difficult. Therefore, for two out of the three examinee groups the item did not satisfy condition one and was identified as aberrant. The second example involves objective 12. Using the empirical approach to item validity would result in identifying three out of four items as not being valid measures of the objective. However, the summary statistics (reported in Table 4.2.2) for the objective subtests shows that this objective had the lowest mean post-test score for two out of the three examinee groups and for the other examinee group it was the second lowest score. On the basis of this information a plausible explanation for poor examinee performance is poor instructional material rather than the items were not measures of the objectives.

If the use of empirical analyses cannot be used to ensure item validity, then what role should these analyses play? It is the contention of this author that empirical analyses can be used to identify both item and objective deficiencies. While examples have been given

for item deficiencies, none have been given for objective deficiencies. One such example involves objectives 2, 9, and 11 of Data Set One. These objectives on the surface appear to be different in that they require different tasks from the examinees. However, in order to respond correctly to the items from either objective the examinee must process the same information. Consequently, it is not surprising to note that the items from these objectives "load" on the same factor.

In summary, while empirical analyses cannot be used to ensure item validity, the information provided by this approach is essential for identifying problems in objective definition, item construction including choice of distractors and instructional effects.

4.3 The Use of the Judgments of Content Specialists in the Assessment of Item Validity

In this section, two studies used to collect the judgments of content specialists on the items from Data Sets One and Two will be described. In Study One, twenty-one science teachers were administered an item validation questionnaire which was designed to determine the extent to which they thought the items were measures of the intended objectives (see Appendix A for a copy of materials given to the teachers). The teachers (or content specialists as we will refer to them) were asked to make judgments on forty items and eleven objectives from Data Set One using the Hemphill-Westie categorizing technique.

In Study Two, a more complex item validation questionnaire was used to obtain the judgments of content specialists on the forty-eight items and twelve objectives from Data Set Two (see Appendix B for a copy of the materials given to the content specialists). For the study, the twelve instructional objectives and their matched items (Test Form A) given in Appendix B were subdivided into three subgroups. Each of these subgroups consisted first of four objectives and their four corresponding items for a total of 16 test items. Next, two additional objectives from the initial pool of twelve objectives, without their corresponding items, were assigned to each subgroup resulting in a final subgroup composition of six objectives and sixteen items. Finally, three forms of an item validation questionnaire were formed by assigning each of the subgroups of items and objectives to one of three judgmental procedures, the Hemphill-Westie categorizing technique, the semantic differential rating technique and the matching technique. All three judgmental procedures were described in Section 3.3. The format of each questionnaire is given as follows:

	Judgmental Procedure		
Questionnaire	Categorizing	Rating	Matching
1	Subgroup One	Subgroup Two	Subgroup Three
2	Subgroup Two	Subgroup Three	Subgroup One
3	Subgroup Three	Subgroup One	Subgroup Two

Ten copies of each form of the questionnaire were randomly assigned to thirty science teachers (not the same teachers from Study One). Thus, for any one subgroup of objectives and items, there is information available from three different groups of content specialists using three different judgmental procedures. The data collected from both studies were examined, where appropriate, with regard to the following questions:

1. Does the data provide information which can be used to assess the extent to which an item is a measure of an instructional objective?
2. Is the information obtained reliable in the sense that there was consitency of agreement amongst the content specialists?
3. Is the data valid?

Both studies will be examined within the context of the judgmental procedures. Since, in Study One only the Hemphill-Westie procedure was used, the discussion of the data collected for that study will be limited.

4.3.1 The Hemphill-Westie Categorizing Procedure

For both Studies One and Two, a decision was made to set the cutoff score for the index of item-objective congruence, the numerical representation of the Hemphill-Westie data, to be .70. That is, items having item-objective congruence indices less than .70 were identified as not being valid measures of their intended objectives.

The results of the calculation of these indices are presented in Tables 4.3.1 and 4.3.2. In Study One, items 3, 4, 7, 8, 9, 10, 15, 18, 19, 20, 26, 31 and 34 from Data Set One were identified as not being valid measures of their intended objectives. In Study Two, items 8, 10, 13, 14, 16, 22, 23, 24, 35, 40 and 41 from Data Set Two were identified as not being valid measures of their intended objectives.

The Hemphill-Westie procedure requires that the content specialists judge each item against all the objectives. If an item is judged to be a measure of more than one objective, its item-objective congruence index will be lowered. For both studies, the item-objective congruence indices were always considerably higher when the items were assessed on the intended objectives than when they were assessed on the other objectives. It would thus appear that the content specialists can make meaningful judgments in the assessment of item validity.

The next analyses were concerned with determining whether or not acceptable levels of item-objective congruence indices were obtained from reliable data. That is, were the content specialists consistent in their judgments? The assessment of the consistency of agreement amongst judges was made by calculating Lu's (1971) coefficient of agreement discussed in Section 3.3. A coefficient of agreement was obtained for each objective

TABLE 4.3.1

VALUES FOR THE INDEX OF ITEM OBJECTIVE CONGRUENCE ON TEST ITEMS IN DATA SET ONE

Test Item	Objectives										
	1	2	3	4	5	6	7	8	9	10	11
1	.80										
2	.70										
3		.57									
4		.61									
5			.77								
6			.77								
7		.56									
8			.50								
9		.50									
10			.63								
11				.93							
12				.93							
13				.93							
14				.91							
15				.50							
16				.93							
17				.95							
18				.54							
19				.35							
20				.21							
21				.85							
22					.82						
23					.80						
24						.92					
25						.92					
26							.62				
27							.89				
28							.73				
29								.81			
30								.85			
31								.38			
32									.82		
33									.72		
34									.59		
35										.94	
36										.94	
37										.82	
38											.87
39											.82
40											.82

TABLE 4.3.2

VALUES FOR THE INDEX OF ITEM-OBJECTIVE CONGRUENCE AND THE SD STATISTIC FOR DATA SET TWO

(Index/SD Statistic)

Objective Subgroup	Test Item	1	2	3	4	5	6	7	8	9	10	11	12
B	1	.81/.69											
	13	.62/.46											
	25	.83/.79											
	37	.72/.78											
A	2		.82/.57										
	14		.50/.47										
	26		.82/.74										
	38		.84/.81										
B	3			.90/.50									
	15			.98/.50									
	27			.92/.82									
	39			.86/.55									
C	4				.83/.61								
	16				.40/.32								
	28				.37/.40								
	40				.60/.30								
C	5					.96/.78							
	17					.85/.59							
	29					.95/.57							
	41					.63/.41							

TABLE 4.3.2 (continued)

Objective Subgroup	Test Item	1	2	3	4	5	6	7	8	9	10	11	12
C	6						.75/.54						
	18						.78/.51						
	30						.80/.49						
	42						.84/.36						
A	7							.77/.63					
	19							.76/.61					
	31							.78/.70					
	43							.78/.68					
A	8								.47/.49				
	20								.75/.61				
	32								.71/.59				
	44								.84/.77				
A	9									.90/.80			
	21									.85/.76			
	33									.83/.74			
	45									.86/.68			
B	10										.62/.43		
	22										.62/.42		
	34										.74/.69		
	46										.80/.54		
C	11											.83/.41	
	23											.50/.42	
	35											.27/.35	
	47											.92/.51	
B	12												.70/.56
	24												.68/.54
	36												.88/.66
	48												.88/.59

subgroup for both Data Sets One and Two. The results are presented in Tables 4.3.3 and 4.3.4. For all twenty-three objectives, the coefficient of agreements were significant. These findings lead to an interpretation that the Hemphill-Westie judgmental data was reliable in the sense that there was consistency of agreement amongst the judges.

For the purposes of this study, validity of the judgmental data is defined as the degree of agreement between different groups of content specialists assessing item validity through the use of different judgmental procedures. For Study One, no estimates of validity were obtained. For Study Two, the degree of agreement was obtained by correlating two rank orderings of the items based on the sizes of judgmental statistics calculated from the categorizing and rating procedures. The first rank ordering of the items was established by using values of the index of item objective congruence. The second rank ordering was established by using values of a statistic (SD) calculated from the semantic differential ratings on the items. This statistic was computed using the following algorithm:

a. Compute the sum (y_1) of the ratings for each item, on the objective to which it was matched, across content specialists.

TABLE 4.3.3

LU'S COEFFICIENT OF AGREEMENT FOR THE OBJECTIVE SUBGROUPS OF DATA SET ONE

Objective	Lu's Coefficient	χ^2 statistic (df)
1	.83	χ^2 (819) = .16*
2	.86	χ^2 (819) = .13*
3	.90	χ^2 (819) = .08*
4	.91	χ^2 (819) = .07*
5	.88	χ^2 (819) = .10*
6	.90	χ^2 (819) = .07*
7	.91	χ^2 (819) = .08*
8	.94	χ^2 (819) = .02*
9	.89	χ^2 (819) = .09*
10	.88	χ^2 (819) = .11*
11	.91	χ^2 (819) = .08*

*$p < .01$

TABLE 4.3.4

LU'S COEFFICIENT OF AGREEMENT FOR THE

OBJECTIVE SUBGROUPS OF DATA SET TWO

Objective	Lu's Coefficient	χ^2 statistic (df)
1	.80	χ^2 (112) = .20*
2	.83	χ^2 (128) = .16*
3	.67	χ^2 (112) = .33*
4	.57	χ^2 (128) = .41*
5	.86	χ^2 (128) = .14*
6	.75	χ^2 (128) = .25*
7	.88	χ^2 (128) = .11*
8	.74	χ^2 (128) = .26*
9	.83	χ^2 (128) = .16*
10	.88	χ^2 (112) = .13*
11	.83	χ^2 (128) = .16*
12	.83	χ^2 (112) = .16*

*$p<.01$

b. Compute the sum (y_2) of the ratings for each item on the remaining objectives across content specialists.

c. Compute the rank order statistic (SD) from the ratio of sum one (y_1) to sum two (y_2). For a rating scale having values from one to k, this statistic (SD) has a maximum value given as

$$\max (SD) = \frac{nk}{n(N-1) \quad (k-(k-1)} \text{ or } \frac{nk}{n(N-1)} = \frac{k}{N-1}$$

The minimum value for SD is given as

$$\min (SD) = \frac{n}{n(N-1)k} = \frac{1}{(N-1)k}$$

where n is the number of content specialists,

N is the number of objectives,

and k is the highest value of the rating scale.

For Study Two, with six objectives per judgmental subgroup, the maximum value for the SD statistic is 1 and the minimum value is .04.

For each of the three subgroups of objectives, consisting of 16 items each, Spearman's coefficient of rank difference was calculated between the item-objective congruence indices and the item SD statistics. The three Spearman coefficients reported in Table 4.3.5 were statistically significant and above .65, suggesting the substantial agreement as to the quality of test items across the two methods for judging items.

TABLE 4.3.5

RANK ORDER CORRELATIONS OF ITEM OBJECTIVE CONGRUENCE INDICES AND THE SD STATISTIC FOR DATA SET TWO

Objective Group	Test Items	Rank Difference Correlation	Statistic
A	2, 7, 8, 9, 14, 19, 20, 21, 26, 31, 32, 33, 38, 43, 44, 45	.82	5.31*
B	1, 3, 10, 12, 13, 15, 22, 24, 25, 27, 34, 36, 37, 39, 46, 48	.66	3.30*
C	4, 5, 6, 11, 16, 17, 18, 23, 28, 29, 30, 35, 40, 41, 42, 47	.67	3.38*

*$p < .01$

4.3.2 The Semantic Differential Rating Procedure

For Study Two, the second judgmental procedure required that the content specialists assign a semantic differential like rating of from one to five to an item depending on whether the item was judged as an irrelevant or relevant measure of the objective in question. The fact that the content specialists consistently rated items higher on the intended objectives than on the other objectives was taken as an indication that this data did provide meaningful information for assessing item validity. However, one problem associated with the use of these ratings is that they do not provide information on whether or not the items were judged to be a measure of more than one objective. Therefore, the SD statistics discussed previously were computed for the items as it takes into consideration the ratings assigned to the item for the other objectives. It was arbitrarily decided that items having SD values less than .50 would be identified as not being valid measures of the objectives to which they were matched. For Data Set Two, items 2, 8, 10, 13, 14, 16, 22, 23, 35, and 40 were identified as invalid.

As assessment of the reliability of these ratings was made through an examination of the standard deviations of the ratings of an item and the objective to which it was matched. With the exception of a few items these standard deviations were quite small which was an indication that the content specialists were making the same

ratings on the item. These results are presented in Table 4.3.6.

4.3.3 The Matching Procedure

For the matching technique the content specialists were asked to match each item to the objective they felt it measured. The data collected from the use of this technique is different from the data collected from the use of the other two techniques in that the content specialists were not required to judge each item on all the objectives.

An (m x n) contingency table of items (m) and objectives (n) was constructed. The mn cell frequencies consisted of the number of times content specialists matched an item to an objective. Discrepancies between the expected matches and the actual matches were used to identify invalid items. A minimum criterion that seventy percent of the content specialists must have correctly matched an item to an objective before the item could be declared valid was established. Using this criterion the results presented in Table 4.3.7 show that for Data Set Two, items 8, 25, 28, 35, 41, and 47 were identified as not having item validity. The relatively high number of correct matches is an indication that this information can be used to assess item validity.

One means for assessing the reliability of the data collected through the use of a matching technique is to calculate the amount of agreement between the expected

TABLE 4.3.6

SEMANTIC DIFFERENTIAL RATINGS ON THE

TEST ITEMS FROM DATA SET TWO

Test Item	SD Rating Coeff.	Mean	Standard Deviation	Test Item	SD Rating Coeff.	Mean	Standard Deviation
1	.69	4.8	.46	25	.79	4.3	.95
2	.57	4.7	.45	26	.74	4.6	.48
3	.50	4.2	.70	27	.82	4.7	.48
4	.61	4.6	.52	28	.40	5.0	.00
5	.78	5.0	.00	29	.57	4.4	.95
6	.54	4.9	.31	30	.49	4.8	.32
7	.63	5.0	.00	31	.70	5.0	.00
8	.49	4.2	1.21	32	.59	4.7	.45
9	.80	5.0	.00	33	.74	4.6	.48
10	.43	4.0	.78	34	.69	4.5	.53
11	.41	5.0	.00	35	.35	4.7	.48
12	.56	4.7	.48	36	.66	5.0	.00
13	.46	4.1	.80	37	.78	4.6	.52
14	.47	4.2	.75	38	.81	5.0	.00
15	.50	4.2	.55	39	.55	4.7	.48
16	.32	4.9	.32	40	.30	4.7	.48
17	.59	4.2	.70	41	.41	3.9	.88
18	.51	4.7	.48	42	.36	3.5	1.27
19	.61	4.7	.45	43	.68	4.7	.45
20	.61	4.5	.50	44	.77	4.8	.40
21	.76	4.8	.40	45	.68	4.7	.45
22	.42	5.0	.60	46	.54	4.8	.40
23	.42	4.8	.40	47	.51	4.0	.82
24	.54	5.0	.00	48	.59	4.9	.31

TABLE 4.3.7

CONTINGENCY TABLES FOR DATA COLLECTED FROM THE CONTENT SPECIALISTS

IN THE TEST ITEMS TO THE OBJECTIVES IN DATA SET TWO

Objective Subgroup A						
Test Item	Objective 1	2	8	7	12	9
9						10
19		1			9	
32	1		9			
26		10				
7		1	1	8		
44		2	8			
33	2					8
14		8	2			
31				10		
45						10
8	4		6			
2		10				
20	1		9			
21						10
38		9	1			
43				10		

Objective Subgroup B						
Test Item	Objective 12	7	4	1	3	10
34					1	9
3					10	
48	10					
13	2			8		
12	10					
46						10
37				10		
15					10	
24	10					
25	5			5		
1	2			8		
27					10	
22						10
39					10	
36	10					
10						10

Objective Subgroup C						
Test Item	Objective 6	11	2	8	5	4
11		8				
42	6		2			
29					8	
18	8					
28		3				5
47	2	4				2
16		1				7
6	8					
35		5				3
17					8	
40		1				7
4	1					7
5					8	
23		7				1
41				4	4	
30	7		1			

matches and the actual standard. Light (1971) has developed a statistic (G) which provides a numerical representation of this amount of agreement which can be tested statistically for significance. However, because of the relatively small number of judgments required of the content specialists, it was not calculated for this data.

The data collected using the matching technique did not lend itself to the assessment of validity as defined in this study. Therefore, no determination of the validity of this data was made.

4.3.4 Summary of the Use of the Judgments of Content Specialists to Assess Item Validity

In Section 4.3, three techniques for collecting and analyzing the judgments of content specialists as a means for assessing item validity were discussed. All three techniques were shown to provide information which can be used to ascertain if an item is a measure of an objective. However, there were differences in the types of data which were collected through the use of these techniques. For example, there were many more low SD statistics than low item-objective congruence indices for the same items. This is an indication that the content specialists when using the semantic differential

rating procedure judged the items to be relevant measures of objectives other than the intended ones more often than when using the categorizing procedure. It appears that these two procedures are tapping different dimensions.

Given the task of judging which items are measures of intended objectives, the Hemphill-Westie procedure is recommended over the other two techniques. Two statements are offered in support of this recommendation. One, the numeric representation of the data, the index of item objective congruence, provides a meaningful interpretation of the extent to which an item is judged to be a valid measure of the intended objective. Two, there are means for determining the reliability and validity of the data collected. Further, these methods can be tested for significance.

There are drawbacks to the use of the Hemphill-Westie procedure which could be rectified through the use of other judgmental techniques. These drawbacks are given as follows:

1. It cannot be used to collect information on such topics as quality of the item, and type of distractors.
2. The dimensionality of the data must be known in advance of its use.

3. Its administration is quite time consuming particularly if the numbers of items and of objectives are large.

Thus, before selecting the type of judgmental procedure to use, the test constructor should take into consideration the information desired and the resources available and then choose the most appropriate procedure.

4.4 Conclusions in the Examination of Two Approaches Used in the Assessment of Item Validity

In this chapter two approaches, empirical analyses and the judgments of content specialists, were used to determine whether individual items in two sets of items were valid measures of their intended objectives. It was determined that empirical analyses could not be used to establish item validity in and of themselves. On the other hand, on the basis of information presented it was concluded that the judgment of content specialists could not be used to assess item validity without other forms of verification either. At this point an attempt will be made to integrate and summarize the results obtained from the use of both of these methods.

An examination of the results of rejected items (see Table 4.4.1) shows that the two approaches did not consistently identify the same items. For example for

Data Set One, only seven out of a total of twenty-three invalid items were identified by both methods. While for Data Set Two, only one out of eleven invalid items was identified by both methods.

It would appear that these two methods for assessing item validity are based on different criteria and therefore provide different information on the items. If this is the case, then one would have to consider the use of both procedures when assessing item validity. In order to examine this possibility a review of the items identified as invalid was conducted. This review revealed the following:

1. Items identified as valid by both procedures could be placed into two groups. In the first group, the items had the same content form and response mode both within and across objective subtests. In the second group, items within the same objective subtest were dissimilar in content and appeared to require different levels of knowledge for correct responses. With regard to the items from the first group, the low judgmental values were caused by the content specialists rating items of similar content as being measures of more than one objective. There was no systematic cause for

TABLE 4.4.1

SUMMARY OF REJECTED TEST ITEMS

FOR DATA SETS ONE AND TWO

Data Set	Analysis	Rejected Test Items		
One	Empirical	Condition 1:	2, 3, 4, 6, 7, 10, 11, 14, 19, 39	
		Condition 2:	2, 3, 6, 9, 17, 19, 20, 28, 30, 32	
		Condition 3:	1, 4, 14, 21, 22, 23, 24, 25, 30	
	Judgemental	3, 4, 7, 8, 9, 10, 15, 18, 19, 20, 26, 31, 34		
Two	Empirical		Examinee Group	
			I	II	III
		Condition 1:	3,27,32, 36,39, 48	8,12,20, 22,24,27, 36,38	20,27, 26,48
		Condition 2:	3,20,32, 36,48	8,12,24, 27,36, 39,48	24,36, 48
		Condition 3:	3,22,34, 44,46	3,15,20, 32,34, 44,46	3,15, 44,46
	Judgemental	Categorizing:	8, 10, 13, 14, 16, 22, 23, 24, 35, 40, 41		
		Rating:	2, 8, 10, 13, 14, 16, 22, 24, 35, 40		
		Matching:	8, 25, 28, 35, 41, 47		

the rejection of these items through empirical analyses. With regard to the second group of items, the low judgmental ratings were caused by the content specialists rating items as not being measures of the intended objectives. As with the first group of items, there was not systematic empirical cause for their rejection.

2. Items identified as invalid only by the use of the judgments of content specialists were different in content from other items within the same objective subtest and were therefore judged not to be measures of their intended objectives.

From these results and other findings in Chapter IV, two conclusions have been reached. One, the rejection of the same item by both methods for assessing item validity is not due to the use of the same item information but to compound deficiencies in the item to which both methods were sensitive. Two, while the use of judgments of content specialists is in and of itself an effective means for assessing which items are measures of their intended objectives, this approach is insensitive to item deficiencies due either to poor construction procedures or an interaction with type or level of instruction. Thus, the information provided by this approach must be augmented by information obtained from the use of empirical analyses.

CHAPTER V

CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

5.1 Introduction

In the first four chapters, the importance of establishing item validity for criterion-referenced tests was delineated; proposed solutions were offered; and the results of empirical investigations were discussed. The pruposes of this final chapter are to interpret the empirical results, to discuss the limitations of the study, to suggest future research, and to propose models for developing valid criterion-referenced tests. On this last point, these test construction models will utilize the approaches to assessing item validity discussed in the study.

5.2 Interpretation of the Empirical Results

Empirical data were collected on two of the three proposed approaches to assessing item validity: the use of empirical analyses and the judgments of content specialists. With regard to the use of empirical analyses, there were a number of important findings obtained from the process of trying to ascertain whether or not items

from two distinct data sets were valid measures of their intended objectives.

First, two of the empirical conditions, two and three, as initially set forth did not provide useful data and accordingly were modified. The revised condition two is as follows: items which measure the same objective should have certain similar statistical indices and other desirable characteristics such as being "positively discriminating."

It was suggested that the indices used to assess item similarity should not be correlational in nature (indices based on the least-squares loss function) but instead should be indices which are concerned with examinee performance across testing occasions. Examples of such indices are the KI coefficient (Sabers and Kania, 1972) and Popham's (1971) median chi-square test.

Condition three was modified to take into consideration that it is unrealistic to assume that criterion-referenced test data will have simple structure, as typically such tests are designed to measure examinee performance on a set of objectives that are related in a hierarchical fashion.

Therefore, condition three was revised to read as follows: items measuring the same objective should correlate at least as high with each other as with items measuring other objectives.

A few words of caution are in order concerning condition three. If the test constructor discovers items which correlate highly even though they measure different objectives, the instructional objectives should be reviewed to ensure that the items are tapping unique aspects of the instructional domain and not just assessing the same information in a different manner. The second finding, shown through a study employing artificial data, indicates that the data reduction techniques such as factor analysis can be used to confirm the dimensional structure of data having minimal variance. However, the extremely low interitem correlations for such data is an indication that the results of an exploratory analyses in which the test constructor does not have a strong feeling for the underlying dimensional structure, should be interpreted with great caution. This call for caution is based on the view that one cannot be certain whether the low interitem correlations are a function of minimal variance or the lack of an underlying structure for the data.

With regard to the use of the judgments of content specialists for assessing item validity, there were also a number of important findings. First, it was shown that content specialists can provide meaningful, reliable, and valid assessments of item validity.

Second, if the task at hand is simply the determination of whether or not items are measures of their intended objectives and the dimensional structure of the

item-objective relationships is known in advance, then the Hemphill-Westie categorizing procedure with the numeric representational formula developed in this study was shown to have certain advantages over the other techniques. However, if any of the following conditions hold then one should probably employ a different judgmental technique:

1. The dimensional structure of the data is not known.
2. Multiple types of information such as the quality of the writing is required.
3. The number of items and objectives is quite large.

Three, in order to assess the validity of the judgments of content specialists, data should be collected on each item from more than one group of specialists employing more than one type of judgmental technique.

Four, there was a strong degree of relationship between the index of item-objective congruence and Lu's coefficient of agreement. This relationship would seem to indicate that for small applications such as in a classroom setting, one can avoid the involved calculations for Lu's coefficient and still insure that the data is reliable by setting a high cutoff score for the index of item-objective congruence.

While the separate findings relating to both approaches to assessing item validity are important, more crucial to the purposes of this study are findings obtained through an integration of the information provided by both techniques.

The first of these findings is that the use of empirical analyses alone cannot provide both necessary and sufficient conditions for establishing item validity. This was demonstrated in two ways. One, even though the items in Data Set Two were constructed through the use of item forms and thereby were given inherently a direct relationship to their intended objectives, a number of these items were identified as invalid through empirical analyses. Two, items identified as invalid by the content specialists were not so identified by the use of empirical analyses.

The second finding is that while the use of the judgments of content specialists provide sufficient information for assessing item validity, such a technique does not ensure that the item will have desirable statistical characteristics. This was demonstrated by the fact that a number of items declared valid by content specialists were identified as invalid by empirical analyses.

5.3. Models for the Construction of Criterion-Referenced Tests Based on the Use of Item Validation Procedures

The fundamental premise underlying the material in this section is that practitioners constructing criterion-referenced tests have one of two primary purposes: the estimation of the examinee's mastery score; or the assignment of the examinee to a mastery state. In order to achieve one of these purposes, the practitioner must generalize from actual test results to the content domain

being tested. If this generalization is to be meaningful the test items must be representative as random subsets of the universe of items from the content domain. As a matter of practical consequence, two means for satisfying this requirement are the use of item generation procedures and less rigid item construction strategies such as amplified objectives provided the question of whether these strategies do indeed produce a homogeneous set of items is empirically examined.

In this section, two models for the construction of criterion-referenced tests will be presented. For model one it will be assumed that extensive resources are available to the test constructor for developing and validating test items. Such an assumption is usually met in large curriculum projects. For model two, it will be assumed that the test constructor (*e.g.*, the classroom teacher) is limited in both the time and resources he or she has to devote to test development. The major objective of both test construction models is to increase the reliability and validity of criterion-referenced test data through the identification, reduction and elimination of sources of error relating to criterion-referenced test items. Four examples of sources of error which effect the interpretation of criterion-referenced test results are:

1. Items are not measures of their intended objectives.

2. Items which are measures of their intended objectives have undesirable statistical characteristics.
3. Items which are measures of the same objective have different statistical characteristics.
4. The ability level of examinees and quality of instruction differentiality effects the item characteristics.

5.3.1 Test Construction Model One

For model one, it has been assumed that time and amount of other resources are not limited and that an important goal is the development of a bank of valid items for a particular content domain. Under these circumstances, there would be a large number of items to validate. If as should be the case, the items were written using item writing guidelines, then the validation study should be concerned with evaluating the forms. Each item form should have the same number of items; therefore, some objectives will be represented by more items than others, in the situation where more than one item form has been developed for some objectives.

In order to validate these item forms, two separate studies should be conducted. One study would be used to collect judgmental data from content specialists and the other to collect empirical data from examinee reference groups.

For the study to collect judgmental data, it is recommended that more than one judgmental technique be used and that multiple types of information on each item form be collected. Further, it is suggested that each item form be represented by at least two items so that more than one judgment per form be available for analysis. The goal of this type of study will be to validate the item form rather than the individual items. The following paradigm provides an example of how a judgmental study can be organized.

Technique	Type of Information	Assessment of Reliability	Assessment of Validity
1. Hemphill-Westie (Categorizing)	1. Item-objective match	1. Lu's coefficient of agreement	1. Correlate with rating data
2. Rating	1. Item-objective match	1. Standard deviations of rating	1. Correlate with category data
	2. Quality of item writing		2. Correlate with empirical data
	3. Difficulty of item		

The following set of rules can be used to apply the data collected from such a judgmental study:

1. If the index of item-objective congruence (IC) is less than the acceptable cutoff level but the rating value (S) is above the acceptance level,

examine the instructional objectives for potential overlap.

2. If IC and S are less than the acceptable cutoff levels, revise the item form.
3. If IC is greater than the cutoff value and S is below the cutoff value, there may be a problem with the judgmental data. That is, the content specialists may be using different criteria for both techniques in judging the item forms.

For the study to collect empirical data, it is recommended that a repeated measure design which can isolate the effects due to instruction be used. Ideally, three measures, corresponding to pre-, post-, and delayed post-test instructional occasions should be obtained from the examinee groups. The unit of measurement must be the objective subtest scores not total test scores for each examinee. The major factor of interest is the effects due to instruction.

The absence of an instructional main effect for any objective subtest should preclude any item analyses on the related items. The essence of this discussion is that empirical analyses relating to the items must take into consideration the context in which the data are collected. Otherwise, a straightforward application of the empirical conditions may result in a number of false decisions concerning the items. Once, the preliminary analyses have been conducted various rules could be

applied to the items in order to identify which of them are in need of revision.

Once these analyses have been conducted, the following rules could be applied to the data:

1. If the average difference index (DI), obtained by summing across items within an item form is less than the average gain for other item forms within the same objective subgroup or the average gain for all the item forms, revise the item form.
2. If the average KI coefficient obtained by summing across items within an item form is less than .80, revise the item form.
3. If Cochran's Q statistic, calculated for the items within an item form, is significant, revise the item form.
4. If any of the discrimination indices for the items from the post- and delayed post-test data within an item form are negative, revise the item form.
5. If the factor loadings for items from the post-test data within an item form appear to be different for the same common factor, revise the item form.
6. If the items from the post-test data within an item form load on the same factor as items from other objective subgroups, examine both the item forms and objectives to determine if different aspects of the content domain are being tapped.

The rules listed above were applied to the item forms for each objective subgroup. Since each item form by definition should produce a very homogeneous set of items, and since the judgments of the content specialist would have already been used as an initial screening, the empirical condition should be rigorously applied in order to identify problem item forms.

5.3.2 Test Construction Model Two

As has been previously indicated, for test construction model two, it has been assumed that the test constructor's time and resources are limited. Other factors which must be taken into consideration are:

1. Items are rarely written according to any guidelines or formats.
2. In many situations, few practitioners conduct item analyses or examine item properties such as difficulty indices.
3. Incorrect responses to items are mostly attributed to lack of knowledge on the examinee's part rather than to sources of error in the item.

Taking these factors and previously stated assumptions into consideration, for a test construction model to be of practical use, it must:

1. Require no sophisticated statistical analyses or significant increases in time to implement;

2. Provide a means of not only determining whether or not items are measures of their intended objectives but also for establishing a degree of generalizability by requiring that items measuring the same objective be relatively homogeneous with regard to content, response mode, and type of distractors.

The use of the judgments of content specialists assumes a role of particular importance for test construction model two as typically little empirical analyses are conducted in classroom settings. It is suggested that items written or selected from prepared sources be judged on two separate levels; the extent to which an item is a measure of an objective; and the extent to which items measuring the same objectives have similar content, type of distractors, level of difficulty and response mode. An academic department would have to cooperate to effectively implement this suggestion as in effect each item would have to receive multiple judgments. One way to utilize judgmental procedures would be to split the content specialists into two groups and have one group judge the items on the extent to which an item is a measure of its intended objective and the second group judge them on relevant characteristics. Two of the judgmental techniques discussed in this study, the Hemphill-Westie categorizing procedure, and the semantic differential rating procedure could be used to collect this dual set of information.

In the first phase, the items should be judged across all negatives. In the second phase, the items measuring the same objectives should be judged against each other as to their similarity on a number of important characteristics.

Empirical analyses should be used but not to the extent that they are used in test construction model one. This de-emphasis of empirical analyses as a means for assessing item validity is an acknowledgement that teachers rarely collect the type of data required for such analyses. Further, even when teachers do collect the required data, limitations on time and statistical sophistication prevent them from effectively utilizing the data.

The following rules are suggested as the essence of test construction model two.

1. If the Hemphill-Westie index of item-objective congruence is less than .80 eliminate the item from the item pool.
2. If an item is rated as more difficult than other items within the same objective subgroup, revise the item.
3. If an item is judged to be poorly written, revise it.
4. If the distractors of an item are judged to be different for items of similar content, revise them.
5. If items within the same objective subgroup have different content or response mode, cluster them on the basis of item form if possible. In any

event, identify such items so that they will not be empirically analyzed together.

For the following rules, all analyses can be done on the post-test data only.

6. If items within an objective subgroup have difficulty levels less than the average item difficulty levels of other objective subgroups, review the instructional material as well as examine the items.
7. If an item has a difficulty level which is different from the other items of similar content with the objective subgroup, revise it.
8. If the average KI coefficient for an objective subgroup is lower than .80, examine the item precision indices. Any items with precision indices less than .80 should be revised.
9. If a measure of item discrimination for an item is negative, revise the item.
10. If the item has been used previously, compare the item data. Discrepancies may provide same insight as to the impact of instruction.

5.4 Limitations and Suggestions for Future Research

While this study, in addressing the problem of assessing item validity for criterion-referenced tests has provided important findings, the analyses were conducted after the items had been administered to examinees. Therefore, the task remains to ascertain whether the

recommendations and methodological considerations discussed in this study can be used to construct criterion-referenced tests that provide more valid data than tests constructed without the benefit of this information.

An additional area of concern is the relatively low performance levels on the tests from both Data Sets One and Two. It is apparent that the students had a great deal of difficulty with the instructional material. This difficulty may have been a function of the amount of time the examinees had to study the material. Future studies should seek to analyze items in situations where the mastery levels for the instructional materials are both known and varied.

The results concerning the use of content specialists as a means for assessing item validity were very encouraging. However, there is a need for much more developmental work to deal with the many methodological issues which were identified in this study. Some examples of these issues are:

1. How to match judgmental techniques with information requirements?
2. How meaningful are the judgments of content specialists on areas such as quality of writing and level of difficulty of the item?
3. How to assess the reliability and validity of the data?

Research to provide answers to questions such as those posed above should become an integral part of any developmental work in the area of item validation for criterion-referenced tests.

5.5 Conclusions

While there is no denying that the objectives-based instruction movement has been accepted by educators as a constructive innovation, many problems hinder effective implementation. For example, few practitioners receive any training in test construction and those that do usually receive it on norm-referenced measured rather than criterion-referenced measures. Also, there has been a shortage of guidelines relating to constructing and validating items for criterion-referenced tests that are used in objectives-based instructional programs. In this study, procedures have been set forth to ascertain whether or not an item is a measure of its intended objective (item validity). These procedures thus represent an important theoretical development and a practical means for enabling practitioners to construct tests suitable for criterion-referenced interpretations of examinee performance.

REFERENCES

Anderson, R. C. How to construct achievement tests to assess comprehension. *Review of Educational Research*, 1972, 42, 145-170.

Boozer, R. F., and Lindvall, C. M. An investigation of selected procedures for the development and evaluation of hierarchical curriculum structures. Learning Research and Development Center, University of Pittsburgh, 1971.

Bormuth, J. R. *On the theory of achievement test items*. Chicago: University of Chicago Press, 1970.

Brennan, R. L., and Light, R. J. Measuring agreement when categories are not predetermined. Boston: Laboratory of Human Development, Harvard University, 1973.

Brennan, R. L., and Stolurow, L. M. An empirical decision process for formative evaluation. *Research Memorandum No. 4*. Harvard CAI Laboratory, Cambridge, Mass., 1971.

Carroll, J. B. A model of school learning. *Teachers College Record*, 1963, 64, 723-733.

Carroll, J. B. Problems of measurement related to the concept of learning for mastery. *Education Horizons*, 1970, 48, 71-80.

Cochran, W. G. The comparison of percentages in matched samples. *Biometrika*, 1950, 37, 256-266.

Cohen, J. A coefficient of agreement for nominal Scales. *Educational and Psychological Measurement*, 1960, 20, 37-46.

Cox, R. C. Evaluative aspects of criterion-referenced measurement. Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, 1970. (ERIC, Ed 038 679).

Cox, R. C., and Graham, G. T. The development of a sequentially scaled achievement test. *Journal of Educational Measurement*, 1966, 3, 147-150.

Cox, R. C., and Vargas, J. C. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, 1966.

Cronbach, L. J. Validation of educational measures. In *Proceedings of the 1969 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service, 1970.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational measurement*. Washington, D.C.: American Council on Education, 1971 (pp. 443-507).

Cronbach, L. J. and Gleser, C. *Psychological tests and personnel decisions*. (2nd ed.) Urbana, Ill.: University of Illinois Press, 1965.

Cronbach, L. J.; Rajaratnam, N. and Gleser, G. C. Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology*, 1963, 16, 137-163.

Dahl, T. Toward an evaluative methodology for criterion-referenced measures: objective-item congruence. Paper presented at the annual meeting of the California Educational Research Association, San Diego, 1971.

Darlington, R. B., and Bishop, C. H. Increasing test validity by considering inter-item correlations. *Journal of Applied Psychology*, 1966, 50, 322-330.

Davis, F. B., and Diamond, J. J. The preparation of criterion-referenced tests. *CSE Monograph Series in Evaluation*. Los Angeles: Center for the Study of Evaluation, UCLA, 1974.

Engelhart, M. D. A comparison of several item discrimination indices. *Journal of Educastional Measurement*, 1965, 2, 69-76.

Ferguson, R. L., and Hsu, T. The application of item generators for individualizing mathematics testing and instruction. Learning Research and Development Center, University of Pittsburgh, 1971.

Findley, W. G. A rationale for evaluation of item discrimination statistics. *Educational and Psychological Measurement*, 1956, 16, 175-180.

Flanagan, J. C. Functional education for the seventies. *Phi Delta Kappan*, 1967, 49, 27-32.

Flanagan, J. C. Program for learning in accordance with needs. *Psychology in the Schools*, 1969, 6, 133-136.

Fleiss, J. L. Measuring nominal agreement among many raters. *Psychological Bulletin*, 1971, 76, 378-382.

Fremer, J. Criterion-referenced interpretations of achievement tests. *Test Development Memorandum TDM-71-1.* Princeton, N.J.: Educational Testing Service, 1972.

Fremer, J. and Anastasio, E. Computer-assisted item writing—I (Spelling Items). *Journal of Educational Measurement,* 1969, 6, 69-74.

Girard, R., and Cliff, N. A comparison of methods for judging the similarity of personality inventory items. *Multivariate Behavioral Research,* 1973, 8, 71-88.

Glaser, R. Adapting the elementary school curriculum to individual performance. In *Proceedings of the 1967 Invitational Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service, 1968 (pp. 3-36).

Glaser, R. Evaluation of instruction and changing educational models. In M. C. Wittrock and D. E. Wiley (Eds.), *The Evaluation of Instruction.* New York, N.Y.: Holt, Rinehart and Winston, 1970.

Glaser, R., and Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational measurement.* Washington, D. C.: American Council on Education, 1971 (pp. 625-670).

Gulliksen, H. *A theory of mental tests.* New York: Wiley, 1959.

Guttman, L. A. The basis for scalogram analysis. In S. A. Stouffer, *et al., Measurement and prediction.* Princeton, N.J.: Princeton University Press, 1950.

Hambleton, R. K. A report on the research and evaluation activities in the Jamesville-DeWitt individualized instructional program in ninth grade science. Final Report. Albany, N.J.: Bureau of School and Cultural Research, New York State Education Department, 1971.

Hambleton, R. K. Testing and decision-making procedures for selected individualized instructional programs. *Review of Educational Research,* 1974, 44, 371-400.

Hambleton, R. K., and Gorth, W. P. Criterion-referenced testing: issues and applications. *Center for Educational Research Technical Report No. 13.* Amherst, Mass.: School of Education, University of Massachusetts, 1971.

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

Hambleton, R. K.; Swaminathan, H.; and Algina, J. Some contributions to the theory and practice of criterion-referenced testing. In D. N. M. de Gruijter, L. J. Th. van de Kamp and H. F. Crombag (Eds.), *Advances in Psychological and Educational Measurement*. New York: Wiley, 1976.

Harman, H. H. Modern factor analysis. (2nd ed.) Chicago: University of Chicago Press, 1967.

Harris, M. J., and Stewart, D. M. Application of classical strategies to criterion-referenced test construction. Paper presented at the Annual Meeting of the American Educational Research Association, New York, 1971.

Hemphill, J., and Westie, C. M. The measurement of group dimensions. *Journal of Psychology*, 1950, 29, 325-342.

Hively, W. Specifying "terminal behavior" in mathematics. *Harvard Committee on Programmed Instruction*, 1962, Unpublished.

Hively, W. Introduction to domain-referenced testing. *Evaluational Technology*, 1974, 14, 5-10.

Hively, W.; Maxwell, G.; Rabehl, G; Sension, D.; and Lunden, S. Domain-referenced curriculum evaluation: a technical handbook and a case study from the Minnemast Project. *Monograph Series in Evaluation, No. 1.* Los Angeles: Center for the Study of Evaluation, University of California, 1973.

Hively, W.; Patterson, H. L.; and Page, S. "A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, 1968, 5, 275-290.

Hotelling, H. The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *Annals of Mathematical Statistics*, 1940, 11, 271-283.

Hoyt, C. J. Test reliability estimated by analysis of variance. *Psychometrika*, 1941, 6, 153-160.

Ivens, S. H. An investigation of item analysis, reliability and validity in relation to criterion-referenced tests. Unpublished doctoral dissertation, Florida State University, 1970.

Ivens, S. H. A pragmatic approach to criterion-referenced measures. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1972.

Jackson, R. Developing criterion-referenced tests. ERIC Clearinghouse on Tests, *Measurement and Evaluation*, 1970.

Joreskog, K. G. Some contributions to maximum likelihood factor analysis. *Psychometrika*, 1967, 32, 443-482.

Joreskog, K. G. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 1969, 34, 183-202.

Kriewall, T. E. Applications of information theory and acceptance sampling principles to the management of mathematics instruction. Unpublished doctoral dissertation, University of Wisconsin, 1969.

Kuder, G. F., and Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, 2, 151-160.

Lawley, D. N., and Maxwell, A. E. *Factor analysis as a statistical method.* London: Butterworths, 1971.

Light, R. J. Issues in the analysis of qualitative data. In R. Travers (Ed.), *Second handbook of research on teaching*. Chicago: Rand McNally, 1973 (pp. 318-381).

Lingoes, J. C. Multicple scalogram analysis: A set theoretic model for analyzing dichotomous items. *Educational and Psychological Measurement*, 1963, 23, 521-523.

Livingston, S. A. Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 1972, 9, 13-26.

Lord, F. M. Estimating norms by item sampling. *Educational and Psychological Measurement*, 1962, 22, 259-267.

Lord, F. M., and Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Lu, K. H. A measure of agreement among subjective judgments. *Educational and Psychological Measurement*, 1971, 31, 75-84.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education: Current practices.* San Francisco: McCutchan Publishers, 1974.

Mulaik, S. *The foundations of factor analysis.* New York: McGraw-Hill, 1972.

Novick, M. R., and Jackson, P. *Statistical methods for educational and psychological research.* New York: McGraw-Hill, 1973.

Osburn, H. G. Item sampling for achievement testing. *Educational and Psychological Measurement,* 1968, 28, 95-104.

Osgood, C. E.; Suci, G. J.; and Tannenbaum, P. H. *The measurement of meaning.* Urbana: University of Illinois Press, 1957.

Ozenne, D. G. Toward an evaluative methodology for criterion-referenced measures: test sensitivity. *CSE Report No. 72.* Los Angeles: Center for the Study of Evaluation, Graduate School of Education, UCLA, 1971.

Popham, W. J. Indices of adequacy for criterion-referenced test items. In W. J. Popham (Ed.), *Criterion-referenced measurement.* Englewood Cliffs, N.J.: Educational Technology Publications, 1971.

Popham, W. J. Selecting objectives and generating test items for objectives-based tests: problems in criterion-referenced measurement. *CSE Monograph Series In Evaluation.* Los Angeles: Center for the Study of Evaluation, UCLA, 1974.

Popham, W. J., and Husek, T. R. Implications of criterion-referenced measurement. *Journal of Educational Measurement,* 1969, 6, 1-9.

Rabehl, G. The experimental analysis of educational objectives. Unpublished doctoral dissertation, University of Minnesota, 1971.

Rahmelow, H. F.; Matthews, J. J.; and Jung, S. M. An empirical investigation of item analysis in criterion-referenced tests. Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, 1970.

Royer, M.; Hambleton, R. K.; and Cadorette, L. Individual differences in the long-term retention of meaningful material. *Laboratory of Psychometric and Evaluative Research Report No. 10.* Amherst, Mass.: School of Education, University of Massachusetts, 1975.

Ryan, J. J. Teacher judgments of test item properties. *Journal of Educational Measurement,* 1968, 5, 301-306.

Sabers, D. L., and Kania, J. G. Item precision in criterion-referenced measurement. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, April 1972.

Saupe, J. L. Selecting items to measure change. *Journal of Educational Measurement*, 1966, 3, 223-228.

Sheehan, D. S., and Hambleton, R. K. An evaluative study of the Jamesville-Dewitt individualized science program (1971-1972). Final Report. Albany, N.Y.: Bureau of School and Cultural Research, New York State Education Department, September 1972.

Siegel, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.

Sirotnik, K. An analysis of variance framework for matrix sampling. *Educational and Psychological Measurement*, 1970, 30, 891-908.

Skager, R. W. Generating criterion-referenced tests from objective-based assessment systems: unsolved problems in test development, assembly, and interpretation. *CSE Monograph Series in Evaluation*. Los Angeles: Center for the Study of Evaluation, UCLA, 1974.

Swaminathan, H.; Hambleton, R. K.: and Algina, J. Reliability of criterion-referenced tests: a decision-theoretic formulation. *Journal of Educational Measurement*, 1974, 11, 263-268.

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

Tinkelman, S. N. Planning an objective test. In R. L. Thorndike (Ed.), *Educational measurement*. Washington, D.C.: American Council on Education, 1971.

Torgerson, W. S. *Theories and methods of scaling*. New York: Wiley, 1959.

Washburne, C. W., and Maryland, S. P., Jr. *Winnetka: the history and significance of an educational experiment*. Englewood Cliffs, N.J.: Prentice-Hall, 1963.

Whitney, D. R., and Sabers, D. L. Two generalizations of the item discrimination index to multi-score items. *Journal of Experimental Education*, 1971, 39, 89-92.

Wilhelms, F. T. The curriculum and individual differences. In N. B. Henry (Ed.), *Individualizing instruction.* Sixty-first Yearbook (Part I) of the National Society for the Study of Education. Chicago: University of Chicago Press, 1962, 62-74.

Wood, R., and Skurnik, L. S. *Item banking: a method for producing school-based examinations and nationally comparable grades.* London: National Foundation for Educational Research in England and Wales, 1969.

APPENDIX A

DATA SET ONE

1. Instructional Objectives
2. Item Validation Questionnaire
3. Answer Sheet
4. Test Items

Instructional Objectives

1. Given a substance, the student will be able to identify it as a mixture by its characteristics.

2. Given the atomic number and weight of an element, the student will be able to select the correct Bohr atomic diagram.

3. Using the Periodic Table, the student will be able to determine the most common valence or oxidation number of an element.

4. Given the chemical formula for the molecule, the student will be able to determine the number of atoms in the molecule.

5. Using the Periodic Table, the student will be able to determine the change in the electron configuration when an atom becomes an ion.

6. Given a substance, the student will be able to identify it as a compound by its characteristics.

7. Given a substance, the student will be able to identify it as an element by its characteristics.

8. Given a table of radicals and a Periodic Table, the student will be able to select the correct chemical formula for a compound.

9. Given the Bohr model of the atom and the Periodic Table, the student will be able to identify the atom.

10. Using the Periodic Table, the student will be able to classify an element as a metal or nonmetal.

11. Using the Periodic Table, the student will be able to determine certain characteristics of an atom such as atomic number, atomic mass or weight, number of protons, electrons or neutrons.

Item Validation Questionnaire

This questionnaire is administered as part of a research project that is designed to study the problem of item validation. Primarily we are interested in the extent to which science teachers and other qualified people are able to match the items in a science pool to the particular instructional objectives that they supposedly measure. Clearly, among curriculum evaluators and evaluators of student achievement this is one of the most basic and important questions to ask since unless there is a detectable match between a test item and an instructional objective there is little that can be inferred about mastery level of a student on the objective from his performance on the particular test item.

In this task there are 11 objectives listed on page 2 of the handout and 40 test items presented in the attached test. Your task is to match each item with each of the instructional objectives. You will indicate your answers by assigning one of the following ratings for each item relative to each objective:

1 - if you feel the item is definitely a measure of the objective

0 - if you cannot make a decision whether the item is a measure of objective

-1 - if you feel the item is definitely not a measure of the objective

Read each item carefully then mark your answers in the appropriate spots on the answer sheet.

It is possible that some items will not measure any of the objectives. Also, some items may measure more than one of the objectives.

Before you begin the task be sure to read the 11 instructional objectives carefully.

Name: _____ Date: _____
(Please Print)

ANSWER SHEET

Item	Objective 1	2	3	4	5	6	7	8	9	10	11
1	—	—	—	—	—	—	—	—	—	—	—
2	—	—	—	—	—	—	—	—	—	—	—
3	—	—	—	—	—	—	—	—	—	—	—
4	—	—	—	—	—	—	—	—	—	—	—
5	—	—	—	—	—	—	—	—	—	—	—
6	—	—	—	—	—	—	—	—	—	—	—
7	—	—	—	—	—	—	—	—	—	—	—
8	—	—	—	—	—	—	—	—	—	—	—
9	—	—	—	—	—	—	—	—	—	—	—
10	—	—	—	—	—	—	—	—	—	—	—
11	—	—	—	—	—	—	—	—	—	—	—
12	—	—	—	—	—	—	—	—	—	—	—
13	—	—	—	—	—	—	—	—	—	—	—
14	—	—	—	—	—	—	—	—	—	—	—
15	—	—	—	—	—	—	—	—	—	—	—
16	—	—	—	—	—	—	—	—	—	—	—
17	—	—	—	—	—	—	—	—	—	—	—
18	—	—	—	—	—	—	—	—	—	—	—
19	—	—	—	—	—	—	—	—	—	—	—
20	—	—	—	—	—	—	—	—	—	—	—

ANSWER SHEET

Item	Objective 1	2	3	4	5	6	7	8	9	10	11
21	—	—	—	—	—	—	—	—	—	—	—
22	—	—	—	—	—	—	—	—	—	—	—
23	—	—	—	—	—	—	—	—	—	—	—
24	—	—	—	—	—	—	—	—	—	—	—
25	—	—	—	—	—	—	—	—	—	—	—
26	—	—	—	—	—	—	—	—	—	—	—
27	—	—	—	—	—	—	—	—	—	—	—
28	—	—	—	—	—	—	—	—	—	—	—
29	—	—	—	—	—	—	—	—	—	—	—
30	—	—	—	—	—	—	—	—	—	—	—
31	—	—	—	—	—	—	—	—	—	—	—
32	—	—	—	—	—	—	—	—	—	—	—
33	—	—	—	—	—	—	—	—	—	—	—
34	—	—	—	—	—	—	—	—	—	—	—
35	—	—	—	—	—	—	—	—	—	—	—
36	—	—	—	—	—	—	—	—	—	—	—
37	—	—	—	—	—	—	—	—	—	—	—
38	—	—	—	—	—	—	—	—	—	—	—
39	—	—	—	—	—	—	—	—	—	—	—
40	—	—	—	—	—	—	—	—	—	—	—

Science IPI-Tests

Instructional Module-The Structure of Matter

1. Of the following substances, the best example of a compound is

   (1) hydrogen (2) dirt (3) gold (4) water

2. Which of the following substances is a compound?

   (1) oxygen (2) sugar (3) iron (4) brass

3. Which of the following substances cannot be broken down by any chemical process?

   (1) water (2) salt (3) air (4) iron

4. Which of the following substances is an element?

   (1) sugar (2) phosphorous (3) carbon monoxide (4) milk

5. Of the following substances, the best example of a mixture is

   (1) hydrogen (2) dirt (3) gold (4) water

6. Of the following substances, the best example of a mixture is

   (1) air (2) silver (3) water (4) salt

7. Which one of the following cannot be broken down into anything simpler by ordinary chemical means?

(1) an element (2) a mixture (3) a compound (4) all of them

8. Which of the following is made up of more than one material with no definite proportions in their composition?

(1) an element (2) a mixture (3) a compound (4) all of them

9. Which of the following is made up of only one kind of atom?

(1) an element (2) a mixture (3) a compound (4) all of them

10. An unknown substance occurs as a powder which appears light green in color. When a student places the powder in water, he finds that some of the powder dissolves, forming a green solution. However, the rest of the substance will not dissolve and settles to the bottom as a white powder. If no chemical reaction occurred, the substance was most likely

(1) an element (2) a compound (3) a mixture (4) an inert material

11. The atomic number of chlorine is

(1) 17 (2) 18 (3) 35 (4) 36

12. The number of protons in an atom of nitrogen is

(1) 7 (2) 14 (3) 28 (4) 31

13. The number of electrons in an atom of sodium is

(1) 11 (2) 12 (3) 19 (4) 23

14. The number of neutrons in an atom of fluorine is

(1) 9 (2) 10 (3) 19 (4) 20

15. In any atom, the number of electrons is

(1) equal to the number of neutrons.
(2) greater than the number of protons.
(3) less than the number of protons.
(4) equal to the number of protons.

16. The number of neutrons in an atom of neon is

(1) 7 (2) 10 (3) 14 (4) 20

17. The number of electrons in an atom of iodine is

(1) 53 (2) 73 (3) 77 (4) 127

18. What is the atomic weight of an element containing 10 protons, 15 neutrons, and 10 electrons?

(1) 10 (2) 15 (3) 25 (4) 35

19. What is the charge on the nucleus of an atom which contains 4 protons, 6 neutrons, and 4 electrons?

(1) 0 (2) +2 (3) +4 (4) +8

20. The particle in the atom which has weight and an electrical charge of +1 is th

(1) electron (2) proton (3) nucleus (4) neutron

21. Which of the following is most likely to be the structure of the nucleus of fluorine, atomic number 9?

(1) 18 protons, 18 neutrons
(2) 9 protons, 10 neutrons
(3) 18 protons, 9 neutrons
(4) 19 protons, 19 neutrons

22. Magnesium has an atomic number of 12 and an atomic weight of 24. Which of these diagrams represents an atom of magnesium?

(1) 

(2) 

(3) 

(4) 

23. The atomic number of a certain element is 5. Its atomic weight is 11. Which of these diagrams represents an atom of this element?

(1)

(3)

(2)

(4)

24. The diagram at the right represents a Bohr model of an atom. What type of atom is it?

(1) sulfur
(2) sodium
(3) oxygen
(4) chlorine

25. The diagram at the right represents the Bohr model of an atom of

(1) helium
(2) beryllium
(3) carbon
(4) hydrogen

26. In general, atoms with 3 electrons in the outer shell belong to which of the following categories?

(1) nonmetals (2) inert gases (3) metals (4) none of these

27. The elements in group II of the Periodic Table are

(1) metals (2) nonmetals (3) inert gases (4) none of these

28. The following choices represent electron configurations of different atoms. Which represents an atom of a metal?

(1) 2-8-6 (2) 2-8-8-7 (3) 2-8-18-18-4 (4) 2-8-3

29. The oxidation or valence number of carbon is usually

(1) +1 (2) +2 (3) +3 (4) +4

30. The oxidation or valence number of calcium is usually

(1) +1 (2) +2 (3) +3 (4) +4

31. The oxidation or valence number of aluminum is +3. This means that, in a chemical reaction, aluminum may

(1) gain 5 electrons
(2) gain 3 electrons
(3) give away 5 electrons
(4) give away 3 electrons

32. How does a fluoride <u>ion</u> differ from a fluorine <u>atom</u>?

(1) it has more electrons
(2) it has more neutrons
(3) it has fewer electrons
(4) it has a positive charge

33. Calcium is the element with atomic number 20. How many electrons would there be in an ion of calcium?

(1) 17 (2) 18 (3) 20 (4) 40

34. When an atom loses an electron, it becomes an ion with a charge of

(1) -2 (2) -1 (3) +1 (4) +2

35. How many atoms are there in a molecule of calcium phosphate, $Ca_3(PO_4)_2$?

(1) 5 (2) 9 (3) 10 (4) 13

36. How many atoms are there in a molecule of ammonium phosphate, $(NH_4)_3PO_4$?

(1) 6 (2) 7 (3) 13 (4) 20

37. How many hydrogen atoms are there in $2H_2O$?

(1) 1 (2) 2 (3) 4 (4) 8

For items (38-40) select the correct chemical formula for the material listed:

38. Sodium phosphate (1) $NaPO_4$ (2) Na_2PO_4 (3) Na_3PO_4 (4) $Na_2(PO_4)_2$

39. Calcium nitrate (1) $CaNO_3$ (2) Ca_2NO_3 (3) $Ca(NO_3)_3$ (4) $Ca(NO_3)_2$

40. Potassium sulfate (1) K_2SO_4 (2) $K(SO_4)_2$ (3) KSO_4 (4 K_3SO_4

APPENDIX B

DATA SET TWO

1. Instructional Objectives

2. Item Validation Questionnaire

   Approach 1, Objectives, Test Items, Answer Sheet
   Approach 2, Objectives, Test Items, Answer Sheet
   Approach 3, Objectives, Test Items, Answer Sheet

3. Test Items

   Form A
   Form B
   Borm C

Instructional Objectives

1. Given a substance or its characteristics, the student will be able to identify it as a compound.

2. Using the Periodic Table, the student will be able to determine the atomic weight and atomic number of an atom.

3. Given the Bohr model of an atom and the Periodic Table, the student will be able to identify the atom.

4. Using the Periodic Table, the student will be able to determine the change in the electron configuration number(s) when an atom becomes an ion.

5. Given a substance or its characteristics, the student will be able to identify it as an element.

6. Using the Periodic Table of the elements, the student will be able to determine the number of protons, electrons and neutrons of an atom.

7. Using the Periodic Table, the student will be able to classify a given element as a metal or nonmetal.

8. Given the chemical formula for the molecule, the student will be able to determine the number of atoms in a molecule.

9. Given a substance or its characteristics, the student will be able to identify it as a mixture.

10. Given the atomic number and atomic weight of an element, the student will be able to select the correct Bohr atom diagram.

11. Using the Periodic Table, the student will be able to determine the most common oxidation or valence number of an element.

12. Given a table of radicals and a Periodic Table, the student will be able to select the correct chemical formula for a compound.

ITEM VALIDATION QUESTIONNAIRE

This questionnaire is being administered as a part of a research project that is designed to study possible ways for determining item validity. Item validity has to do with the extent to which a test item measures an instructional objective. Primarily, we are interested in the opinions of science teachers and other qualified people on the appropriateness of a set of test items for measuring some science objectives. Clearly, the problem of item validity is an important one for classroom teachers since it is essential to establish the validity of a test item before it is used in a classroom test.

One important question concerns the best approach for collecting your evaluative judgements of a set of test items. In this questionnaire you will be given an opportunity to use three different approaches. Through our research we hope to establish which of the three is the best.

The questionnaire is divided into three sections: In each section is a set of directions describing an approach for rating the test items, a list of six science objectives, 16 test items, and an answer sheet.

Please go ahead now and complete the questionnaire.

Item Validation Task

-Approach 1-

In this section of the questionnaire, your task is to judge the relationships among 16 test items and six instructional objectives. For each of the 16 items, indicate the extent to which you feel it is a relevant measure of each objective. That is, your task is to indicate how relevant or appropriate you feel each item is as a measure of each of the objectives. Use the rating scale below:

1-Irrelevant	2-Slightly relevant	3-Somewhat relevant	4-Relevant	5-Highly relevant

Before you begin the rating task be sure to read the six instructional objectives and the 16 test items on the next few pages very carefully. Then go ahead and mark your ratings for each item on the answer sheet for approach one.

Objectives
-Form A-

1. Given a substance or its characteristics, the student will be able to identify it as a compound.

2. Using the Periodic Table of Elements, the student will be able to determine the atomic weight and atomic number of an atom.

3. Given the chemical formula of a molecule, the student will be able to determine the number of atoms number of an atom.

4. Using the Periodic Table, the student will be able to classify a given element as a metal or nonmetal.

5. Given a table of Radicals and a Periodic Table the student will be able to select the correct chemical for a compound.

6. Given a substance or its characteristics, the student will be able to identify it as a mixture.

Test Items
-Form A-

1. Which of the following substances is an example of a mixture?

   (1) air (2) gold (3) oxygen (4) water

2. The elements in group VII of the Periodic Table belong to which of the following categories?

   (1) metals (2) nonmetals (3) inert gases (4) none of these

3. How many oxygen atoms are there in $2H_2O$?

   (1) 1 (2) 2 (3) 4 (4) 8

4. The atomic weight of an element is approximately equal to the weight of which of the following particles?

   (1) protons (2) neutrons (3) nucleus (4) electrons

5. In general, atoms with 5 electrons in the outer shell belong to which of the following categories?

   (1) metals
   (2) nonmetals
   (3) neither metals nor nonmetals
   (4) both metals and nonmetals

6. How many atoms are there in a molecule of $(NH_4)_4SiO_4$?

   (1) 2 (2) 12 (3) 14 (4) 25

7. Which of the following is a mixture?

   (1) carbon (2) sugar (3) sodium chloride (4) blood

8. Which of the following particles is equal to the atomic number of an atom?

   (1) number of neutrons
   (2) number of electrons
   (3) number of protons
   (4) both (2) and (3) are correct

9. Which one of the following is a characteristic of metals?

   (1) form negative ions
   (2) combine with other metals
   (3) form positive ions
   (4) add electrons to the outer shell

10. Which one of the following is not a mixture?

(1) steel (2) cake (3) milk (4) oil

11. How many atoms are there in a molecule of sulfuric acid, H_2SO_4?

(1) 3 (2) 7 (3) 9 (4) 10

12. What is the atomic number of sodium?

(1) 11 (2) 19 (3) 23 (4) 39

13. How many atoms are there in a molecule of calcium hydroxide, $Ca(OH)_2$?

(1) 2 (2) 3 (3) 4 (4) 5

14. Which of the following substances is a mixture?

(1) salt (2) hydrogen (3) dirt (4) silver

15. What is the atomic weight of calcium?

(1) 20 (2) 40 (3) 45 (4) 60

16. In general, atoms with 2 electrons in the outer shell belong to which of the following categories?

(1) metals
(2) nonmetals
(3) neither metals nor nonmetals
(4) both metals and nonmetals

Answer Sheet
(Approach 1)

Remember that you should assign a rating of "1" to "5" (1 = irrelevant, 2 = slightly relevant, 3 = somewhat relevant, 4 = relevant, 5 = highly relevant) to indicate the extent to which you feel each item is a relevant measure of each objective.

	Objective					
Item	1	2	3	4	5	6
1	___	___	___	___	___	___
2	___	___	___	___	___	___
3	___	___	___	___	___	___
4	___	___	___	___	___	___
5	___	___	___	___	___	___
6	___	___	___	___	___	___
7	___	___	___	___	___	___
8	___	___	___	___	___	___
9	___	___	___	___	___	___
10	___	___	___	___	___	___
11	___	___	___	___	___	___
12	___	___	___	___	___	___
13	___	___	___	___	___	___
14	___	___	___	___	___	___
15	___	___	___	___	___	___
16	___	___	___	___	___	___

Item Validation Task

-Approach 2-

In this section of the questionnaire, your task is again to judge the relationships among 16 test items and six instructional objectives. However, in this second approach, indicate the extent to which you feel an item is a relevant measure of an objective by using the rating scale below:

3 - If you feel the item is definitely a measure of the objective.

2 - If you cannot make a decision whether the item is a measure of the objective.

1 - If you feel the item is definitely not a measure of the objective.

Before you begin the rating task be sure to read the six instructional objectives and the 16 test items on the next few pages very carefully. Then go ahead and mark your ratings for each item on the answer sheet for approach two.

Objectives
-Form B-

1. Given a table of Radicals and a Periodic Table, the student will be able to select the correct chemical formula for a compound.

2. Using the Periodic Table, the student will be able to classify a given element as a metal or nonmetal.

3. Using the Periodic Table, the student will be able to determine the change in the electron configuration number(s) when an atom becomes an ion.

4. Given a substance or its characteristics, the student will be able to identify it as a compound.

5. Given the Bohr model of an atom and the Periodic Table, the student will be able to identify the atom.

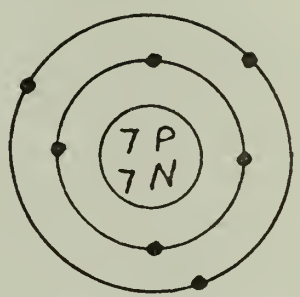
6. Given the atomic number and atom weight of an element, the student will be able to select the correct Bohr atom diagram.

Test Items
-Form B-

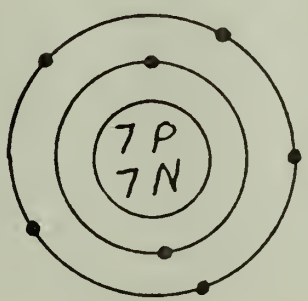
1. Nitrogen has an atomic number of 7 and an atomic weight of 14. Which diagram represents an atom of nitrogen?

(1)

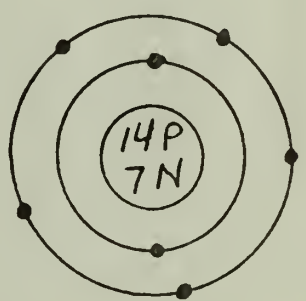
(3)



(2)

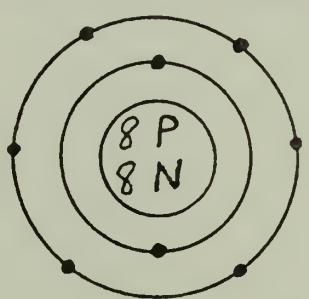


(4)



2. Which atom is represented by the diagram below?

(1) sodium
(2) copper
(3) oxygen
(4) carbon



3. What is the correct formula for sodium silicate?

(1) NaSiO (2) Na_2SiO_3 (3) Na_4SiO_3 (4) Na_6SiO_3

4. Which of the following substances is an example of a compound?

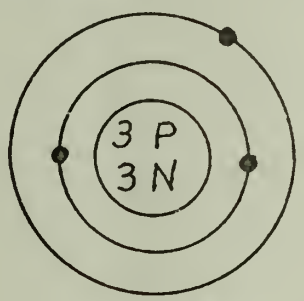
(1) hydrogen (2) oxygen (3) water (4) none of these

5. What is the correct chemical formula for aluminum hydroxide?

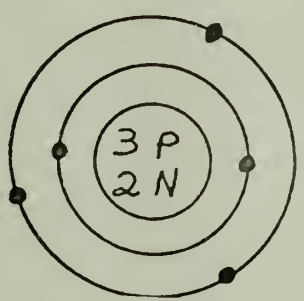
(1) AlOH (2) $AlOH_3$ (3) Al_3OH (4) $Al(OH)_3$

6. The atomic number of an imaginary element is 3. Its atomic weight is 5. Which of the diagrams below represents an atom of this element?

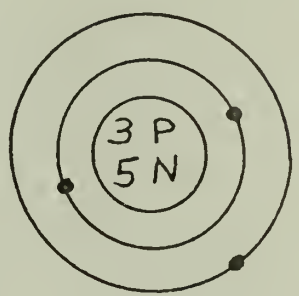
(1)



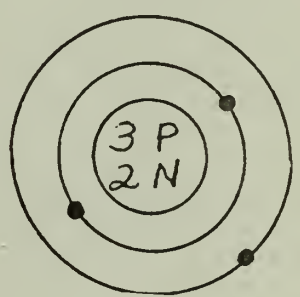
(2)



(3)



(4)

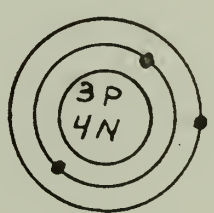


7. Which of the following is a substance that can only be broken down into two or more simpler substances by a chemical reaction?

(1) element (2) mixture (3) compound (4) none of these

8. Which atom is represented by the diagram below?

(1) hydrogen
(2) helium
(3) lithium
(4) beryllium



9. What is the correct formula for sodium carbonate?

(1) NaCO (2) $NaCO_3$ (3) Na_2CO_3 (4) $Na(CO_3)_2$

10. Which of the following is a material made of more than one kind of atom in a definite proportion?

(1) element (2) mixture (3) compound (4) none of these

11. Which of the following substances is an example of a compound?

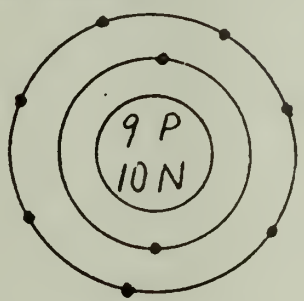
(1) air (2) silver (3) milkshake (4) salt

12. Which atom is represented by the diagram below?

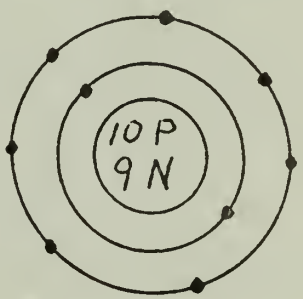
(1) hydrogen
(2) lithium
(3) helium
(4) carbon

13. The atomic number of a certain element is 9. Its atomic weight is 19. Which of the diagrams below represents an atom of this element?

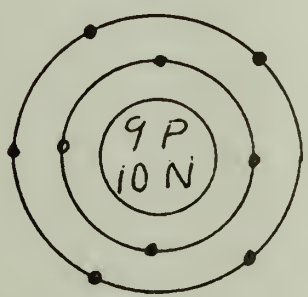
(1)



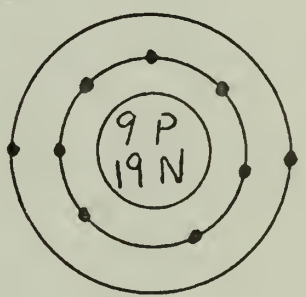
(3)



(2)

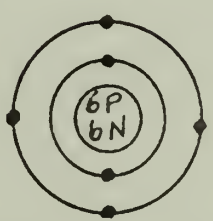


(4)



14. Which atom is represented by the diagram below?

(1) hydrogen
(2) lithium
(3) helium
(4) carbon

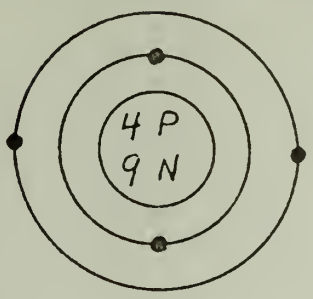


15. What is the correct formula for nickel chloride? chloride?

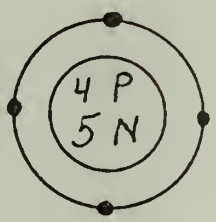
(1) NiCl (2) $NiCl_2$ (3) Ni_2Cl (4) $(NiCl)_2$

16. Beryllium has an atomic number of 4 and an atomic weight of 9. Which of the diagrams below represents an atom of beryllium?

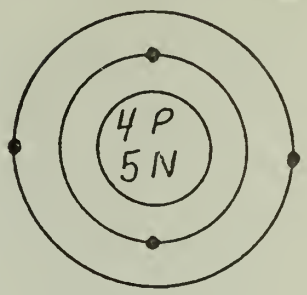
(1)



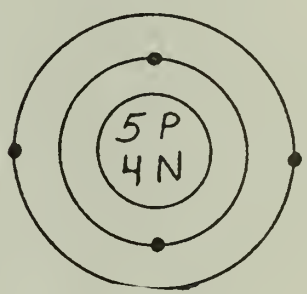
(2)



(3)



(4)



Answer Sheet
(Approach 2)

Remember that you should use the following rating scale:

3 - If you feel the item is definitely a measure of the objective.

2 - If you cannot make a decision whether the item is a measure of the objective.

1 - If you feel the item is definitely not a measure of the objective.

	Objective					
Item	1	2	3	4	5	6
1	___	___	___	___	___	___
2	___	___	___	___	___	___
3	___	___	___	___	___	___
4	___	___	___	___	___	___
5	___	___	___	___	___	___
6	___	___	___	___	___	___
7	___	___	___	___	___	___
8	___	___	___	___	___	___
9	___	___	___	___	___	___
10	___	___	___	___	___	___
11	___	___	___	___	___	___
12	___	___	___	___	___	___
13	___	___	___	___	___	___
14	___	___	___	___	___	___
15	___	___	___	___	___	___
16	___	___	___	___	___	___

Item Validation Task
-Approach 3-

In this final section of the questionnaire, the task is somewhat different from that of the first two sections. Proceed in the following way: First, read the complete list of six objectives and 16 test items on the next few pages. Then on the answer sheet, beside each objective number, write in the item numbers corresponding to the items that you think measure each of the objectives. Keep in mind that it is possible that some objectives will not be measured by <u>any</u> test items, and that some items will <u>not</u> measure any of the available objectives.

Objectives
-Form C-

1. Using the Periodic Table of the Elements, the student will be able to determine the number of protons, electrons, and neutrons of an atom.

2. Using the Periodic Table, the student will be able to determine the most common oxidation or valence number of an element.

3. Using the Periodic Table of the elements, the student will be able to determine the atomic weight and atomic number of an atom.

4. Given the chemical formula for the molecule, the student will be able to determine the number of atoms in a molecule.

5. Given a substance or its characteristics, the student will be able to identify it as an element.

6. Using the Periodic Table, the student will be able to determine the change in the electron configuration number(s) when an atom becomes an ion.

Test Items
-Form C-

1. What is the oxidation or valence number of aluminum?

   (1) +1 (2) +2 (3) +3 (4) +4

2. How many protons, neutrons and electrons are there in an atom of an element with an atomic weight of 30 and an atomic number of 12?

   (1) 12p 18n 12e
   (2) 12p 12n 18e
   (3) 30p 12n 30e
   (4) 18p 30n 12e

3. What do the "symbols" in the Periodic Table represent?

   (1) elements (2) mixtures (3) radicals (4) compounds

4. How many neutrons are there in an atom of nitrogen?

   (1) 7 (2) 14 (3) 28 (4) 31

5. What is the charge on the <u>ion</u> formed when an atom gains an electron?

   (1) +2 (2) -2 (3) +1 (4) -1

6. The oxidation or valence number of calcium is +2. What will a calcium atom do in a chemical reaction?

   (1) gain 6 electrons
   (2) gain 2 electrons
   (3) give away 6 electrons
   (4) give away 2 electrons

7. Fluorine is an element with an atomic number of 9. How many electrons would there be in an <u>ion</u> of fluorine?

   (1) 7 (2) 8 (3) 9 (4) 10

8. How many electrons are there in an atom of aluminum?

   (1) 13 (2) 14 (3) 26 (4) 27

9. The oxidation or valence number of iodine is -1. What will an iodine atom do in a chemical reaction?

(1) gain 7 electrons
(2) gain 1 electron
(3) give up 7 electrons
(4) give up 1 electron

10. Which of the following substances is an element?

(1) carbon dioxide (2) zinc (3) cement (4) steel

11. Beryllium is an element with an atomic number of 4. How many electrons would there be in an ion of beryllium?

(1) 2 (2) 4 (3) 6 (4) 9

12. How does a sodium ion differ from a sodium atom?

(1) ion has fewer electrons
(2) ion has more neutrons
(3) ion has more electrons
(4) ion has a negative charge

13. Which of the following substances cannot be broken down by any chemical process?

(1) salt (2) air (3) gold (4) water

14. What is the oxidation or valence number of sodium?

(1) +1 (2) +2 (3) +3 (4) +4

15. Which one of the following is not an element?

(1) iron (2) lithium (3) helium (4) none of these

16. How many protons are there in an atom which contains 12 neutrons and 10 electrons?

(1) 2 (2) 10 (3) 12 (4) 22

Answer Sheet

(Approach 3)

Beside each objective, indicate the item numbers corresponding to the items that you feel measure it.

Objective	Items Measuring the Objective
1	
2	
3	
4	
5	
6	

Concluding Comments

If you would like to share with us any reactions you had to the three approaches or suggest any other approaches that you feel may be useful please indicate them below:

Structure of Matter

Module Test, Form A

This test is designed to determine your knowledge of some selected science concepts in the area of the "Structure of Matter." Before beginning the test, print your name in the boxes provided on the answer sheet. In the space marked "TEST" on the answer sheet, write the letter "A". Be sure to use a soft-lead pencil for all marks on the answer sheet.

In completing the test, remember to blacken in the rectangle below the number corresponding to your answer and beside the appropriate question number on the answer sheet. Be sure to erase completely any answers you wish to change.

Please do not try to guess the answers to questions you do not know the answers for.

Do not rush since you will be allowed enough time to complete the test. Please do not make any marks on the test booklet.

1. Which of the following substances is an example of a compound?

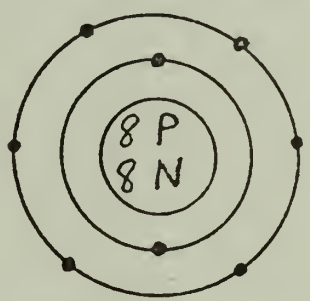
(1) air (2) silver (3) milkshake (4) salt

2. What is the atomic number of sodium?

(1) 11 (2) 19 (3) 23 (4) 39

3. Which atom is represented by the diagram below?

(1) sodium
(2) copper
(3) oxygen
(4) carbon



4. How does a sodium <u>ion</u> differ from a sodium <u>atom</u>?

(1) ion has fewer electrons
(2) ion has more neutrons
(3) ion has more electrons
(4) ion has a negative charge

5. Which of the following substances cannot be broken down by any chemical process?

(1) salt (2) air (3) gold (4) water

6. How many electrons are there in an atom of aluminum?

(1) 13 (2) 14 (3) 26 (4) 27

7. In general, atoms with 5 electrons in the outer shell belong to which of the following categories?

(1) metals
(2) nonmetals
(3) neither metals nor nonmetals
(4) both metals and nonmetals

8. How many atoms are there in a molecule of sulfuric acid, H_2SO_4?

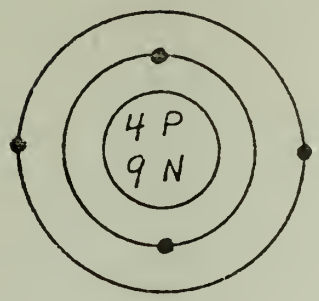
(1) 3 (2) 7 (3) 9 (4) 10

9. Which of the following substances is an example of a mixture?

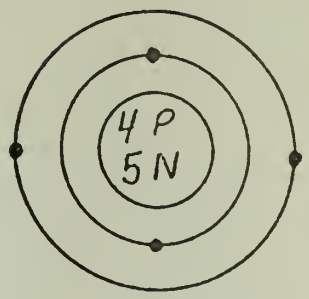
(1) air (2) gold (3) oxygen (4) water

10. Beryllium has an atomic number of 4 and an atomic weight of 9. Which of the diagrams below represents an atom of beryllium?

(1)

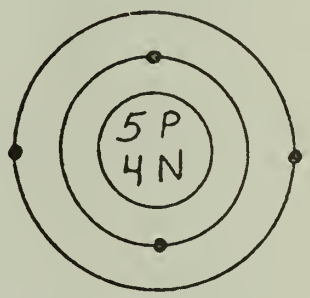


(3)



(2)

(4)



11. What is the oxidation or valence number of aluminum?

(1) +1 (2) +2 (3) +3 (4) +4

12. What is the correct chemical formula for aluminum hydroxide?

(1) AlOH (2) $AlOH_3$ (3) Al_3OH (4) $Al(OH)_3$

13. Which of the following substances is an example of a compound?

(1) hydrogen (2) oxygen (3) water (4) none of these

14. Which of the following particles is equal to the atomic number of an atom?

(1) number of neutrons
(2) number of electrons
(3) number of protons
(4) both (2) and (3) are correct

15. Which atom is represented by the diagram below?

(1) hydrogen
(2) helium
(3) lithium
(4) beryllium

16. Fluorine is an element with an atomic number of 9. How many electrons would there be in an ion of fluorine?

(1) 7 (2) 8 (3) 9 (4) 10

17. Which of the following substances is an element?

(1) carbon dioxide (2) zinc (3) cement (4) steel

18. How many neutrons are there in an atom of nitrogen?

(1) 7 (2) 14 (3) 28 (4) 31

19. The elements in group VII of the Periodic Table belong to which of the following categories?

(1) metals (2) nonmetals (3) inert gases (4) none of these

20. How many atoms are there in a molecule of calcium hydroxide, $Ca(OH)_2$?

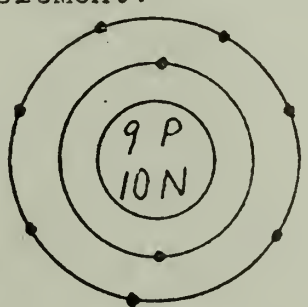
(1) 2 (2) 3 (3) 4 (4) 5

21. Which of the following substances is a mixture?

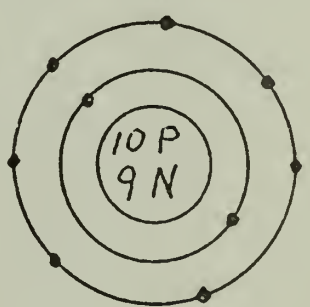
(1) salt (2) hydrogen (3) dirt (4) silver

22. The atomic number of a certain element is 9. Its atomic weight is 19. Which of the diagrams below represents an atom of this element?

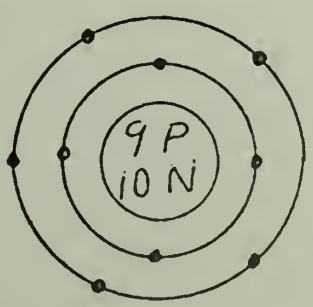
(1)



(3)



(2)



(4)

23. What is the oxidation or valence number of sodium?

(1) +1 (2) +2 (3) +3 (4) +4

24. What is the correct formula for sodium carbonate?

(1) NaCO (2) $NaCO_3$ (3) Na_2CO_3 (4) $Na(CO_3)_2$

25. Which of the following is a material made of more than one kind of atom in a definite proportion?

(1) element (2) mixture (3) compound (4) none of these

26. The atomic weight of an element is approximately equal to the weight of which of the following particles?

(1) protons (2) neutrons (3) nucleus (4) electrons

27. Which atom is represented by the diagram below?

(1) hydrogen
(2) lithium
(3) helium
(4) carbon

28. What is the charge on the <u>ion</u> formed when an atom gains an electron?

(1) +2 (2) -2 (3) +1 (4) -1

29. What do the "symbols" in the Periodic Table represent?

(1) elements (2) mixtures (3) radicals (4) compounds

30. How many protons are there in an atom which contains 12 neutrons and 10 electrons?

(1) 2 (2) 10 (3) 12 (4) 22

31. Which one of the following is a characteristic of metals?

(1) form negative ions
(2) combine with other metals
(3) form positive ions
(4) add electrons to the outer shell

32. How many oxygen atoms are there in $2H_2O$?

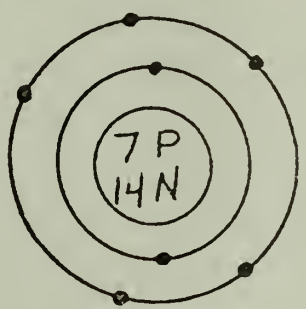
(1) 1 (2) 2 (3) 4 (4) 8

33. Which of the following is a mixture?

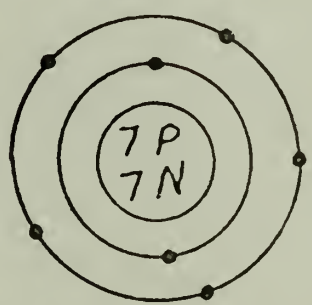
(1) carbon (2) sugar (3) sodium chloride (4) blood

34. Nitrogen has an atomic number of 7 and an atomic weight of 14. Which diagram represents an atom of nitrogen?

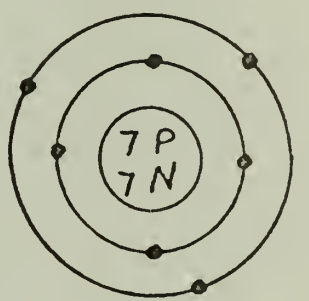
(1)



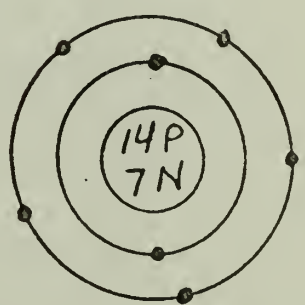
(2)



(3)



(4)



35. The oxidation or valence number of iodine is -1. What will an iodine atom do in a chemical reaction?

(1) gain 7 electrons
(2) gain 1 electron
(3) give up 7 electrons
(4) give up 1 electron

36. What is the correct formula for nickel chloride?

(1) NiCl (2) $NiCl_2$ (3) Ni_2Cl (4) $(NiCl)_2$

37. Which of the following is a substance that can only be broken down into two or more simpler substances by a chemical reaction?

(1) element (2) mixture (3) compound (4) none of these

38. What is the atomic weight of calcium?

(1) 20 (2) 40 (3) 45 (4) 60

39. Which atom is represented by the diagram below?

(1) hydrogen
(2) lithium
(3) helium
(4) carbon

40. Beryllium is an element with an atomic number of 4. How many electrons would there be in an ion of beryllium?

(1) 2 (2) 4 (3) 6 (4) 9

41. Which one of the following is not an element?

(1) iron (2) lithium (3) helium (4) none of these

42. How many protons, neutrons and electrons are there in an atom of an element with an atomic weight of 30 and an atomic number of 12?

(1) 12p 18n 12e
(2) 12p 12n 18e
(3) 30p 12n 30e
(4) 18p 30n 12e

43. In general, atoms with 2 electrons in the outer shell belong to which of the following categories?

(1) metals
(2) nonmetals
(3) neither metals nor nonmetals
(4) both metals and nonmetals

44. How many atoms are there in a molecule of $(NH_4)_4SiO_4$?

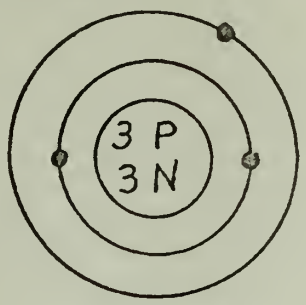
(1) 2 (2) 12 (3) 14 (4) 25

45. Which one of the following is not a mixture?

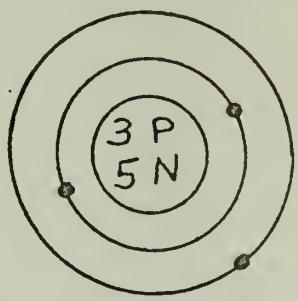
(1) steel (2) cake (3) milk (4) oil

46. The atomic number of an imaginary element is 3. Its atomic weight is 5. Which of the diagrams below represents an atom of this element?

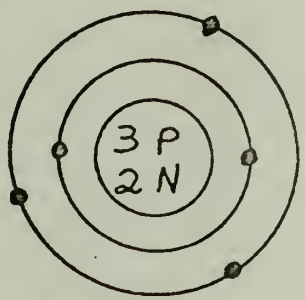
(1)



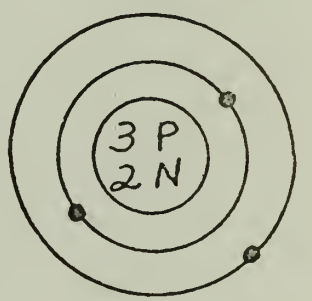
(3)



(2)



(4)



47. The oxidation or valence number of calcium is +2. What will a calcium atom do in a chemical reaction?

(1) gain 6 electrons
(2) gain 2 electrons
(3) give away 6 electrons
(4) give away 2 electrons

48. What is the correct formula for sodium silicate?

(1) NaSiO (2) Na_2SiO_4 (3) Na_4SiO_4 (4) Na_6SiO_4

Structure of Matter

Module Test, Form B

This test is designed to determine your knowledge of some selected science concepts in the area of the "Structure of Matter." Before beginning the test, print your name in the boxes provided on the answer sheet. In the space marked "TEST" on the answer sheet, write the letter "B". Be sure to use a soft-lead pencil for all marks on the answer sheet.

In completing the test, remember to blacken in the rectangle below the number corresponding to your answer and beside the appropriate question number on the answer sheet. Be sure to erase completely any answers you wish to change.

Please do not try to guess the answers to questions you do not know the answers for.

Do not rush since you will be allowed enough time to complete the test. Please do not make any marks on the test booklet.

1. Which of the following substances is a compound?

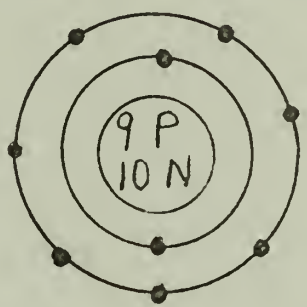
   (1) sodium (2) mercury (3) tin (4) amido chloride

2. What is the atomic number of zinc?

   (1) 20 (2) 30 (3) 45 (4) 65

3. What atom is represented by the diagram below?

   (1) oxygen
   (2) fluorine
   (3) potassium
   (4) none of these



4. How does an oxygen <u>ion</u> differ from an oxygen <u>atom</u>?

   (1) ion has fewer electrons (3) ion has more electrons
   (2) ion has more protons (4) none of these

5. Which of the following substances cannot be broken down by a chemical reaction?

   (1) water (2) glass (3) dirt (4) potassium

6. How many electrons are there in an atom of iodine?

   (1) 51 (2) 73 (3) 77 (4) none of these

7. In general, atoms with 1 electron in the outer shell belong to which of the following categories?

   (1) metals (3) neither metals nor nonmetals
   (2) nonmetals (4) both metals and nonmetals

8. How many atoms are there in a molecule of NH_4OH?

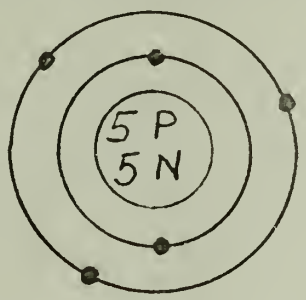
   (1) 1 (2) 4 (3) 6 (4) 7

9. Which of the following is a mixture?

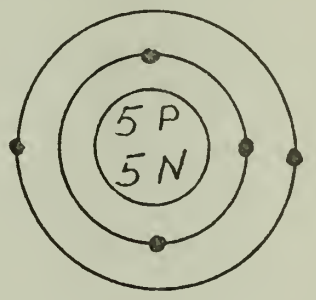
(1) sugar (2) steam (3) carbon (4) butter

10. Boron has an atomic number of 5 and an atomic weight of 10. Which of the following diagrams below represents an atom of boron?

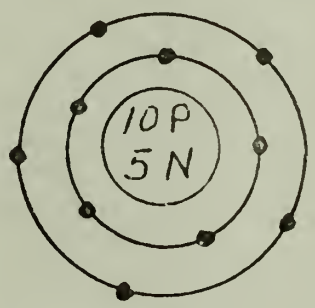
(1)



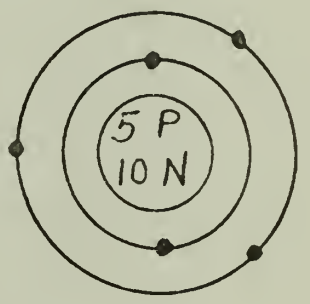
(3)



(2)



(4)



11. What is the oxidation or valence number of aluminum?

(1) +1 (2) +2 (3) +3 (4) +4

12. What is the correct formula for magnesium sulfate?

(1) $MgSO_4$ (2) Mg_2SO_4 (3) $Mg(SO_4)_2$ (4) $Mg(SO_4)_3$

13. Which of the following is an example of a compound?

(1) lead (2) sodium chloride (3) ammonia (4) lithium

14. Which element has an atomic number of 19?

(1) fluorine (2) carbon (3) oxygen (4) potassium

15. Which atom is represented by the diagram below?

(1) iron
(2) cobalt
(3) cadmium
(4) lithium

16. How many electrons should there be in an <u>ion</u> of rubidium?

(1) 36 (2) 37 (3) 44 (4) none of these

17. Which of the following cannot be broken down by a chemical reaction?

(1) milk (2) sand (3) carbon (4) steel

18. How many neutrons are there in an atom of neon?

(1) 7 (2) 10 (3) 14 (4) 20

19. Elements in group VI of the Periodic Table belong to which of the following categories?

(1) metals (2) nonmetals (3) inert gases (4) none of these

20. How many atoms are there in a molecule of $Ca(ClO_3)_2$?

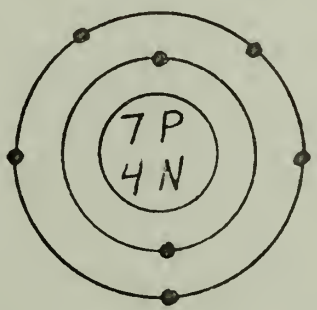
(1) 10 (2) 9 (3) 6 (4) 5

21. Which of the following substances is an example of a mixture?

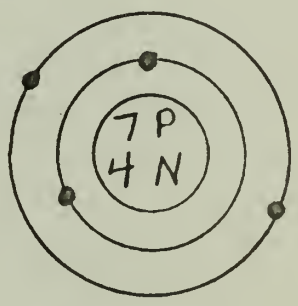
(1) neon (2) lard (3) copper (4) zinc

22. The atomic number of a certain element is 7. Its atomic weight is 11. Which of the diagrams below represents an atom of this element?

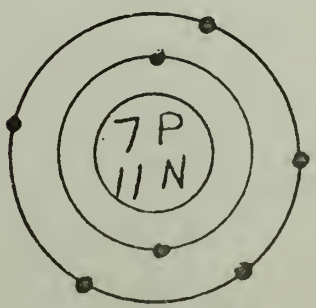
(1)



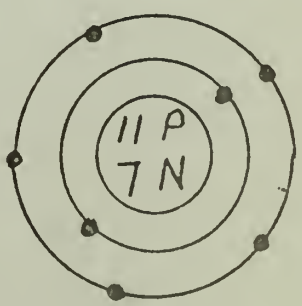
(3)



(2)



(4)



23. What is the usual oxidation or valence number of nitrogen?

(1) +3 (2) -3 (3) +5 (4) -5

24. What is the correct formula for ammonium bicarbonate?

(1) NH_3HCO_3 (2) NH_4CO_3 (3) $(NH_4)_2CO_4$ (4) NH_4HCO_3

25. Which of the following is an example of a compound?

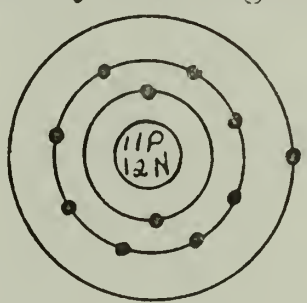
(1) iron (2) phosphorus (3) glass (4) oxygen

26. What element has an atomic weight of 12?

(1) carbon (2) magnesium (3) iron (4) steel

27. Which atom is represented by the diagram below?

(1) calcium
(2) vanadium
(3) sodium
(4) magnesium



28. How does a bromine <u>ion</u> differ from a bromine <u>atom</u>?

(1) ion has fewer electrons (3) ion has more electrons
(2) ion has more neutrons (4) ion has fewer protons

29. Which of the following is <u>not</u> an example of an element?

(1) ammonia (2) gold (3) oxygen (4) phosphorus

30. Which particle of the atom has an electrical charge of +1?

(1) proton (2) electron (3) neutron (4) none of these

31. Which of the following is a characteristic of nonmetals?

(1) form negative ions (3) form positive ions
(2) combine with nonmetals (4) give up electrons

32. How many atoms are there in anhydrous aluminum chloride, $2Al(H_2O)_6Cl_3$?

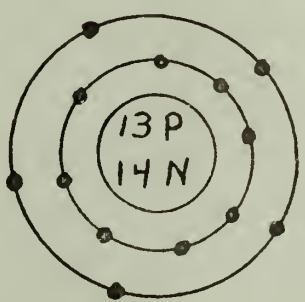
(1) 13 (2) 18 (3) 30 (4) 44

33. Which of the following is an example of a mixture?

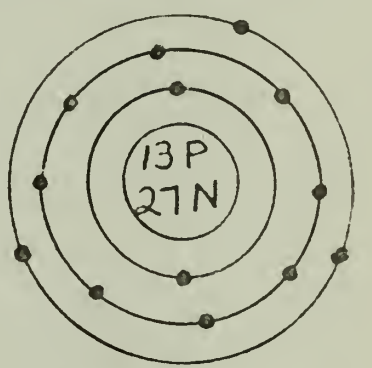
(1) silicon (2) hydrogen (3) cream (4) none of these

34. Aluminum has an atomic number of 13 and an atomic weight of 27. Which of the diagrams below represents an atom of aluminum?

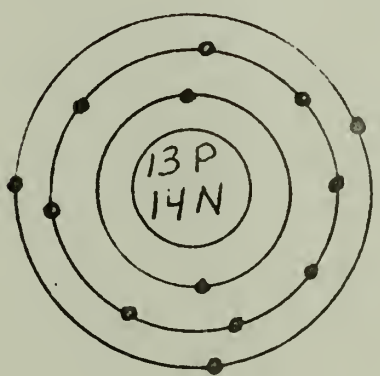
(1)



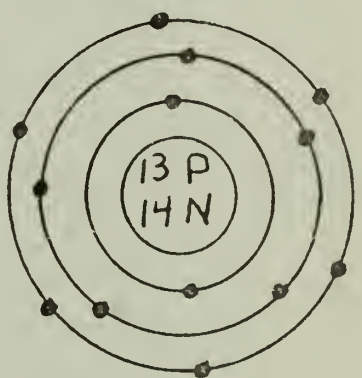
(3)



(2)



(4)



35. The oxidation or valence number of lithium is +1. What will a lithium atom do in a chemical reaction?

(1) gain 7 electrons
(2) gain 1 electron
(3) give up 7 electrons
(4) give up 1 electron

36. Which is the correct chemical formula for barium chlorate?

(1) $Ba(ClO_3)$ (2) $Ba(ClO_3)_2$ (3) $Ba_2(ClO_3)$ (4) none of these

37. Which of the following substances is a compound?

(1) ice (2) potassium (3) silicon (4) pluto

38. What is the atomic weight of gold?

(1) 100 (2) 197 (3) 207 (4) 212

39. What atom is represented by the diagram below?

(1) silicon
(2) nickel
(3) iron
(4) aluminum

40. What is the charge on the ion formed when an atom gains 2 electrons?

(1) 0 (2) +2 (3) -2 (4) none of these

41. Which one of the following is <u>not</u> an element?

(1) silicon (2) uranium (3) nobelium (4) none of these

42. How many protons, neutrons and electrons are there in an atom with an atomic weight of 42 and an atomic number of 17?

(1) 17p 42n 17e
(2) 25p 17n 25e
(3) 17p 25n 17e
(4) 42p 17n 25e

43. In general, atoms with 7 electrons in the outer shell belong to which of the following categories?

(1) metals (2) nonmetals (3) inert gases (4) none of these

44. How many atoms are there in a molecule of $Pb(C_2H_3O_2)_4$?

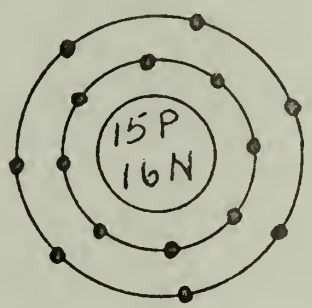
(1) 5 (2) 11 (3) 29 (4) 32

45. Which of the following is an example of a mixture?

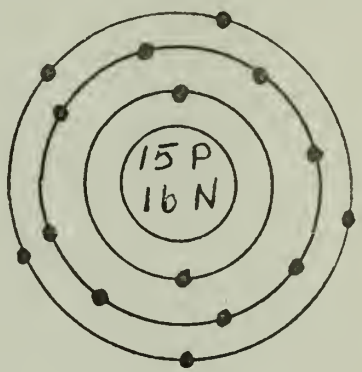
(1) iodine (2) mercury (3) glass (4) none of these

46. The atomic number of an imaginary element is 15. Its atomic weight is 31. Which of the diagrams below represents an atom of this element?

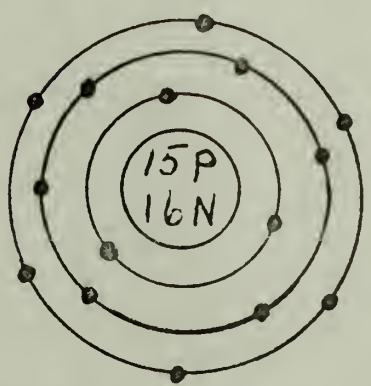
(1)



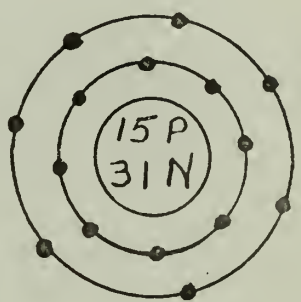
(3)



(2)



(4)



47. The oxidation or valence number of sulfur is -2. What will a sulfur atom do in a chemical reaction?

(1) gain 6 electrons
(2) gain 2 electrons
(3) give up 6 electrons
(4) give up 2 electrons

48. What is the correct chemical formula for aluminum sulfate?

(1) $AlSO_4$ (2) $Al_2(SO_4)$ (3) $Al_2(SO_4)_3$ (4) $Al(S)_3)O_4$

Structure of Matter

Module Test, Form C

This test is designed to determine your knowledge of some selected science concepts in the area of the "Structure of Matter." Before beginning the test, print your name in the boxes provided on the answer sheet. In the space marked "TEST" on the answer sheet, write the letter "C". Be sure to use a soft-lead pencil for all marks on the answer sheet.

In completing the test, remember to blacken in the rectangle below the number corresponding to your answer and beside the appropriate question number on the answer sheet. Be sure to erase completely any answers you wish to change.

Please do not try to guess the answers to questions you do not know the answers for.

Do not rush since you will be allowed enough time to complete the test. Please do not make any marks on the test booklet.

1. Which of the following substances is a compound?

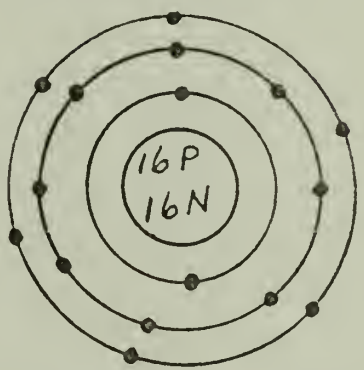
   (1) hydrogen (2) dirt (3) gold (4) water

2. What is the atomic number of chlorine?

   (1) 17 (2) 13 (3) 35 (4) 36

3. Which atom is represented by the diagram below?

   (1) sulfur
   (2) sodium
   (3) oxygen
   (4) chlorine



4. How does a fluorine <u>ion</u> differ from a fluorine <u>atom</u>?

   (1) ion has more electrons
   (2) ion has more neutrons
   (3) ion has fewer electrons
   (4) ion has a positive charge

5. Which of the following substances cannot be broken down by a chemical reaction?

   (1) water (2) salt (3) air (4) iron

6. How many electrons are there in an atom of sodium?

   (1) 11 (2) 12 (3) 19 (4) 23

7. In general, atoms with 3 electrons in the outer shell belong to which of the following categories?

   (1) nonmetals (2) inert gases (3) metals (4) none of these

8. How many atoms are there in a molecule of calcium phosphate, $Ca_3(PO_4)_2$?

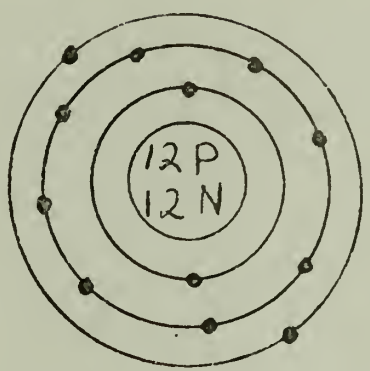
   (1) 5 (2) 9 (3) 10 (4) 13

9. Which of the following substances is an example of a mixture?

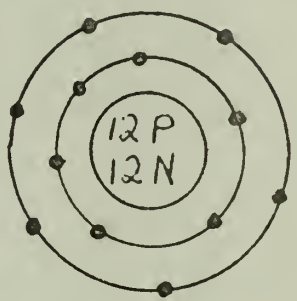
(1) hydrogen (2) dirt (3) gold (4) water

10. Magnesium has an atomic number of 12 and an atomic weight of 24. Which of the diagrams below represents an atom of magnesium?

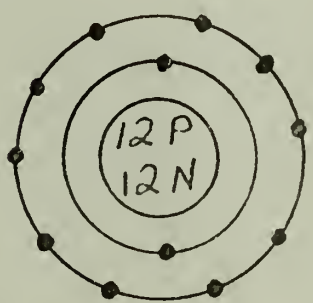
(1)



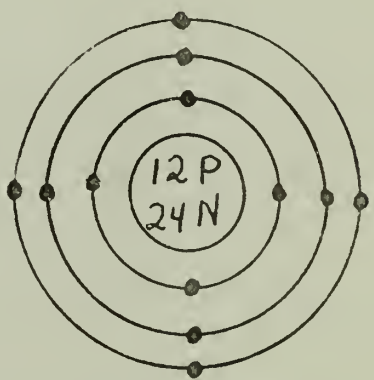
(3)



(2)



(4)



11. What is the usual oxidation or valence number of carbon?

(1) +1 (2) +2 (3) +3 (4) +4

12. What is the correct chemical formula for sodium phosphate?

(1) $NaPO_4$ (2) Na_2PO_4 (3) Na_3PO_4 (4) $Na_2(PO_4)_2$

13. Which of the following substances is a compound?

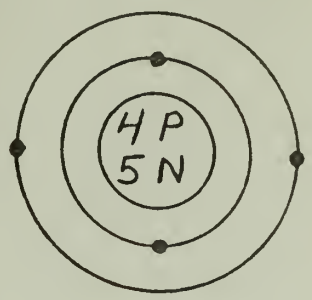
(1) oxygen (2) sugar (3) iron (4) brass

14. What is the atomic weight of an element containing 10 protons, 15 neutrons, and 10 electrons?

(1) 10 (2) 15 (3) 25 (4) 35

15. What atom is represented by the diagram below?

(1) helium
(2) beryllium
(3) carbon
(4) hydrogen



16. Calcium is an element with an atomic number of 20. How many electrons would there be in an ion of calcium?

(1) 17 (2) 18 (3) 20 (4) 40

17. Which of the following substances is an element?

(1) sugar (2) phosphorus (3) carbon monoxide (4) milk

18. How many neutrons are there in an atom of lead?

(1) 82 (2) 207 (3) 289 (4) none of these

19. The elements in group II of the Periodic Table belong to which of the following categories?

(1) metals (2) nonmetals (3) inert gases (4) none of these

20. How many atoms are there in a molecule of ammonium phosphate, $(NH_4)_3PO_4$?

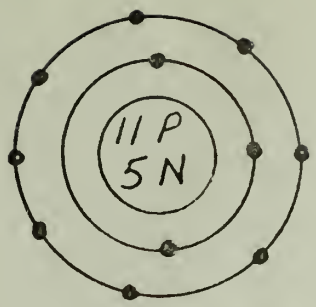
(1) 6 (2) 7 (3) 13 (4) 20

21. Which of the following substances is an example of a mixture?

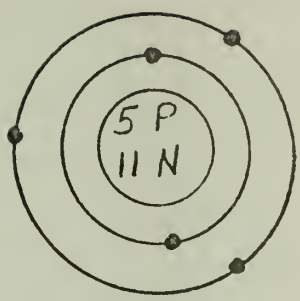
(1) air (2) silver (3) water (4) salt

22. The atomic number of a certain element is 5. Its atomic weight is 11. Which of the diagrams at the top of the next page represents an atom of this element?

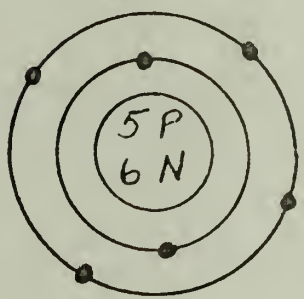
(1)



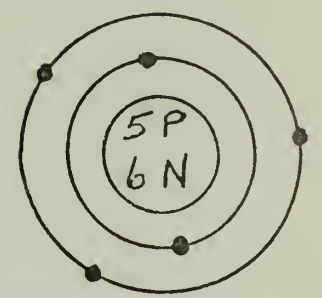
(2)



(3)



(4)



23. What is the usual oxidation or valence number of calcium?

(1) +1 (2) +2 (3) +3 (4) +4

24. What is the correct chemical formula for calcium nitrate?

(1) $CaNO_3$ (2) Ca_2NO_3 (3) $Ca(NO_3)_3$ (4) $Ca(NO_3)_2$

25. Which of the following is a compound?

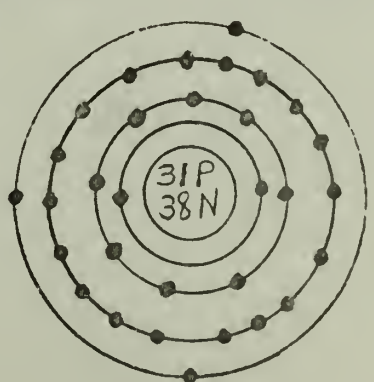
(1) sodium (2) carbon dioxide (3) mud (4) iron

26. What is the atomic number of magnesium?

(1) 6 (2) 12 (3) 18 (4) 24

27. Which atom is represented by the diagram below?

(1) zinc
(2) nickel
(3) cerium
(4) none of these



28. What is the charge of the ion formed when an atom loses an electron?

(1) -2 (2) -1 (3) +1 (4) +2

29. Which of the following substances cannot be broken down into something simpler by a chemical reaction?

(1) element (2) mixture (3) compound (4) all of these

30. How many protons are there in an atom of nitrogen?

(1) 7 (2) 14 (3) 28 (4) 31

31. Listed below are four examples of electron configurations of different atoms. Which configuration represents an atom of a metal?

(1) 2-8-6 (2) 2-8-8-7 (3) 2-8-18-18-5 (4) 2-8-3

32. How many hydrogen atoms are there in $2H_2O$?

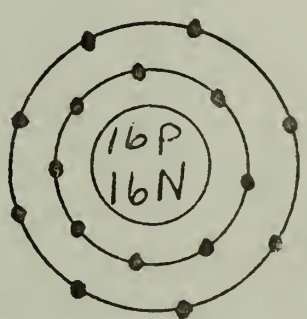
(1) 1 (2) 2 (3) 4 (4) 8

33. Which of the following substances consists of several materials in which the proportions of materials in the substance may vary?

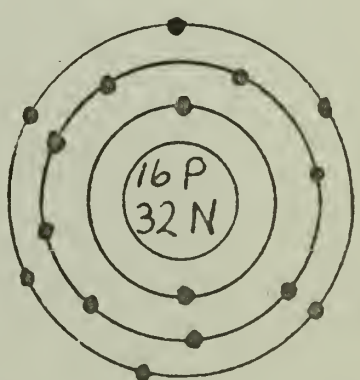
(1) element (2) mixture (3) compound (4) none of these

34. Sulfur has an atomic number of 16 and an atomic weight of 32. Which of the diagrams represents an atom of sulfur?

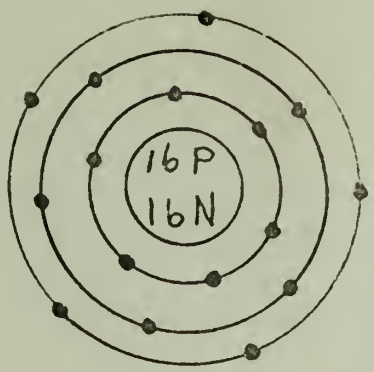
(1)



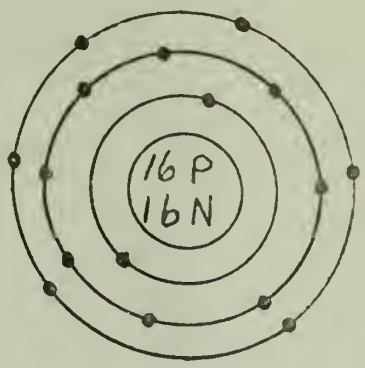
(3)



(2)



(4)



35. The oxidation or valence number of aluminum is +3. What will an aluminum atom do in a chemical reaction?

(1) gain 5 electrons
(2) gain 3 protons
(3) give up 5 electrons
(4) give up 3 electrons

36. What is the correct chemical formula for potassium sulfate?

(1) K_2SO_4 (2) $K(SO_4)_2$ (3) KSO_4 (4) K_3SO_4

37. Which of the following is a compound?

(1) iron (2) nitrogen (3) hydrogen peroxide (4) soup

38. What is the atomic weight of an element containing 13 protons, 14 neutrons and 13 electrons?

(1) 13 (2) 26 (3) 27 (4) 40

39. Which atom is represented by the diagram below?

(1) carbon
(2) nitrogen
(3) oxygen
(4) none of these

40. How does a potassium <u>ion</u> differ from a potassium atom?

(1) ion has more neutrons
(2) ion has more electrons
(3) ion has fewer electrons
(4) ion has a negative charge

41. Which of the following is made up of only one kind of atom?

(1) element (2) mixture (3) compound (4) all of these

42. How many protons, neutrons, and electrons are there in an atom of an element with an atomic weight of 40 and an atomic number of 20?

(1) 20p 20n 40e
(2) 20p 40n 20e
(3) 40p 20n 20e
(4) 20p 20n 20e

43. The elements in group VI of the Periodic Table belong to which of the following categories?

(1) metals (2) nonmetals (3) both (1) and (2) (4) none of these

44. How many atoms are there in a molecule of $Be_3(PO_4)_2$?

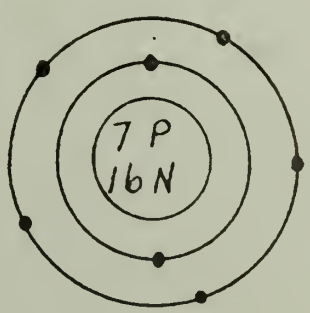
(1) 1 (2) 5 (3) 13 (4) 16

45. An unknown substance appears as a light green powder. When a student places the powder in water, he finds that some of the powder dissolves, forming a green solution. However, the remainder of the substance will not dissolve and settles to the bottom as a white powder. If a chemical reaction had not occurred, what was the substance?

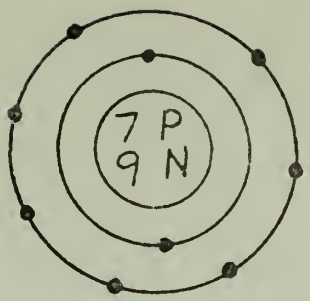
(1) element (2) compound (3) mixture (4) inert material

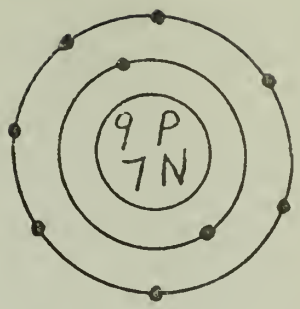
46. The atomic number of a certain element is 7. Its atomic weight is 16. Which of the diagrams below and at the top of the next page represents this element?

(1)



(3)

(2) 

(4) 

47. What is the usual oxidation or valence number of strontium?

(1) +1 (2) +2 (3) +3 (4) +4

48. What is the correct chemical formula for aluminum nitrate?

(1) $AlNO_3$ (2) $Al(NO_3)_2$ (3) $Al(NO_3)_3$ (4) $Al_2(NO_3)_3$