Statistical methods for the determination of content validity

Olagunju, Amos Omotayo, Ed.D.

The University of North Carolina at Greensboro, 1987

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

STATISTICAL METHODS FOR THE DETERMINATION OF CONTENT VALIDITY

by

Amos Omotayo Olagunju

A Dissertation Submitted to
the Faculty of the Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment of the Requirements
for the Doctorate in Educational Administration
with a Concentration in Research and Evaluation

Greensboro
1987

Approved by

Dissertation Advisor

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of the Graduate School at The University of North Carolina at Greensboro.

Dissertation Advisor Richard M. Jaeger

Committee Members Linda J. DeGuire

Joseph E. Bryson

Rita O'Sullivan

John Christian Busch

29 May 1987
Date of Acceptance by Committee

1 May 1987
Date of Final Oral Examination

OLAGUNJU, AMOS O., Ed.D. Statistical Methods for the Determination of Content Validity. (1987)
Directed by Dr. Richard M. Jaeger. 265 pp.

Determination of the content validity of standardized tests is a central problem at all levels of education and in the professions. The problem investigated in this research was that of developing operational rules and statistical guidelines for estimating the content validity of standardized achievement tests. Rating and matching techniques were examined as alternative methods for eliciting judgments about the content validity of test items. These methods of eliciting judgments and newly-developed quantitative indices of the content validity of test items and tests were used to validate the Greensboro Public Schools' Mathematics Promotion Standard tests for Grade 4. Rating items on the basis of their "Overall" quality and a matching method were found to be more accurate procedures for eliciting judgments than was a method that required rating items on a given set of dimensions of judgment; however, judges were more consistent in their judgments of test items over domains with the latter method. The assumptions and implications of using the newly-developed indices of content validity with tests designed for criterion-referenced interpretations were discussed.

ACKNOWLEDGMENTS

The work described herein originated through a series of dialogues with Professor Richard Jaeger. I would like to express my sincere gratitude to Dr. R. M. Jaeger for his constructive critical commentaries and valuable suggestions towards the succesful outcome of this dissertation.

My thanks go to Dr. Linda DeGuire and Dr. Rita O'Sullivan for their contributions during the development and validation of the research instruments. Thank you, Dr. Rita O'Sullivan, for providing the inservice training for sampled Greensboro Public Schools' teachers.

I gratefully acknowledge the support provided by the Greensboro Public Schools' personnel. Sampled teachers were sponsored and duplicate copies of the research instruments were made at the expense of the Greensboro Public Schools. Thank you, Mr. Ron Hann and Mr. Daniel Watkins, for making all arrangements with the elementary school mathematics teachers. Mr. Ron Hann also validated the research instruments for appropriateness and completeness. I would like to thank all Greensboro Public Schools' teachers who participated in this project.

Dr. Joseph Bryson, I do not know how to thank you enough for your help and encouragement over the years. Thank you, Dr. John Busch, for introducing me to this doctoral program and for your valuable comments at various stages of this dissertation. I owe much tribute to Daddy S. O. Olagunju for his prayerful support throughout this work.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

CHAPTER 1

INTRODUCTION

1.1 Rationale

Standardized tests play an important role in American life. Standardized tests are used to monitor individual progress through objectives-based instructional programs, to evaluate educational and social action programs, to diagnose learning deficiencies, and to assess competence on certification and licensing examinations. Today, an increasing emphasis on mastery, proficiency, and competency is permeating all levels of education and the professions, in particular, medicine and the allied health fields. Every day, many individuals at every level of education are administered standardized tests in diverse settings (e.g., schools, businesses, and the military). The usefulness of these standardized tests depends directly on the validity of the descriptions, decisions, and interpretations that result from the test scores. Unfortunately, methods for assessing the validity of the content of standardized tests are still being scrutinized. Thus, it is useful to establish operational rules and statistical guidelines for determining the content validity of standardized tests.

Recent attention and confusion in the courtrooms of the nation call for consensus on the term "content validity" within educational and psychological professions. The current use of the term "content validity" in the literature often leads to ambiguity in the meaning of the term. Presently, content validation of standardized tests in education and the professions relies primarily on expert judgment. Unfortunately, numerical figures representing content validity are not usually provided. In order to explicate the meaning of the term "content validity," it is useful to establish methods for eliciting judgments of content validity of test items and to provide quantitative indices for the determination of the content validity of individual test items and of tests as a whole.

There are two predominant kinds of standardized test interpretations: norm-referenced and criterion-referenced. Norm-referenced interpretations rely mainly on the relative status of an examinee's performance in relation to the performances of those in a normative group. Criterion-referenced interpretations describe an individual's performance in terms of what he or she can and cannot do, irrespective of the performance of other examinees. In a norm-referenced world, it is easy to consult a norms table when we are only interested in what a person's score means on a test, compared to those of other people. But if we want to know what a person with a particular score "can or can't do," norms tables don't offer much solace. In order to determine what an examinee's score really means, the need exists to develop test items that (a) represent a relatively homogeneous collection of instances for the examinee to exhibit the tested skill, and (b) are described well enough that we really know what the items are trying to measure.

Content validation is of primary interest when criterion-referenced interpretations of test scores are to be made. This research study has examined quantitative indices of content validity in the context of specially constructed tests designed for criterion-referenced interpretation.

1.2 Purpose

The primary aim of this resesearch study was to develop operational rules and statistical guidelines for estimating the content validity of standardized tests. The specific goals established for the research are as follows.

1) To examine several alternative methods for eliciting judgments of the content validity of test items.
2) To develop some quantitative indices of the content validity of test items.
3) To demonstrate the usefulness of these indices of content validity.

1.3 Theoretical Framework

Traditionally, content validation seeks to accumulate evidence to support the assertion that a test samples the domain of subject matter about which inferences are to be made. To demonstrate the content validity of a test, it is necessary to show that the behaviors tested constitute a representative sample of behaviors to be exhibited in a desired performance domain. Definition of the performance domain, the users' objectives, and methods of sampling are critical to claims of content validity. If the purpose of a content validity study is to demonstrate what Cronbach (1971) has characterized as "showing how well the content of the test samples the class of situations or subject matter about which conclusions were drawn," it is essential that the study be solidly grounded in a body of relevant theory.

The criterion-referenced test is one of the examples of measures that require content validation. A major part of the criterion-referenced test plan is an outline of content domains for the test which is to be constructed. Since content validity depends on a rational appeal to adequate coverage of important content, an explicit outline of theoretical content domains is a useful basis for discussing content validity. Such an outline should describe the types of test items, state the approximate number of items to be selected from each theoretical domain and each objective of the test, and provide examples of the types of test items to be used. In order to explicate the meaning of content validity, it is important that the theoretical domains be adequately sampled by the test items. Thus the method for sampling items must be based on a theory.

To the extent that a measure possessing construct validity is an operational definition of a theoretical domain (Cronbach and Meehl, 1955), it seems appropriate that an indication of the degree to which an item reflects the particular theoretical domain would be valuable in item selection and content validation. One way to gather information for a content validity study might be to teach a theoretical population of judges about the nature of the theoretical domain under consideration, and have them rate the degree to which the various items reflect this theoretical domain. These content validity scale values, and their dispersions, could then be used to evaluate the degree to which the theory is reflected in the measure, as well as the extent to which the population of judges agree upon the content validity of the item. Such an approach would offer an explicit means of determining the degree of relation between the theory and the measure, which would appear to be an important aspect of the establishment of content validity.

1.4 Research Questions

The determination of content validity generally involves consideration of elicited examinee behaviors and three features of test items: (1) the extent to which each test item actually measures some aspect of the content included in a domain specification, (2) representativeness of the test items, and (3) technical quality of the test items. The present research has investigated the following questions:

(1) To what extent are the results of content validation dependent on procedures used for eliciting judgments?

(2) Does the degree of accuracy of judges (in defining the content validity of items) vary with the proportion of "bad/good" items presented to them?

(3) Do various indices of content validity increase or decrease as the proportion of "bad" items provided to judges increases?

1.5 Limitations of the Study

Test performance must be interpreted in terms of the particular items included. Usually, test items are written from domain specifications. When the domain of items measuring an objective is unclear, only the weakest form of test interpretation is possible (Popham, 1974). Because of our special interest in working with well-written domain specifications, the present study has investigated the content validity of test items from the field of mathematics. Since clear domain specifications are more easily produced in mathematics and the physical sciences, we recognize that our results might not generalize to many other content areas of interest to educators.

1.6 Review of the Literature

1.6.1 Characterization and Requirements of Content Validity

Validation of the content of standardized tests has been an ever present, driving concern for psychological and educational researchers and practitioners. Content validity is usually characterized as follows:

> "Content validity is indicated by a description of the universe of items from which selection was made, including a description of the selection" (APA, 1954, p.216), or "Content validity is evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions were drawn" (APA, 1954, p.213; Cronbach, 1971, p.444).

Cronbach (1971) explains content validity as follows:

> Content validity has to do with the test as a set of stimuli and as a set of observing conditions. The measuring procedure is specified in terms of a class of stimuli, an injunction to the subject that defines his task (i.e., what he is to try to do with the stimuli), and a set of rules for observing the performance and reducing it to a score" (p.452).

Cronbach (1971) asserts that "if the content is validly selected, the test is content-valid for persons of all kinds" (p.453). Arguing that construct validity is an important consideration for the validity of test scores, Messick (1975) cautions that interpretations claiming content validity in this official sense should be carefully restricted to task language.

Educational researchers have expressed diverse opinions about content validity. For instance, Guion (1977) has provided a number of reasons for his discontent about content validity. Of importance among his reasons is his conclusion that "judgments of content validity have been too swiftly, glibly and easily reached in accepting tests that otherwise would never be deemed acceptable" (p.8). Of particular interest to our proposed research work are the five minimal conditions that Guion proposed as a measure of content validity. These conditions are as follows:

(1) The content domain ought to include "behavior with a generally accepted meaning" (p.6);

(2) The definition of the domain should be specified unambiguously;

(3) The domain ought to be relevant to the intentions of the measurement;

(4) The measure must be reliable; and

(5) "Qualified judges must agree that the domain has been adequately sampled" (p.7).

In partial disagreement with Guion's notion of content validity, Linn (1980) argued that concerns for "relevance" and "meaning" go far beyond content validation and involve constructs or external criteria and require other kinds of validity evidence. According to Linn (1980) content validity is derived only from a domain definition and representativeness. Convincingly, Linn argued that "judgments about sampling adequacy or representativeness require clarity of definition of the item domain. Indeed, domain definition provides the key to item generation and content validity" (p.549).

Recognizing that content validation encompasses a series of activities which take place after the initial test has been developed, Crocker and Algina (1986) recommended four major content validation tasks. These tasks are as follows:

(1) The performance domain must be defined;

(2) A panel of qualified experts in the content domain must be selected;

(3) A well-defined procedure for matching test items to the performance domain must be provided; and

(4) The matching process must provide data that must be collected and analyzed.

As advocated by Crocker and Algina (1986), the content validation researcher should provide definitions of content validity terms rather than allowing judges to use idiosyncratic definitions.

1.6.2 Scope and Problems of Criterion-Referenced Test Validity

Since the term "criterion-referenced measurement" was applied to proficiency assessment by Glaser and Klaus (1962), numerous useful contributions to the criterion-referenced testing literature have been made (for reviews, see Hambleton, et al., 1978; Millman, 1974; Popham, 1978). Although some researchers consistently use the term "criterion-referenced test," it is not uncommon to find the terms "domain-referenced test," "proficiency test," "objectives-referenced test," "competency-based test" and "mastery test" used interchangeably in the literature.

Various comparisions of the descriptions of a criterion-referenced test suggest there is general agreement that the test is intended to reference an individual's score to a well-defined domain of behaviors (Hambleton, et al., 1978). Popham's (1978) definition most accurately reflects that conceptualization: "A criterion-referenced test is used to ascertain an individual's status with respect to a well-defined behavioral domain". An alternative conceptualization of criterion-referenced measurement derived from mastery learning theory (Mayo, 1970) is represented by the mastery test. In order to expedite individualized instruction, a mastery test can be used to classify students as masters and nonmasters of an objective.

Unfortunately, criterion-referenced test score validity remains an essentially unexplored topic. Only a few researchers have attempted to clarify the scope of the topic, to resolve any of the complex problems, or to offer practitioners guidelines for validating criterion-referenced test scores (Linn, 1979; Messick, 1975; Millman, 1974; Popham, 1975).

Many criterion-referenced test developers have argued that, to validate their tests and test scores, it is sufficient to assess their "content validity". Usually, this means that judgments are made regarding the match between the objectives to be measured by a test and the item content of that test. Indeed, an index of content validity should not vary in different samples of examinees or over time. Even though such an index is important, it is doubtful if it presently exists in the literature. Despite its stated importance, it cannot be argued that the nature of content validation studies with criterion-referenced tests is well understood. Guion (1977), for one, discusses many of the problems surrounding the topic. Guion (1977) and Messick (1975) prefer the term "content representativeness" to "content validity" because they are not convinced that content representativeness is a validity issue.

Empirical test validation procedures involve an examination of test item statistics, such as the difficulty index and discrimination index; they can be applied to criterion-referenced tests in much the same way as empirical procedures are applied in norm-referenced test development. Popham (1980), clearly identified three problems involved with the use of empirical test validation procedures. First, most empirical procedures depend upon the characteristics of the group of examinees and the effects of instruction. Second, there is a considerable risk of obtaining a nonrepresentative set of items from the domains measuring the objectives included in a test because item statistics are derived from empirical analyses of test data that are used to select items for a criterion-referenced test. Third, empirical techniques in many instances require pretest and posttest data on the same items even though pretest data are rarely collected. Despite these problems, empirical methods do have one important use in content validation. According to Rovinelli and Hambleton (1977), "In situations where the test constructor is interested in identifying aberrant items, not for elimination from the item pool but for correction, the use of an empirical approach to item validation should provide important information with regard to the assessment of item validity" (p.51). However, it seems more appropriate to establish the content validity of test items by seeking the opinions of content specialists on the (1) extent of match between the test items and the domains they are designed to measure, and (2) the degree to which the test items in a criterion-referenced test are representative of the domain of items specified in the domain specification.

1.6.3 Alternative Approaches to Domain Specification and Operational Definition

The general notion of validity of the content of criterion-referenced tests has been challenged on the grounds that it is difficult to talk about how well the test "samples" the subject matter or class of situations because there are no populations of items or testing conditions (Loevinger, 1965). Responding to this attack, Cronbach (1971) suggested that the important requirement is that the boundaries of the universe or domain be well-defined. Essentially, this is a requirement of operational definition which might be accomplished in a variety of ways such as by specifying categories of learning outcomes in a subject-matter area through task analysis or other means (Glaser and Nitko, 1971; Gagné, 1974).

If the proper domain of test items measuring an objective is clear, it is possible to select a representative sample of test items from that domain. Representative samples of test items measuring each objective included in a test are necessary to obtain unbiased estimates of examinee performance in the full domain of behaviors measuring each objective. If the proper domain of test items measuring an objective is not clear, it is impossible to select a representative sample of test items from that domain.

In recent years, it has been very popular to write instructional objectives in "behavioral" terms. Behavioral objectives are definitely better than no objectives at all. However, given their terse form, behavioral objectives leave too many decisions to the item writer. Fortunately, more useful behavioral objectives are "amplified" objectives. According to Millman (1974), "An amplified objective is an expanded statement of an educational goal which provides boundary specifications regarding testing situations, response alternatives, and criteria of correctness." Even though some ambiguity is still left in the domain definition, the additional guidelines introduced in an amplified objective help to define a domain of items.

Hively, Patterson, and Page (1968) have developed a scheme called an "item form". An item form can be used to generate a "universed-defined" test. With this scheme, a content area is conceptually analyzed into a hierarchical arrangement of item forms. An item form generates items with a fixed syntactical structure that contains one or more variable elements and defines a class of item sentences by specifying the replacement sets for the variable elements. This allows a specification of items for a test in advance. An item form scheme is a highly detailed set of rules for creating what is hoped to be homogeneous test items. Thus, tests sampling this explicitly defined universe of content can be constructed. Item forms have been found useful in an assessment of mathematics and science skills (Hively et al., 1968). However, an item form scheme can lead to too many item forms (Popham, 1980).

Another method for specifying a content domain is called "limited-focus" (Popham, 1978). While retaining the descriptive rigor of item forms, a limited focus strategy limits measurement focus to a smaller number of assessed behaviors. To conceptualize these behaviors so that they are of larger scale, important behaviors that subsume lesser, "en route" behaviors are developed. By using a limited focus measurement strategy, it is possible to create a small enough number of test descriptors so that item writers would attend to them (Popham, 1978).

In general terms, validity is concerned with the accuracy of estimates of universe scores (Kane, 1982). According to Kane, "validity involves the interpretation of the observed score as representative of some external property" (p.125). Kane clearly defined the concepts of "basic," "derived," "theoretical," attributes, and "operational definition." On the basis of some property, a basic attribute is a representation of an observed ordering. The constants in empirical laws stating relationships among the basic attributes are called derived attributes. Theoretical attributes involve a few postulates defining a theory underlying the basic and derived attributes. Operational definitions specify "the kind of observations that are to be used and the way in which numbers are derived from these observations" (structural rules), and "the range of conditions that may be tolerated for the various characteristics of observations" (selection rules), (p.128). Thus, if the content validity of criterion-referenced tests is operationally defined, the structural rules and the selection rules can provide an interpretation for the numbers assigned as values of the basic and derived attributes of content validity.

Kane (1982) also discussed the concept of "errors of measurement" in terms of classical test theory. By establishing the definition of an attribute for an object of measurement, its value (true score) is the expected value over all observations, and the expected value of the errors is zero. Though they indicate the accuracy of estimates of the true score for each object of measurement, object-specific error variances are difficult to estimate because they require repeated observations on each object of measurement (Kane, 1982). However, the average error variance over all objects of measurements is more widely used because it can be estimated with pairs of observations on each object of measurement. Thus if structural rules are established in the operational definition of content validity, it is possible to estimate the accuracy of the numbers that are derived from the observations in a content validity study.

1.6.4 Judgmental and Quantitative Content Validity Procedures

The use of judges to assess content congruence offers a promising method of assessing the content validity of criterion-referenced test items. For example, Rogers (1973) had three groups of undergraduates from psychology courses rate the degree to which each of 60 Personality Research

Form items reflected particular personality characteristics. The 60 items contained 20 items from each of three scales: Impulsivity, Autonomy, and Desirability. The first group of 54 students rated the desirability of all 60 items on a seven-point scale. The second group of 54 students received an Autonomy instructional set, and, on a seven-point scale, rated the degree to which each item reflected this particular characteristic. The third group of 54 students received Impulsivity rating instructions and, on a seven-point scale, rated the degree to which each of the 60 items reflected this characteristic. An analysis of variance was performed on the average ratings and the dispersions of the ratings. Results indicated that judges were able, through their ratings, to identify the scales to which the items belonged. Therefore, judgmental procedures might be effective in the determination of the validity of the various content components of criterion-referenced tests.

Recently, researchers are beginning to show interest in the statistical determination of content validity. Lawshe (1975) conceptualized the problem of validating a job performance test as that of identifying the segment of the total universe from which a job performance domain could be sampled and operationally defined. Lawshe operationally defined content validity as "the extent to which members of the Content Evaluation Panel perceive overlap between the test and the job performance domain" (p.566). The Content Evaluation Panel was composed of job incumbents and supervisors who judged whether or not the knowledge of a given bit of job information was relevant to the job performance domain. The consensus of the panel was quantified to yield a content validity ratio for each test item. Contending that content validation requires judgment as to the correspondence of abilities requisite for job success, Lawshe (1975) devised his content validity ratio as a direct linear transformation from the percentage of experts judging a skill measured by an item to be "essential".

Quantitative content validity techniques have been found useful in the development of behavioral rating scales for use as job-related criteria for selection validation (Distefano, Pryer, and Erffmeyer, 1983; Distefano, Pryer and Craig, 1980). Distefano, Pryer and Craig (1980) have demonstrated the use of Lawshe's (1975) content validity procedures to establish the job-relatedness of a post-training test criterion for psychiatric aides. The test consisted of 60 items that were designed to be representative of the content of a training program and of the work required of aides. A 60-statement questionnaire describing the specific type of knowledge required by an item on the aide test, was administered to a sample of 18 incumbent aides and 19 aide supervisors. The subjects evaluated each statement to make judgments on the importance of each particular job knowledge or skill to the job performance. The three rating response choices for each statement were "essential," "useful but not essential," and "not essential." By analyzing the judges' responses, the content validity ratio, (CVR = (2 x N_e- N_i)/N_i , where N_e is the number of judges indicating "essential" and N_i is the number of judges indicating judgment on the ith item, Lawshe (1975)) was calculated for each of the 60 items. According to Lawshe's content validity procedures, reponses to 41 of the 60 items yielded statistically significant CVR's (Ho: CVR = 0 against Ha: CVR > 0) and the mean CVR for the 60 items was significant, indicating significant content validity for the overall test. Note that the significance test for Lawshe's content validity ratio (CVR) for each test item is an approximation to the Binomial test. While the application of Lawshe's method provided significant quantitative evidence of the content validity of the aide test criterion, the method has not been applied to the validation of the content of criterion-referenced educational tests.

Matching individual items to a list of objectives has been recommended as a reasonable approach to content validation. Crocker and Algina (1986) for one proposed "percentages of items matched to objectives," "percentage of items matched to objectives with high importance ratings," and "percentage of objectives not assessed by any of the items on the test" as three indices of content validity. For meaningful interpretations, the first two indices require one hundred or more test items and the third index would be low whenever all test items match only one of many relevant objectives (Crocker and Algina, 1986). Another index of item-objective congruence has been developed by Rovinelli and Hambleton (1977). In the data collection procedure, content specialists are instructed to match items to each objective and for the appropriate objective to assign a +1 if an item measures the objective, 0 if the item is questionable as a measure of the objective, and -1 if the item definitely does not measure the objective. Given N objectives and n content specialists, the index of congruence I_{ik}, of the item i to objective k is given by $I_{ik} = (N \sum_{j=1}^{n} X_{ijk} - \sum_{i=1}^{N} \sum_{j=1}^{n} X_{ijk})/(2N-1)n$, where X_{ijk} is the jth content specialist's rating of item k on the ith objective. In the ideal situation, this statistic assumes that an item clearly matches one and only one objective in the set. According to Hambleton (1980), one major drawback of this approach is that "it is very time consuming." The practical use of this technique in content validation of educational achievement tests is questionable. For what use is it to arbitrarily separate test items from the specific objective for which they have been designed?

Of particular relevance to the validation of the content of test items is a statistical index of content validity that Aiken (1980) invented. Aiken's index of content validity assumes that each of N judges will rate a single item on a c-category ordinal rating scale. The content validity index is then defined as the sum of weighted categories of ratings by the N judges (Aiken-V $= \sum_{i=1}^{c-1}((i \text{ x } N_i \ / \ N(c-1))$, where c is the number of categories on an ordinal rating scale, and i is the weight assigned to the N_i ratings in the highest (ith) category). Aiken (1980) has also developed a procedure for assessing the probability that the observed categories of ratings have occurred at random. The procedure employs the multinomial probability distribution for small samples and normal curve probability estimates for large samples. If judgments of the content validity of criterion-referenced test items can be made on ordinal rating scales, Aiken's index might be found useful as a measure of content validity for criterion-referenced test interpretations. However, the usefulness of Aiken's index as a measure of content validity for criterion-referenced test interpretations is not of interest to the present discussion because the index assumes an ordinal rating scale. Of what use is a "bad" test item that is rated as "fair" or "partially acceptable?"

Distefano, Pryer and Craig (1983) have found Lawshe's (1975) and Aiken's (1980) quantitative content validity procedures useful in the development of a job-related behavioral rating scale criterion for entry-level psychiatric aides. Eighty-three work behavior items were developed and a panel of aides (20 aides and 18 aide supervisors with work experience that enabled them to give informed opinions about the work behaviors required of aides after completing a basic aide training program) rated each item as either "essential," "useful, but not essential," or "not necessary" in the performance of the job. The consensus of the panel was quantified to yield a content validity ratio and Aiken validity coefficient. Seventy-eight of the 83 items were found to be significantly job-relevant using the computational procedures of both Lawshe and Aiken. While Laswhe's and Aiken's quantitative indices provided evidence of the content validity of the 83-work-behavior item pool for entry-level psychiatric aides, it is not evident that the indices will provide quantitative content validity evidence for criterion-referenced educational test interpretations.

Arguing that current validation procedures which concentrate on an item analysis are insensitive, Jones and Szatrowski (1983) have suggested several criteria related to individual exposure-nonexposure responses for consideration in content validity studies. The criteria involved validation of a test for a population or subpopulation by considering minimum exposure responses to each topic covered by the test. Jones and Szatrowski's criteria for validation of tests has nothing to do with content validity since analysis of exposure-nonexposure response sequences has more to do with the concept of "instructional validity." In content validity studies, a researcher is interested in the extent to which test items sample the domain of behaviors about which inferences are to be made, whereas in an instructional validity study, the question of interest is, what is the likelihood that a population would have been exposed to test samples (i.e., to what extent does the instruction sample the test?). The present work is completely different from the conceptualizations of content validity by Lawshe (1975) and Jones and Szatrowski(1983), since the study has developed and examined quantitative indices of the content validity of educational tests rather than employment tests.

1.7 Operational Definition of Terms

1.7.1 Content Validity

Content validity is the degree to which members of the Panel of Expert Judges agree in defining the test items as representative of a relatively homogeneous collection of instances for the examinee to exhibit behaviors measured by the tested domain. The Panel of Expert Judges is composed of certified teachers of a particular subject who judge whether or not knowledge of a given test item is relevant to the tested domain. The consensus of the panel is quantified to produce a content validity index for each item and an overall content validity index for the test.

1.7.2 Test Reliability

Test reliability is the degree to which a test is internally consistent in its measurements. Operationally defined, reliability of the measure of content validity is the degree to which the Panel of Expert Judges consistently agree in defining the content validity of items over the domains.

1.8 Overview of Dissertation

As stated earlier, the purpose of this dissertation is to investigate statistical techniques for estimating the content validity of criterion-referenced standardized tests.

Of the few quantitative content validity techniques available in the literature, none has been applied to a realistic collection of educational achievement tests. Morever, many of the conceptualizations of available quantitative content validity indices are of questionable value in assessing the content validity of educational tests. Thus we have chosen to develop statistical guidelines and apply them to a collection of educational criterion-referenced tests.

In Chapter 2, we briefly describe two alternative methods for eliciting judgments about the content validity of test items. In this same chapter, we present the various research instruments and describe the approach to the design of the research experiment. It is then shown how indices of the content validity of tests can be estimated. Measures of the statistical significance of the values derived from the content validity indices are presented. We then detail the approaches to data analysis. The research hypotheses are then defined.

In Chapter 3, we report and interpret the results of our investigations. Finally, in Chapter 4, we consider the implications of the research results and evaluate the usefulness of the content validity indices that we have developed.

CHAPTER 2

METHODOLOGY

An important problem in statistical estimation of content validity relates to obtaining information about the degree to which individual test items fit within the domain specification. Once data are gathered about the extent to which individual test items fit within the domain specification, quantitative indices and measures are needed for content validity estimation and measurement. In this chapter, we present two alternative methods that were used to elicit judgments about the content validity of test items. The sample of judges with which the methods were used, is described. The various indices and measures of content validity are derived. The approaches to data analysis are discussed.

2.1 Project Design and Activities

In these sections, we present two alternative methods for eliciting judgments about the content validity of test items. We also describe a sample of elementary school mathematics teachers in Grades 3-6 with which these methods were used. For about three hours in a controlled setting, this sample of teachers was asked to judge the content validity of criterion-referenced elementary mathematics test items, some of which had been specially constructed for the research.

When constructing criterion-referenced tests, a method for specifying operationally defined domains is needed. These sections outline a newly-developed method for establishing test content domains. In these same sections, the data collection instruments, the pilot study for the research, and the approach to the design of our experiment are presented.

2.1.1 Population and Sample of Judges

The population of judges for the study consisted of teachers who teach mathematics in Grades 3, 4, 5 and 6 of the Greensboro, North Carolina Public Schools. This population of teachers was not sampled because of the small number of teachers in the four grade levels (68 teachers in Grade 3, 66 teachers in Grade 4, 69 teachers in Grade 5, and 52 teachers in Grade 6). Moreover, since those who teach elementary mathematics are certified to teach in Grades 3-6, all Greensboro teachers in Grades 3-6 define the target population for this research.

For the entire population of teachers, Table 1 contains the frequency of the number of years of teaching experience by grade level. In the population, the years of teaching experience ranges from a minimum of 1 to a maximum of 40. The average number of years of teaching experience is 16.2, the median is 16 and the mode is 17.

Number of Years of Teaching Experience	Grade Level				Row Total
	3	4	5	6	
1 - 5	10	10	6	4	30
6 - 10	12	8	11	9	40
11 - 15	15	13	12	12	52
16 - 20	15	14	17	12	58
21 - 25	6	8	7	10	31
26 - 30	6	9	11	4	30
31 - 35	4	3	4	1	12
36 - 40	0	1	1	0	2
Column Total	68	66	69	52	255

Table 1 Frequency of Number of Years of Teaching Experience by Grade Level, for the Entire Population of Teachers

For the population of teachers, Table 2 shows the frequency of highest degrees earned, by grade level. Just over sixty-seven percent of the population have earned a bachelor's degree "as the highest degree" because actually 100 percent have earned a bachelor' degree, 32.5 percent have earned a master's degree and only one (0.4%) teacher has earned a doctorate.

For the sample of teachers who voluntarily participated in the research activities, the frequencies of the number of years of teaching experience are tabulated by grade level in Table 3. In the sample of teachers, the average number of years of teaching experience is 16.9, the median is 16.5 and the mode is 17.

Highest Degree Earned	Grade Level				Row Total
	3	4	5	6	
Bachelor	48	45	41	37	171
Master	20	20	28	15	83
Doctorate	0	1	0	0	1
Column Total	68	66	69	52	255

Table 2 Frequency of Highest Degree Earned by Grade Level, for the Entire Population of Teachers

The breakdown of the highest degrees earned by the sampled teachers is contained in Table 4. Sixty point seven percent of the sampled teachers have earned a bachelor's degree as the highest degree and 39.3% have a master's degree.

Number of Years of Teaching Experience	Grade Level				Row Total
	3	4	5	6	
1 - 5	3	2	1	0	6
6 - 10	3	2	2	2	9
11 - 15	4	3	2	2	11
16 - 20	3	1	5	0	9
21 - 25	3	4	1	1	9
26 - 30	2	2	2	0	6
31 - 35	2	2	0	1	5
36 - 40	0	0	1	0	1
Column Total	20	16	14	6	56

Table 3 Frequency of Number of Years of Teaching Experience by Grade Level, for Sampled Teachers

Highest Degree Earned	Grade Level				Row Total
	3	4	5	6	
Bachelor	12	9	9	4	34
Master	8	7	5	2	22
Column Total	20	16	14	6	56

Table 4 Frequency of Highest Degree Earned by Grade Level, for Sampled Teachers

2.1.2 Eliciting Judgments

Two techniques were used to elicit judgments about test items. First, a rating scale was developed to tap the perceptions of each content specialist (Grade 3-6 teacher). Each specialist was asked to judge each test item corresponding to each of various test objectives along five dimensions: (1) "format," whether the way the facts were arranged for an item was or was not appropriate for measuring the domain, (2) "wording," whether or not the words used to state the problem for an item were simple enough and within the recognition vocabulary of Grade 4 students, (3) "numbers," whether or not the numbers in an item agreed with the range of the numbers required for the domain, (4) "behavior," whether or not an item elicited the behavior or knowledge to be measured by the domain, and (5) "overall," whether an item was or was not a measure of the domain for which it had been written. Each dimension of judgment was rated on a dichotomous scale with values "Yes" and "No". A rating of "Yes" indicated a judgment that an item was a measure of the objective with respect to the dimension under consideration; a "No" signified a judgment that an item was not a measure of the objective on the rating dimension being considered.

The second procedure that was used to obtain the judgments of content specialists involved the use of a matching task. Content specialists were presented with two lists, one with test items and the other with domain specifications. Each content specialist was then asked to indicate which domain he or she thought each item measured (if any).

It is desirable to have test items that are representative of the domain of items specified in a domain specification. It is only in highly special cases (such as mathematics, vocabulary and spelling) that it is possible to specify completely criteria for a pool of valid test items. In this project, relevant content domains were described clearly to allow content specialists to make a judgment about the representativeness of items included in a test. Judgments were made by content specialists about the representativenesss of the selected test items in the same way that judgments were made about item-objective congruence, above. Specifically, each content specialist was asked to make a judgment concerning the representativeness of the collection of items that were rated or matched as measuring a given domain specification.

2.1.3 Content Domain and Test Specifications

The features of three techniques for specifying content domains were used. Instructional objectives of elementary school mathematics can be written in "behavioral" terms and representative samples of test items measuring each objective can be included in an item pool. Examinees' performances on these item samples provide unbiased estimates of their performances in the full domain of behaviors measured by each objective. However, in our research work, when defining a domain of items for a Grade 4 test, behavioral objectives of Grade 4 elementary school mathematics (see Appendix A.1) that are defined in the Greensboro Public Schools (1982) were supplemented with guidelines regarding testing situations, response alternatives, and criteria of correctness (also called "amplified objectives").

When specifying the domains for the Grade 4 test, rules for generating test items were established and their most noticeable characteristics were described (also called "item forms"). This latter feature made it unnecessary to store individual items, because it allowed items to be generated when needed, by substituting a set of written rules. Also, item forms enabled the relationships among items to be traced by giving clear specifications of relevant item characteristics. In the process of specifying content domains, while retaining the descriptive rigor of item forms, we limited our measurement focus to a small number of assessed elementary school mathematics behaviors. We conceptualized these behaviors so that domain specifications were of larger scale than is typical for behavioral objectives; each domain specification was developed as a large behavior consisting of smaller behaviors (also called "limited-focus").

In summary, we used a newly-developed scheme called "Test Construction Rules" to specify content domains for the Grade 4 test (see Appendix A.2). A Test Construction Rule is one which defines a content domain from an aggregate of a small number of assessed behaviors and specifies testing conditions, rules for generating test items, number of items to be sampled for assessing each behavior, item format, and boundaries for item construction. Note that a test construction rule is an embodiment of "amplified objectives," "item forms," and a "limited-focus strategy."

2.1.4 Data Instruments

Two types of instruments were developed for data collection. The first type contained test items for Grade 4 elementary school mathematics. The Grade 4 test instrument included all of the Mathematics Promotion Standard test items that are defined for Grade 4 in the Greensboro Public Schools (1982), in addition to some specially constructed items (see Appendix A.3). The test instrument covered twelve domain specifications (see Appendix A.2) and included both "good" and specially constructed "bad" test items. A "good" item is one that actually measures some aspect of the content included in a domain specification; a "bad" item fails to measure some aspect of the content included in any domain specification. The test instrument was designed to contain four mixtures of "good" and "bad" items: 91% good-9% bad, 77% good-23% bad, 64% good-36% bad, and 55% good-45% bad.

The test instrument used for domain-item matching and for rating the content of test items along a number of dimensions contained randomly organized "good" and "bad" items within each domain. The test instrument contained a total of two hundred twenty-four items, sixty-eight or approximately 30% of which were "bad" items. The "bad" items were purposefully designed and distributed over twelve Grade 4 tests, in a way that allowed the tests to be split into two sub-tests of equal length and proportion of "good" and "bad" items. Since we investigated two methods for eliciting judgments of test items, this constraint eliminated test treatment effects that might have confounded effects due to method treatment. Within each test, the "bad" items were designed to sample all dimensions of judgment in equal proportions. This latter constraint enabled us to identify and discuss problems associated with content validation of test items using well-defined dimensions of judgment.

The second type of instrument consisted of the data collection instruments used in assessing the content validity of the Grade 4 test that had been constructed for the research. An "Item-Domain Rating Instrument" was designed to allow each content specialist to indicate whether or not each item satisfied requirements of each dimension of judgment within the domain specification (see Appendix B.1). An "Item/Domain Matching Instrument" was designed to allow each content specialist to specify the domain (if any) that each item had been written to measure (see Appendix B.2). A "Domain Representativeness Instrument" was designed to allow each content specialist to indicate whether or not an entire collection of items measuring a domain was representative of the domain (see Appendix B.3).

2.1.5 Pilot Study and Project Activities

Inservice training on techniques for specifying content domains and methods for eliciting judgments concerning content validity was provided for the Grade 3-6 teachers in the study, the teachers' mathematics coordinators of the Greensboro Public Schools and the Grades 3-6 teachers who participated in the pilot study. The Grade 3 inservice training materials are presented in Appendix B.4. The instruments consist of (1) a brief introduction to the concepts and terms of achievement testing, (2) two objectives of Grade 3 tests, (3) domain specifications for the two objectives of Grade 3 tests, (4) a checklist for validating individual test items, (5) sampled Grade 3 test items, and (6) Item-Domain Rating Instrument, Item/Domain Matching Instrument, and Domain Representativeness Instrument for the Grade 3 inservice training tests.

In the initial training, teachers were introduced to the concepts and terms of achievement testing, to apprise them of the reliability and validity issues surrounding achievement test development and test score interpretation. The teachers were then asked to evaluate the two objectives of Grade 3 tests, both in terms of scope and essentiality. Next, the teachers were presented with the three domain specifications corresponding to the two Grade 3 test objectives and were asked to (1) match the domains to their corresponding objectives and (2) comment on the adequacy of the domain specifications. Given that the teachers may find it cumbersome to keep track of all information buried into the domain specifications as they proceed through the evaluations of the worth of test items as measures of domains, teachers were introduced to a technique for developing a checklist for each domain. By making use of one Grade 3 training test, teachers were taught how to assess the content validity of individual test items using a well-defined set of rating dimensions. Also, by making use of another Grade 3 training test, teachers were exposed to the technique of matching items to their corresponding domains. Finally, a collection of self-evaluated content-valid test items was used to demonstrate the process of assessing domain representativeness. Throughout the training we engaged the teachers in discussions of various validity issues and solicited their reasons on (1) why individual test items were evaluated as "valid" or "invalid" and (2) why a collection of self-evaluated content-valid test items was assessed as representative or not representative of the domain.

The Grade 4 test domain specifications were developed after carefully reviewing (1) the mathematics curriculum and objectives that are defined for Grade 4 in the Greensboro Public Schools (1982), and (2) the Teacher's Edition Holt Mathematics (1981) and the Student's Edition Holt Mathematics (1981) that are used respectively by the teachers and the students at the Greensboro Public Schools.

The question naturally arises as to the adequacy of our Grade 4 test domain specifications. Three content specialists in the area of elementary school mathematics validated our Grade 4 test domain specifications. A professor of educational research and evaluation with previous years of teaching experience in elementary mathematics and a professor of mathematics education at the

University of North Carolina at Greensboro, validated the initial version of our domain specifications and suggested changes as appropriate. The initial domain specifications were revised and the teachers' mathematics coordinator was asked to match the content domains that had been specified for this project with the Greensboro Public Schools' Mathematics Promotion Standards (skills). After reviewing a summary of the coordinator's judgments of the adequacy of our domain specifications, the domain specifications that were found to be inadequate were revised. The Mathematics Supervisor of the Greensboro Public Schools then validated our final domain specifications. These activities enabled us to confirm the adequacy of our domain specifications both in terms of scope and essentiality.

A pilot study was conducted to provide an estimate of the amount of time that teachers in Grades 3-6 needed to complete the project survey instruments. Also, the pilot test illuminated weaknesses in our project instruments. In the pilot study, the teachers' mathematics coordinators made judgments concerning the content validity of six tests that covered six mathematics skills for Grades 4 and 5 (three skills were selected from each grade).

In predefined order, random samples of twenty-eight teachers from Grades 3-6 made judgments concerning the content validity of test items currently part of the Greensboro Mathematics Tests and those that had been specially constructed for the project.

2.1.6 Experimental Design

Fifty-six Grade 3-6 teachers (20 Grade 3 teachers, 16 Grade 4 teachers, 14 Grade 5 teachers and six Grade 6 teachers) made judgments concerning the content validity of twelve Grade 4 tests and other test items that had been specially constructed for this research. The teachers were divided into two groups. In order to equate the the two groups of teachers, teachers were paired randomly within each grade level, and one teacher was randomly assigned to each group. The two groups of teachers received the two method treatments in a different order. The same groups of teachers were exposed to both methods of eliciting judgments, thereby controlling for treatment order.

The order in which the groups of teachers were asked to apply the methods for eliciting judgments was determined randomly. One group was asked to judge the first half of this project's test items (Test 1-6 items) using the "item-domain matching" method of eliciting judgments, and then judged the remaining test items (Test 7-12 items) along a number of dimensions. The second group was asked to judge Test 1-6 items along a number of dimensions, and was then asked to indicate the domain (if any) which each of Test 7-12 items measured. There is potential multiple-treatment interference when each group of teachers receives two treatments. However, since for a given group of judges, different tests were used with the two treatments and treatments were balanced across orders of application, exposure to one method should not have affected evaluation of the effectiveness of the other.

Figure 1 shows the counterbalanced, factorial experiment with one within-subjects factor. The tests are nested within method-order. Alternate forms of Grade 4 tests were used with the two methods of eliciting judgments. When dividing the Grade 4 tests, deliberate attempts were made to ensure that the two sub-tests were of equal difficulty. This helped control for any instrumentation effects that might have occurred.

		Methods	
		Rating	Matching
Groups	G1	Tests1-6	Tests7-12
	G2	Tests7-12	Tests1-6

Figure 1 Factorial Experiment with Method-Order Nested within Groups.

Since this study was conducted in a controlled setting with only Grade 3-6 teachers at the Greensboro Public Schools, we recognize a threat to the generalizability of our results. However, our content validity indices and other variables are operationally defined in ways that have meaning outside the Greensboro Public Schools. In the following sections we consider the content validity indices and measures that were investigated in this research.

2.2 Content Validity Estimation and Measurement

In the content validation process, specialists were asked to indicate the domain (if any) that each Grade 4 mathematics test item measured. Also, the specialists were asked to rate each test item along a number of dimensions. The question naturally arises as to the validity of the judgments of the specialists. Whenever specialists fail to agree that a test item measures a domain, serious questions can be raised. On the other hand, if all specialists agree, it can be concluded that they are either "all correct" or "all wrong." Since all specialists were certified, and were engaged in teaching the tested knowledge or skills, it is difficult to reject a strong consensus. When all specialists say that an item fits the domain specification for which it has been written, or when none say that it measures the domain, we have reason to believe that the item is or is not a measure of the domain. However, problems arise when the strength of the consensus diminishes, and judgments approach a fifty-fifty split. In the following sections, we present various indices and measures of content validity.

The proportion of content specialists agreeing that an item is a measure of a domain was examined as an index of item-domain congruence. The proportion of content specialists indicating that each item fit the requirements of each dimension of the domain it had been written to measure was considered as an index of item-domain-dimension congruence. The proportion of content specialists agreeing that test items adequately represented each domain was treated as an index of domain representativeness. Using these proportions, the Chi-square test of proportions and the normal approximation to the binomial distribution were used as bases for quantitative indices of content validity.

Purposefully, "bad" items were introduced into the research test instrument to allow an assessment of the degree of "accuracy" of the content specialists' judgments. Rates of false-positive and false-negative errors were used as indices of accuracy for each test. A measure of judgment reliability is presented.

2.2.1 Index and Significance Test for Individual Items

Suppose N_j independent specialists judge and classify the jth test item into one of two categories: whether the item is or is not a measure of a domain. Define a "success" as a judge expressing correct judgment that an item is or is not a measure of a domain, and a "failure" as his or her incorrect judgment that the item is or is not a measure of the domain. If indeed no conclusion can be reached concerning the worth of an item as a measure of a domain, it is reasonable to expect half of the N_j specialists to judge the jth item as a measure of the domain, while the other half will disagree. Thus, in this circumstance, the expected number of judges in each category for the jth item would be $0.5N_j$.

Let Θ represent the probability of success in the population of judges. For the jth item, the null hypothesis that a success and a failure are equally likely in the population, that is Ho: $\Theta = 0.5$, and the alternative hypothesis that the probability of success is different from 0.5, that is, Ha: $\Theta \neq 0.5$, were formulated. Note that, under Ho, the number of successes for N_j judges follows the binomial distribution with $\Theta = 0.5$, and the expected number of successes is then $0.5N_j$.

Let P_j be the proportion of judges expressing successful judgments that the jth item is (or is not) a measure of a domain. One index of the content validity of the jth item is defined as the proportion P_j. Clearly, to the extent that $(P_j \times N_j)$ is larger than $(1-P_j)N_j$ or P_j is greater than 0.5 or $(2 \times P_j - 1)N_j$ is greater than zero, the jth item can be said to be a measure of a domain. Conversely, the extent to which $(1-P_j)N_j$ is larger than $(P_j \times N_j)$ or P_j is less than 0.5 determines the degree to which the jth item is not a measure of a domain. With $(P_j \times N_j)$ successes, $(1-P_j)N_j$ failures and $\Theta = 0.5$, the test statistic for the jth item is defined as the following standardized random variable.

$$z = \frac{P_j \times N_j - 0.5 - 0.5N_j}{0.5\sqrt{N_j}}$$

The normal approximation to the binomial distribution is used to test the statistical significance of the P_j for the jth test item. Given the small sample size available for this kind of study, the Type I error level α was set to 0.05 to ensure reasonable power of the test of Ho.

2.2.2 Index and Measure of Accuracy of Method

Of particular interest to the present study is the accuracy of each method used to elicit judgments of the content validity of test items. Consider the 2 x 2 contingency table in Figure 2 constructed for a method of eliciting judgments. The state of "Reality" in Figure 2 refers to the actual status of the items in a test or collection of tests --the status of each item is either "valid" or "invalid" depending on whether or not the item has been constructed to measure the skill tested. In Figure 2, N1 is the number of valid test items in "Reality" for which a particular method of eliciting judgments of content validity resulted in a judgment of "valid," N2 is the number of invalid items for which the method resulted in a judgment of "valid," N3 is the number of valid items for which the method resulted in a judgment of "invalid," and N4 is the number of invalid items for which the method resulted in a judgment of "invalid."

		Reality	
		Valid	Invalid
Method of Eliciting Judgments	Valid	N1	N2
	Invalid	N3	N4

Figure 2 2 x 2 Contingency Table for the Accuracy of a Method of Eliciting Judgments

For each test and for the entire collection of tests, the numbers N1, N2, N3 and N4 are computed by making use of the results of the normal approximation to the binomial distribution tests for individual test items. The p-values of the P_j's (proportions of judges expressing successful judgments on test items) derived in Section 2.2.1 are used to determine the numbers N1, N2, N3 and N4.

We define a false-negative error as a valid test item for which a method of eliciting judgments resulted in an incorrect judgment of "invalid," and define a false-positive error as an invalid test item for which a method of eliciting judgments resulted in an incorrect judgment of "valid." For each test and for the entire collection of tests, the following hypotheses are relevant to the determination of the accuracy of each method of eliciting judgments:

(i) The null hypothesis that the population proportion of false-positive errors equals zero when a particular method is used for eliciting judgments of content validity is tested against the alternative hypothesis that the population proportion of false-positive errors is greater than zero when the method is used to elicit judgments of content validity.

(ii) The null hypothesis that the population proportion of false-negative errors equals zero when a particular method is used for eliciting judgments of content validity is tested against the alternative hypothesis that the population proportion of false-negative errors is greater than zero when the method is used to elicit judgments of content validity.

(iii) The null hypothesis that judgments are uncorrelated with state of "Reality" using a particular method, is tested against the alternative hypothesis that there is a positive correlation between judgments which resulted from using the method and the state of "Reality."

Formally, let Pfp denote the proportion of false-positive errors and let Pfn represent the proportion of false-negative errors which resulted from using a particular method to elicit judgments of the content validity of test items. With V1 valid items and V2 invalid items in a test or a collection of tests, Pfp is given by Pfp = N2/V2, and Pfn =N3/V1. Let N=V1+V2, T1.= N1+N2, T2.=N3+N4, T.1=N1+N3, and T.2 = N2+N4. Also, let C=N*(N1*N4-N2*N3)^2/(T1.*T2.*T.1*T.2). The correlation between judgments which resulted from using a method and "Reality," Phi, is given by Phi =$\sqrt{(C/N)}$ (see Glass and Hopkins, 1984). For each test and for the entire collection of tests, the formal hypotheses to be tested are

(i) Ho: Pfp = 0
Ha: Pfp > 0

(ii) Ho: Pfp = 0
Ha: Pfn > 0

(iii) Ho: $\rho_o = 0$
Ha: $\rho_o > 0$

The Sign Test procedure was used to test the formal hypotheses (i) and (ii) above if the number of test items in a domain was not more than 20, otherwise the normal approximation the binomial distribution procedure was used. A small α-level was used (α=0.01) to guard against a large experimentwise Type I error rate. The statistical significance of Phi was tested using the Chi-Square test.

The 2 x 2 contingency table in Figure 2 illustrates the data for determining the accuracy of a method of eliciting judgments. This table represents the ideal situation where all judgments are conclusive on the worth of all items as measures or non-measures of a domain. More realistically however, it is reasonable to expect that there are test items for which there is no consensus on the worth of the items. In that case, we will have the contingency table in Figure 3 since there is no "inconclusive" category in "Reality."

One way to treat the inconclusive data would be to eliminate them from subsequent hypothesis testing. It seemed more logical, however, to use the inconclusive data to examine the worst case and the best case when (1) deriving the correlation between judges' judgments and the state of reality, and (2) computing the proportions of the false-positive and false-negative errors.

In order to compute the best-case proportions of false-positive and false-negative errors, the test items that were valid in "Reality" but for which a method resulted in a judgment of "inconclusive" (N5), were added to the valid items in "Reality" for which the method also resulted in a judgment of "valid" (N1 = N1 + N5); the test items pronounced "inconclusive" but were invalid in "Reality" (N6), were added to the invalid items in "Reality" for which the method also resulted in a judgment of "invalid" (N4 =N4 +N6). To estimate the worst-case proportions of false-negative and false-positive errors, the test items pronounced "inconclusive" but are valid in "Reality," were added to the valid test items in "Reality" but for which a method resulted in a judgment of "invalid" (N3 = N3 + N5); the test items for which the method resulted in a judgment of "inconclusive" but were invalid in "Reality," will be added to the test items for which the method resulted in a judgment of "valid" but were invalid in "Reality" (N2 = N2 + N6).

		Reality	
		Valid	Invalid
Method of Eliciting Judgments	Valid	N1	N2
	Invalid	N3	N4
	Inconclusive	N5	N6

Figure 3 3 x 2 Contingency Table for the Worst- and Best-Case Accuracy of a Method of Eliciting Judgments

2.2.3 Measure of Similarity of Rating and Matching Methods

In this research, we have developed two alternative methods of eliciting judgments: Rating and Matching methods. The question before us is: are the two methods of eliciting judgments of content validity alike? That is, with regard to eliciting judgments of the content validity of tests, are the rating method and the matching method equally accurate?

One way to compare the Rating and Matching methods is to examine the proportion of false-positive errors and the proportion of false-negative errors. For each test let Pfpr and Pfpm denote the respective proportions of false-positive errors derived from judges' rating and matching judgments, respectively. Also, for each test, let Pfnr and Pfnm denote the proportions of false-negative errors derived from judges' rating and matching judgments, respectively. The bivariate plots Pfpr vs Pfpm and Pfnr vs Pfnm across tests were examined to provide some conclusions on the respective distributions. Overall Sign Tests were performed on the pairs across tests.

Another way to examine the similarity between the Rating and the Matching methods is to examine the proportions of false-positive and false-negative errors for all items in the entire collection of tests. For all items in all tests, let Ofpr and Ofpm denote the respective proportions of false-positive errors derived from judges' rating and matching judgments. Also, for all items in all tests, let Ofnr and Ofnm denote the proportions of false-negative errors derived from judges' rating and matching judgments, respectively. For each pair of proportions of false-positive and false-negative errors, the normal approximation to the binomial distribution is appropriate for performing a test of equality of proportions. The Type I error level α was set to 0.01 to guard against a large experimentwise Type I error rate.

Finally, we used a measure of similarity between the Rating and Matching methods of eliciting judgments. For each Grade 4 test and for the entire collection of Grade 4 tests, consider the 3 x 3 contingency table in Figure 4. In Figure 4, Al is the number of items rated valid and also matched as valid; A2 is the number of items rated valid but matched as invalid; A3 is the number of items for which no consensus agreement could be reached on the match between the items and the domains but for which the ratings resulted in judgments of valid; A4 is the number of items rated invalid but matched as valid; A5 is the number of items rated invalid and also matched as invalid; A6 is the number of items for which no consensus agreement could be reached on the match between the items and the domains but for which ratings resulted in judgments of invalid; A7 is the number of items for which the ratings resulted in no consensus agreement on the content validity of the items but for which the Matching method resulted in judgments of valid; A8 is the number of items for which the ratings resulted in no consensus agreement on the content validity of the items but for which the Matching method resulted in judgments of invalid; and A9 is the number of items for which the ratings resulted in no consensus agreement on the content validity of the items and for which the matching judgments were inconclusive. For each Grade 4 test and for the entire collection of Grade 4 tests, the numbers Al-A9 were calculated from the proportions Pr's (proportions of judges rating items) and the proportions Pm's (proportions matching items to domains).

		MATCHING METHOD		
		Valid	Invalid	Inconclusive
RATING METHOD	Valid	A1	A2	A3
	Invalid	A4	A5	A6
	Inconclusive	A7	A8	A9

Figure 4 3 x 3 Contingency Table for Comparing the Similarity of the Rating and Matching Methods

The null hypothesis that there is no correlation between Rating judgments and Matching judgments was then tested against the alternative hypothesis that there is a positive correlation between Rating judgments and Matching judments. Formally, With N items in a test or a collection of tests, Let $T1.=A1+A2+A3$, $T2.=A4+A5+A6$, $T3.=A7+A8+A9$, $T.1=A1+A4+A7$, $T.2=A2+A5+A8$ and $T.3=A3+A6+A9$. Also, let $Q=N*[(A1)^2/(T1.*T.1) + (A2)^2/(T1.*T.2) + (A3)^2/(T1.*T.3) + (A4)^2/(T2.*T.1) + (A5)^2/(T2.*T.2) + (A6)^2/(T2.*T.3) + (A7)^2/(T3.*T.1) + (A8)^2/(T3.*T.2) + (A9)^2/(T3.*T.3) -1]$. One measure of association between rating and matching judgments is the Cramér statistic (see Cramér, 1946; Gibbons, 1976; and Conover, 1980). Consistent with the Phi statistic, the Cramér statistic is defined as $C = \sqrt{(Q/(2*N))}$. For each test and for the entire collection of tests, the formal hypothesis to be tested is: Ho: $\rho_o = 0$ against Ha: $\rho_o > 0$. The statistical significance of each value of the Cramer statistic was tested using the Chi-Square test.

2.2.4 Measure of Effects of Proportions of Bad Items on Accuracy of Judgments and Content Validity Indices

In the specially constructed Grade 4 test instrument, there were four samples of "bad" and "good" items (9% bad-91% good, 23% bad-77% good, 36% bad-64% good and 45% bad-55% good). Each item in each sample was classified into one of two fixed categories --"bad" or "good." In other words, each sample was drawn from a dichotomous population of items. Two questions of particular interest to the content validity of mathematics achievement test items are: (1) Is there any effect of the proportions of "bad" items on the accuracy of judgments? and (2) Is there any effect of the proportions of "bad" items on the accuracy of each content validity index?

Let the parameters PB1, PB2, PB3 and PB4 denote the respective probabilities of an item being classified as "bad" in the four populations of items. In order to investigate the effects of the proportions of "bad" items on the accuracy of judgments, the null hypothesis that the four population proportions of false-positive or false-negative errors are equal was tested against the alternative hypothesis that at least two of the population proportions of false-positive or false-negative errors differed from each other.

The data to be analyzed were enumerative, representing the numbers of false-positive or false-negative errors in each of the four samples of "bad" and "good" test items. We denote these observed frequencies by F1, F2, F3 and F4. If the probability of false-positive or false-negative errors occuring is the same in the four populations of mixtures of "bad" and "good" items, the logical sample estimate of this common probability, denoted by P, is the total number of false-positive or false-negative errors observed, divided by the total number of test items, or P =(F1+F2 +F3+F4)/(N1+N2+N3+N4), where N_j is the number of items in the jth sample of "bad" and "good" items. The corresponding estimate of the number of false-positive or false-negative errors in sample number j then is N_j *P.

Formally, the hypothesis to be tested is

Ho: PB1 = PB2 = PB3 = PB4
Ha: At least two of the PB_j's differ from each other.

If the null hypothesis is true, there should be close agreement between the observed frequency of false-positive or false-negative errors F_j , and the expected frequency of errors N_j*P. The Chi-square test statistic for determining the equality of the four population proportions of false-positive or false-negative errors is

$$Q = \sum_{j=1}^{4} \frac{(F - N *P)^2}{N_j *P} ,$$ with 3 degrees of freedom.

A large value of this statistic reflects heterogeneity among the four population proportions of false-positive or false-negative errors, and hence among the effects of the proportions of "bad" items on the accuracy of judgments. Similarly, since the four samples of "bad" and "good" items were independent, the procedures outlined above were used to test equality of proportions for each pair of false-positive or false-negative errors.

We define a "correct" decision as a valid test item for which the value of an index of content validity resulted in a correct judgment of "valid" or an invalid test item for which the index value resulted in a correct judgment of "invalid;" define an "incorrect" decision as a valid test item for which the index value resulted in an incorrect judgment of "invalid" or an invalid test item for which the index value resulted in an incorrect judgment of "valid;" and define an "inconclusive" decision as a valid or invalid item for which the index value resulted in a judgment of "inconclusive."

One way to examine the effects of the proportions of "bad" items on the accuracy of an index of content validity is to calculate and plot the proportions of "correct" decisions versus the proportions of "bad" items. Ideally, such a graph should exhibit a monotonic decreasing sequence since the accuracy of a "good" index should decrease (the proportion of "correct" decisions decreases) as the proportion of "bad" items increases. Alternatively, the proportions of "incorrect" or "inconclusive" decisions can be computed and plotted against the proportions of "bad" items. In that case, the graph should exhibit a monotonic increasing pattern since the accuracy of a "good" index should increase (the proportion of "incorrect" or "inconclusive" decisions decreases) as the proportion of "bad" items increases.

Another way to determine the effects of the proportions of "bad" items on the accuracy of an index of content validity is to use a Chi-square procedure to test equality of proportions for each pair of "correct" or "incorrect" decisions. The data to be analyzed were enumerative, representing numbers of "correct" or "incorrect" or "inconclusive" decisions in each of the four samples of "bad" and "good" test items. Let F1, F2, F3 and F4 denote these observed frequencies. Note that F1, F2, F3 and F4 are computed separately for the "correct," "incorrect" and "inconclusive" decisions.

The null hypothesis that the four population proportions of "correct" or "incorrect" or "inconclusive" decisions are equal was tested against the alternative hypothesis that at least two of the population proportions of "correct" or "incorrect" or "inconclusive" decisions differed from each other. Also, each pair of proportions of "correct" or "incorrect" or "inconclusive" decisions were tested for equality. The Chi-square test statistic and the formal hypothesis to be tested were derived in a manner that is analogous to the derivations given above for determining heterogeneity among the four population proportions of false-positive or false-negative errors.

2.2.5 Index and Measure of Representativeness

One major question in content validation of tests is: to what extent do the test items cover the scope of a domain specification? Let PR_j be the proportion of judges indicating judgments that a collection of items adequately covered the scope of a domain specification and let N_j denote the number of judges who made judgments on whether or not a collection of items adequately covered the scope of the jth domain. Define a "success" as a judge expressing judgment that a collection of items adequately covered the scope of a domain, and a "failure" as his or her judgment that the collection of items did not cover the scope of the domain.

If no conclusion could be reached on whether or not a collection of items adequately covered the scope of a domain, it would be reasonable to expect half of the specialists to judge the collection of items as representative of the domain, while the other half would disagree. Thus, the expected number of judges in each category for the jth domain would be $0.5N_j$. Let Θ represent the probability of success in the population of judges. For each domain, the null hypothesis that a success and a failure are equally likely in the population, that is Ho: $\Theta = 0.5$, and the alternative that a collection of items adequately covers or does not cover the scope of a domain, that is, Ha: $\Theta \neq 0.5$ were formulated. The proportion PR_j, of judges indicating judgments that a collection of items adequately covered the scope of the jth domain was used as the index of domain content representativeness. To the extent that PR_j is greater than 0.5 or $(PR_j \times N_j)$ is larger than $(1-PR_j)N_j$, the scope of the jth domain can be said to be adequately covered by the collection of items. Alternatively, the degree to which PR_j is less than 0.5 or $(1-PR_j)N_j$ is larger than $(PR_j x N_j)$ determined the extent to which a collection of items did not cover the scope of the jth domain. To determine statistically, whether or not the scope of the jth domain was adequately covered by a collection of items, the normal approximation to the binomial distribution was used to test the significance of the PR_j for the jth domain.

2.2.6 Index and Measure of Interjudge Reliabilities

In order to examine the degree of consistency of the judges in their judgments over the domains, define the jth item's score S_j as the number of judges who identified the item correctly as a measure or non-measure of a domain. Let N denote the number of teachers who judged the jth item and let I_i denote the score for the ith judge derived from the data which resulted from using a particular method to elicit judgments. Each of the I_i equals one for an item identified correctly as a measure or non-measure of the dth domain by the ith judge, and zero otherwise. Thus, the score S_j is computed as the number of judges who used a particular method to judge the jth item correctly as a measure or non-measure of a domain. $S_j = \sum_{i=1}^{N} I_i$.

Let N_d be the number of items in the dth domain and let I_{kj} denote the score for the kth judge on the jth item. Each of the I_{kj} equals one for the jth item identified correctly as a measure or non-measure of the dth domain by the kth judge, and zero otherwise. The proportion PJ_k of items identified correctly by the kth judge as measures or non-measures of a domain using a particular method of eliciting judgments is given as $PJ_k = (\sum_{j=1}^{Nd} I_{kj})/N_d$. The S_j's and PJ_k's were used to determine the internal consistency of the teachers' judgments of items over domains as follows.

Let TR-87 be an index of interjudge reliability. The index TR-87 is defined as

$$\text{TR-87} = \frac{N}{(N - 1)} \left\{ 1 - \frac{\sum_{k=1}^{N} [PJ_k * (1 - PJ_k)]}{S^2} \right\}$$

where

$$S^2 = \frac{N_d \sum_{j=1}^{N_d} S_j^2 - (\sum_{j=1}^{N_d} S_j)^2}{N_d (N_d - 1)} .$$

In subsequent sections, we consider in more detail the use of the content validity indices and measures outlined above. These sections detail the analyses of data for testing the hypotheses introduced earlier.

2.3 DATA ANALYSIS PROCEDURES

The intent of this analysis was to obtain information on the extent to which the results of content validation were dependent on the techniques used for eliciting judgments, to examine whether the degree of accuracy of teachers (in defining the content validity of items) varied with the proportion of "bad/good" items presented to them, and to investigate whether various indices of content validity increased or decreased as the proportion of "bad" items provided to teachers increased.

In order to obtain answers to each of the basic research questions listed in the Research Questions Section (1.4) of this dissertation, the Statistical Package for the Social Sciences was used to obtain crosstabulations of the number and percent of the responses in each of the categories under consideration. In these sections we describe our approaches to data reduction and analyses.

2.3.1 Data Editing, Coding and Reduction

Procedures used for data editing, coding, and reduction included the following:

(1) Each completed instrument was assigned a unique number which was stamped on the first page of the instrument. This number was used for analysis purposes.

(2) All judges' instruments were edited for completeness and appropriateness of responses. In order to facilitate data coding and quantitative analysis, guidelines for error detection and resolution consisted of the following:

(a) Given the small sample size of judges, each instrument was validated for completeness to ensure 100 percent response rate --we ensured that each teacher completed all sections of the instrument prior to its acceptance.

(b) A codebook was developed and pilot tested for mutual exclusivity. All judgment data were coded directly into numerical form as each instrument was being screened.

(c) All invalid or unacceptable judgments were circled for coding into "inappropriate judgments" or "no judgments" categories. For instance, if a judgment was made on the content validity of an item where a dimension of judgment was not applicable, that judgment was coded into an "inappropriate judgments" category.

(3) All coded data from edited instruments were entered into a database. Code checks were completed on all data field values in the database by using a data validation program. The computer validation program

identified missing coded data and also determined if coded judgments were within appropriate ranges.

(4) For each logical record in the database, there was associated, a field to designate the order in which each subject applied the methods for eliciting judgments, a domain number field, and an item number indicator field. The order indicator field enabled comparison of the two methods of eliciting judgments. The domain and test item indicator fields made it possible to analyze the accuracy of teachers' judgments of item-domain congruence and the consistency of their judgments of items over domains.

(5) For each test item, a correct judgment of an item as a measure or non-measure of a domain was scored as "1," an incorrect judgment was scored as "0."

A preliminary analysis was performed to summarize data for subsequent data analysis. Within each group of teachers and for the "rating method" treatment, data consistent with instructions were analyzed to estimate the proportion of teachers who (1) judged the format of an item as appropriate for measuring a domain, (2) judged the words used to state the problem (where applicable) for an item as simple enough and appropriate for students in Grade 4, (3) judged the magnitudes of the numbers in an item as consistent with a domain specification, (4) judged an item as eliciting the behavior or knowledge to be measured by a domain, and (5) made summative judgments that, overall, an item was a measure of a domain. Within each group of teachers and for the "matching method" treatment, the proportion of teachers who identified an item as a measure of a domain (i.e., matched the item and the domain) was computed. Using the combined data for the two groups of teachers, the proportion of teachers who made judgments that the scope of a domain was adequately covered, was calculated. Also, the proportion of teachers who made judgments that the scope of a domain was adequately covered, was computed separately using the judgment data for each method of eliciting judgments.

2.3.2 Analysis of Accuracy of Methods and Computation of Indices

The first set of analyses was designed to provide answers to two major questions: (1) To what extent are the results of content validation dependent on methods for eliciting judgments? and (2) Do the various indices and measures of content validity increase or decrease as the proportion of "bad" items provided to judges increases?

Three techniques were used to derive estimates of the content validity of each item. First, from the item domain rating data, the proportion of teachers P1, who expressed judgments that an item satisfied all of the requirements of a domain (with regard to "format," "wording" --where appropriate, --"sizes of numbers," and "behavior") was derived. Second, the proportion of teachers P2, who made summative judgments that, overall, an item was a measure of a domain was calculated. Third, from the item/domain matching data, the proportion of teachers P3, who identified each item as a measure of a domain was computed. The proportions Pl, P2, and P3 were used separately as content validity indices of individual test items. The validity of the content of individual test items was determined using the indices Pl, P2, and P3 separately with the test procedure outlined in Section 2.2.1.

In order to determine whether or not Rating and Matching methods were accurate for eliciting judgments of the content validity of tests, the accuracy of each method of eliciting judgments was investigated separately. For the Rating method, Nl was computed as the number of valid test items in "Reality" for which the Rating method of content validation resulted in a judgment of "valid," N2 was calculated as the number of invalid items for which the Rating method resulted in a judgment of "valid," N3 was computed as the number of valid items for which the Rating method resulted in a judgment of "invalid," N4 was calculated as the number of invalid items for which the Rating method resulted in a judgment of "invalid," N5 was calculated as the number of valid items for which the Rating method resulted in a judgment of "inconclusive," and N6 was computed as the number of invalid items for which the Rating method resulted in a judgment of "inconclusive." For each test and for the entire collection of tests, the numbers N1, N2, N3, N4, N5 and N6 were computed by making use of the results of the normal approximation to the binomial distribution tests of individual test items. The p-values of the P1's (proportions of judges rating items on all dimensions) and the p-values of the P2's (proportions of judges rating items on the overall dimension) were used separately to determine the numbers N1, N2, N3, N4, N5 and N6.

For each test and for the entire collection of tests, the proportions of false-positive and false-negative errors were calculated for the Rating method using the computational procedures outlined in Section 2.2.2. By making use of the statistical test procedures outlined in Section 2.2.2, the following hypotheses were tested.

(i) The null hypothesis that the population proportion of false-positive errors equalled zero when the Rating method was used for eliciting judgments of content validity was tested against the alternative hypothesis that the population proportion of false-positive errors was greater than zero when the Rating method was used to elicit judgments of content validity.

(ii) The null hypothesis that the population proportion of false-negative errors equalled zero when the Rating method was used for eliciting judgments of content validity was tested against the alternative hypothesis that the population proportion of false-negative errors was greater than zero when the Rating method was used to elicit judgments of content validity.

(iii) The null hypothesis that teachers' judgments using the Rating method were uncorrelated with the state of "Reality" was tested against the alternative hypothesis that there is a positive correlation between rating judgments and the state of "Reality."

Notice that two sets of statistical tests were performed for the Rating method. The first set of statistical tests was performed using the proportions of false-positive and false-negative errors derived from the p-values of the P1's (proportions of judges rating items on all dimensions). The second set of statistical tests was performed using the proportions of false-positive and false-negative erros derived from the p-values of the P2's (proportions of judges rating items on the "Overall" dimension). Similarly, the accuracy of the Matching method of eliciting judgments of content validity was determined using the procedures described above. However, the p-values of the P3's (proportions of judges matching items to domains) were used to determine the numbers N1, N2, N3, N4, N5 and N6 when calculating the worst-case and the best-case proportions of false-positive and false-negative errors.

2.3.3 Analysis of Similarity of the Rating and Matching Methods

The next analysis was designed to provide information on the extent to which the Rating and the Matching methods were alike with regard to eliciting judgments of the content validity of tests. First, to compare the accuracy of the Rating and Matching methods, the proportions of false-positive errors and the proportions of false-negative errors were examined. For each test, we denoted the respective proportions of false-positive errors derived from judges' ratings on all dimensions, judges' ratings on the "Overall" dimension, and matching judgments by Pfp1, Pfp2, and Pfp3, respectively. Also, for each test, we represented the proportions of false-negative errors derived from judges' ratings on all dimensions, judges' ratings on the "Overall" dimension, and matching judgments, by Pfn1, Pfn2 and Pfn3, respectively. The bivariate plots Pfp1 vs Pfp2, Pfp1 vs Pfp3, Pfp2 vs Pfp3, Pfn1 vs Pfn2, Pfn1 vs Pfn3, and Pfn2 vs Pfn3 across tests were constructed and examined. Overall Sign Tests were performed on the pairs across tests.

Second, for all items in all tests, we denoted the respective proportions of false-positive errors derived from judges' ratings on all dimensions, judges' ratings on the "Overall" dimension, and matching judgments by Ofp1, Ofp2 and Ofp3, respectively. Also, for all items in all tests, we represented the proportions of false-negative errors derived from judges' ratings on all dimensions, judges' ratings on the "Overall" dimension, and matching judgments, by Ofn1, Ofn2 and Ofn3, respectively. For each possible pair of proportions of false-positive and false-negative errors, the normal approximation to the binomial distribution was used in performing a z-test of equality of proportions. The Type I error level α was set to .01 to guard against a large experimentwise Type I error rate.

Third, for each Grade 4 test and for the entire collection of Grade 4 tests, we computed A1 as the number of items rated valid and also matched as valid, A2 as the number of items rated valid but matched as invalid, A3 as the number of items rated valid but resulting in a judgment of "inconclusive" when matched, A4 as the number of items rated invalid but matched as valid, A5 as the number of items rated invalid and also matched as invalid, A6 as the number of items rated invalid but resulting in a judgment of "inconclusive" when matched, A7 as the number of items which resulted in a judgment of "inconclusive" but were matched as valid, A8 as the number of items which resulted in a judgment of "inconclusive" but were matched as invalid, and A9 as the number of items which resulted in a judgment of "inconclusive," both when rated and matched. For each Grade 4 test and for the entire collection of Grade 4 tests, the numbers A1, A2, A3, A4, A5, A6, A7, A8 and A9 were calculated using two different methods. First, the proportions P1's (proportions of judges rating items on all dimensions) were used with the proportions P3's (proportions matching items to domains). Second, the proportions P2's (proportions of judges rating items on the "Overall" dimension) were used with the proportions P3's (proportions matching items to domains). By making use of the computational procedures outlined in Section 2.2.3 with the numbers A1, A2, A3, A4, A5, A6, A7, A8 and A9, the similarity between the Rating and Matching methods was determined.

2.3.4 Analysis of Content Domain Representativeness

In order to obtain information on the extent to which test items covered the scope of a domain, the proportion of judges PR, indicating judgments that a collection of items adequately covered the scope of a domain specification, was computed using two techniques. First, the proportion PR was calculated separately for each method of eliciting judgments. Second, the proportion PR was computed from the responses of all judges. By making use of the proportion PR's with the statistical test procedure outlined in Section 2.2.4 the representativeness of each collection of items measuring a domain was determined.

2.3.5 Analysis of Accuracy of Judgments and Content Validity Indices

In Section 2.1.3 we discussed a set of test instruments which contained four mixtures of "good" and "bad" items (91% good- 9% bad, 77% good-23% bad, 64% good-36% bad, and 55% good-45% bad) that was specially constructed for this research study. Two questions of interest to the present research are: (1) Does the degree of accuracy of judges (in defining the content validity of items) vary with the proportion of "bad/good" items presented to them? and (2) Does the degree of accuracy of each content validity index increase or decrease as the proportion of "bad" items provided to judges increases?

In order to determine whether the proportions of errors and the methods used to elicit judgments were independent, the Chi-square test of association between the proportions of false-negative judgments and the methods of eliciting judgments was performed.

The Chi-square test of equality of proportions was used to determine the effects of the proportions of "bad" items on the degree of accuracy of judges, separately for the Rating and Matching methods. Because the item samples were small, the proportions of false-positive items were determined across all tests from the proportions P1's (calculated from the ratings on all dimensions), P2's (computed from the ratings on the "Overall" dimension) and P3's (calculated from the matching of items).

First, a Chi-square test for equality was performed on the proportions of false-positive items across the proportions of "bad" items. Second, the functional relationship between false-positive judgments and the proportions of "bad" items was examined graphically.

The Chi-square test of equality of proportions was used to determine the effects of the proportions of "bad" items on the degree of accuracy of the content validity indices that were derived separately from the rating and matching judgments. Because the item samples were small, the proportions of "correct," "incorrect" and "inconclusive" decisions were determined across all tests from the index values P1's (calculated from the ratings on all dimensions), P2's (computed from the ratings on the "Overall" dimension) and P3's (calculated from the matching of items). A Chi-square test for equality was performed on the proportions of "correct," "incorrect" and "inconclusive" decisions across the proportions of "bad" items. The functional relationship between each of these categories of decisions and the proportions of "bad" items was determined graphically.

2.3.6 Analysis of Interjudge Consistency

In order to examine the degree of consistency of the judges in their judgments over the domains, the reliability index, TR-87, and statistical test procedures outlined in Section 2.2.5 were used. Each individual item's score was derived as the number of judges who identified the item correctly as a measure or non-measure of a domain. We denoted the number of teachers who judged the jth item by N_j and represented the respective scores for the ith judge derived from "rating," "summative judgment," and "item-domain matching" data by $I1_i$, $I2_i$ and $I3_i$. Each of $I1_i$, $I2_i$, and $I3_i$ was set to one for an item identified correctly as a measure or non-measure of the dth domain by the ith judge, and to zero otherwise. Three separate scores were calculated for each item:

(1) A score $S1_j$, was computed as the number of judges who rated the jth item correctly as a measure or non-measure (with regard to "format," "wording," "sizes of numbers," and "behavior") of a domain. $S1_j = \sum_{i=1}^{N_j} I1_i$.

(2) A score $S2_j$, was calculated as the number of judges who rated the jth item correctly (with regard to "Overall" fit) as a measure or non-measure of a domain. $S2_j$ was defined as $S2_j = \sum_{i=1}^{N_j} I2_i$.

(3) A score $S3_j$, was computed as the number of judges who matched the jth item correctly to a domain. $S3_j = \sum_{i=1} I3_i$.

The proportions $PJ1_k$, $PJ2_k$, and $PJ3_k$ of items identified correctly by the kth judge as measures or non-measures of a domain using (1) dimensional criteria as bases for judgment, (2) summative judgments and (3) a match between items and domains respectively, were computed. Let N_d be the number of items in the dth domain and let $I1_{kj}$, $I2_{kj}$ and $I3_{kj}$ denote the respective scores derived from judgments on all dimensions of rating, the "Overall" dimension of rating and matching judgments for the kth judge on the jth item. Each of the $I1_{kj}$, $I2_{kj}$ and $I3_{kj}$ was set to one for the jth item identified correctly as a measure or non-measure of the dth domain by the kth judge, using all dimensions of rating, the "overall" dimension of rating and matching method respectively, and to zero otherwise. For the kth judge, $PJ_{ik} = (\sum_{j=1}^{N_d} I_{ikj})/N_d$, where i=1, 2 and 3. The $S1_j$'s, $S2_j$'s, $S3_j$'s, $PJ1_k$'s, $PJ2_k$'s, and $PJ3_k$'s were used with the index TR-87 to determine the internal consistency of the teachers' judgments of items over domains.

In Chapter 3, we present the results of the various hypothesis tests used in this research.

CHAPTER 3

HYPOTHESIS TESTING AND RESULTS

In this Chapter we present and test hypotheses that are relevant to the determination of the content validity of Grade 4 mathematics test items. The results of the statistical tests of the hypotheses are evaluated and interpreted.

3.1 Content Validity of Individual Test Items

The number and percentage of teachers expressing judgments on whether individual test items are measures of the domains for which they have been constructed, are tabulated in Appendix C.1 for the Rating and Matching methods of eliciting judgments. The number and proportion of teachers expressing correct judgments on whether individual test items are measures of these domains are presented in Appendix C.3, for both the Rating and the Matching methods.

A question naturally arises as to the worth of each individual test as a measure of the domain for which it was written. For each test item, the percentages of "Yes" and "No" or "Match" and "No Match" in Appendix C.1 are used with the proportion information in Appendix C.3 to test the null hypothesis that no decision can be made on the worth of a test item as a measure of the domain against the alternative hypothesis that the test item is (or is not) a measure of the domain. Note that the percentages of "Yes" and "No" or "Match" and "No Match" in Appendix C.1 were used to determine whether or not the judges agreed that an item was a measure of the domain. For instance, if the percentage expressing indicated "Yes" judgments was significantly larger than the percentage expressing indicated "No" judgments, the conclusion was that the judges agreed that the item was a measure of the domain; however, if the percentage of indicated "No" judgments was significantly larger than the percentage expressing "Yes" judgments, the conclusion was that the judges agreed that the item was not a measure of the domain. Note also that the alternative hypothesis above is directional. When testing each null hypothesis, the Type I error α was set to .05 to provide reasonable power for the test.

Three sets of results were derived from the item ratings and the matching data in Appendix C.1. First, the proportion P1 of teachers (in Appendix C.3) expressing correct judgments on whether, overall, an item is a measure of the domain, was used to assess the worth of each test item. Figure 5 contains the results of using an "Overall" rating to assess the content validity of the Grade 4 test items and other test items that were specially constructed for the research. Notice that the results are tabulated by domain. To illustrate the interpretation of the values in each of the tables in Figure 5, consider the results tabulated for Domain 1. According to the numbers in the table for Domain 1, we conclude with five percent chance of committing a Type I error for each item rated, that:

(a) Twenty test items which were valid in reality were rated correctly on the "Overall" dimension as measures of the domain,

(b) fourteen test items which were invalid in reality were rated correctly on the "Overall" dimension as non-measures of the domain, and

(c) two items which were invalid in reality resulted in inconclusive decisions with regard to the worth of the items as measures of the domain.

Figure 6 shows the table of results for the entire collection of items when the individual test items were rated on the "Overall" dimension. With five percent probability of rejecting a null hypothesis that is true for each item rated, we conclude that:

(a) one-hundred and fifty-three valid test items in reality were rated correctly on the "Overall" dimension as measures of their domains,

(b) three valid test items resulted in inconclusive assessment of the worth of the items as measures of the domains,

(c) two invalid test items in reality were rated incorrectly on the "Overall" dimension as measures of the domains,

(d) fourty-eight invalid test items were rated correctly on the "Overall" dimension as non-measures of the domains, and

(e) eighteen invalid test items resulted in inconclusive evaluation of the worth of the items as measures of their domains.

Second, the proportion P2 (in Appendix C.3) of teachers expressing correct judgments on all dimensions ("format," "numbers," "wording," and "behavior") of rating was used to estimate the content validity of each test item. The results of using all dimensions of rating to estimate the content validity of the individual test items are tabulated by domain in Figure 7. The table of results for the entire collection of items is shown in Figure 8. Note that the numbers in the tables of Figures 7 and 8 are interpreted in a manner that is analogous to the interpretations given above for numbers for Domain 1 of Figure 5 and the results in Figure 6. For example, consider the table for Domain 6 in Figure 7. With five percent risk of concluding that an item whose worth cannot be determined, was or was not content valid, we find that:

(a) Ten valid test items in reality were rated correctly as measures of the domain, on all dimensions of judgment,

(b) three invalid test items were rated incorrectly as measures of the domain, on all dimensions of judgment, and

(c) five invalid test items resulted in inconclusive assessment of the worth of the items as measures of the domain.

Third, the proportion P3 of teachers (in Appendix C.3) who correctly matched an item to a domain was used to establish a decision on the content validity of each test item. The results of using the Matching method to obtain information for estimating the content validity of the individual test items are tabulated by domain, in Figure 9. For the entire collection of items, the table of results is presented in Figure 10. Again, interpretations given above for the numbers for Domain 1 of Figure 5 and results in Figure 6 also apply to the numbers in the tables of Figures 9 and 10. For example, consider the table for Domain 11 in Figure 9. At the five percent level of significance for each item considered, we conclude that:

(a) Ten test items which were valid in reality were matched correctly as measures of their domain, and

(b) eight invalid test items were matched correctly as non-measures of their domain.

DOMAIN:- 1

		REALITY	
		Valid	Invalid
RATING (P1)	Valid	20	0
	Invalid	0	14
	Inconclusive	0	2

DOMAIN:- 2

		REALITY	
		Valid	Invalid
RATING (P1)	Valid	9	0
	Invalid	0	2
	Inconclusive	0	3

DOMAIN:- 3

		REALITY	
		Valid	Invalid
RATING (P1)	Valid	10	1
	Invalid	0	0
	Inconclusive	0	2

DOMAIN:- 4

		REALITY	
		Valid	Invalid
RATING (P1)	Valid	20	0
	Invalid	0	2
	Inconclusive	0	0

DOMAIN:- 5

		REALITY	
		Valid	Invalid
RATING (P1)	Valid	10	0
	Invalid	0	1
	Inconclusive	0	0

DOMAIN:- 6

		REALITY	
		Valid	Invalid
RATING (P1)	Valid	10	0
	Invalid	0	5
	Inconclusive	0	3

Figure 5 3 x 2 Contingency Tables for the Accuracy of Rating Judgments on the "Overall" Dimension

Figure 5 continued--

DOMAIN:- 7

		REALITY	
		Valid	Invalid
RATING (P1)	Valid	20	0
	Invalid	0	5
	Inconclusive	0	1

DOMAIN:- 8

		REALITY	
		Valid	Invalid
RATING (P1)	Valid	9	0
	Invalid	0	5
	Inconclusive	0	0

DOMAIN:- 9

		REALITY	
		Valid	Invalid
RATING (P1)	Valid	10	0
	Invalid	0	0
	Inconclusive	0	1

DOMAIN:- 10

		REALITY	
		Valid	Invalid
RATING (P1)	Valid	15	1
	Invalid	0	6
	Inconclusive	3	3

DOMAIN:- 11

		REALITY	
		Valid	Invalid
RATING (P1)	Valid	10	0
	Invalid	0	7
	Inconclusive	0	1

DOMAIN:- 12

		REALITY	
		Valid	Invalid
RATING (P1)	Valid	10	0
	Invalid	0	1
	Inconclusive	0	2

Figure 5 3 x 2 Contingency Tables for the Accuracy of Rating Judgments on the "Overall" Dimension

ALL DOMAINS

		REALITY	
		Valid	Invalid
RATING (P1)	Valid	153	2
	Invalid	0	48
	Inconclusive	3	18

Figure 6 3 x 2 Contingency Table for the Accuracy of Rating Judgments on the "Overall" Dimension, for the Entire Collection of Test Items

DOMAIN:- 1

		REALITY	
		Valid	Invalid
RATING (P2)	Valid	20	3
	Invalid	0	4
	Inconclusive	0	9

DOMAIN:- 2

		REALITY	
		Valid	Invalid
RATING (P2)	Valid	9	2
	Invalid	0	0
	Inconclusive	0	3

DOMAIN:- 3

		REALITY	
		Valid	Invalid
RATING (P2)	Valid	10	1
	Invalid	0	0
	Inconclusive	0	2

DOMAIN:- 4

		REALITY	
		Valid	Invalid
RATING (P2)	Valid	20	0
	Invalid	0	0
	Inconclusive	0	2

Figure 7 3 x 2 Contingency Tables for the Accuracy of Rating Judgments on all Dimensions

Figure 7 continued--

DOMAIN:- 5

		REALITY	
		Valid	Invalid
RATING (P2)	Valid	10	0
	Invalid	0	0
	Inconclusive	0	1

DOMAIN:- 6

		REALITY	
		Valid	Invalid
RATING (P2)	Valid	10	3
	Invalid	0	0
	Inconclusive	0	5

DOMAIN:- 7

		REALITY	
		Valid	Invalid
RATING (P2)	Valid	20	0
	Invalid	0	1
	Inconclusive	0	5

DOMAIN:- 8

		REALITY	
		Valid	Invalid
RATING (P2)	Valid	9	1
	Invalid	0	1
	Inconclusive	0	3

DOMAIN:- 9

		REALITY	
		Valid	Invalid
RATING (P2)	Valid	10	0
	Invalid	0	0
	Inconclusive	0	1

DOMAIN:- 10

		REALITY	
		Valid	Invalid
RATING (P2)	Valid	15	4
	Invalid	0	0
	Inconclusive	3	6

Figure 7 3 x 2 Contingency Tables for the Accuracy of Rating Judgments on all Dimensions

Figure 7 continued--

DOMAIN:- 11

		REALITY	
		Valid	Invalid
RATING (P2)	Valid	10	0
	Invalid	0	1
	Inconclusive	0	7

DOMAIN:- 12

		REALITY	
		Valid	Invalid
RATING (P2)	Valid	10	2
	Invalid	0	0
	Inconclusive	0	1

Figure 7 3 x 2 Contingency Tables for the Accuracy of Rating Judgments on all Dimensions

ALL DOMAINS

		REALITY	
		Valid	Invalid
RATING (P2)	Valid	153	16
	Invalid	0	7
	Inconclusive	3	45

Figure 8 3 x 2 Contingency Table for the Accuracy of Rating Judgments on all Dimensions, for the Entire Collection of Test Items

DOMAIN:- 1

		REALITY	
		Valid	Invalid
MATCHING (P3)	Valid	20	1
	Invalid	0	15
	Inconclusive	0	0

DOMAIN:- 2

		REALITY	
		Valid	Invalid
MATCHING (P3)	Valid	9	0
	Invalid	0	1
	Inconclusive	0	4

DOMAIN:- 3

		REALITY	
		Valid	Invalid
MATCHING (P3)	Valid	10	0
	Invalid	0	1
	Inconclusive	0	2

DOMAIN:- 4

		REALITY	
		Valid	Invalid
MATCHING (P3)	Valid	20	0
	Invalid	0	2
	Inconclusive	0	0

DOMAIN:- 5

		REALITY	
		Valid	Invalid
MATCHING (P3)	Valid	10	0
	Invalid	0	1
	Inconclusive	0	0

DOMAIN:- 6

		REALITY	
		Valid	Invalid
MATCHING (P3)	Valid	10	0
	Invalid	0	5
	Inconclusive	0	3

Figure 9 3 x 2 Contingency Tables for the Accuracy of the Matching Method

Figure 9 continued--

DOMAIN:- 7

		REALITY	
		Valid	Invalid
MATCHING (P3)	Valid	20	0
	Invalid	0	5
	Inconclusive	0	1

DOMAIN:- 8

		REALITY	
		Valid	Invalid
MATCHING (P3)	Valid	9	0
	Invalid	0	3
	Inconclusive	0	2

DOMAIN:- 9

		REALITY	
		Valid	Invalid
MATCHING (P3)	Valid	10	0
	Invalid	0	0
	Inconclusive	0	1

DOMAIN:- 10

		REALITY	
		Valid	Invalid
MATCHING (P3)	Valid	15	0
	Invalid	0	6
	Inconclusive	3	4

DOMAIN:- 11

		REALITY	
		Valid	Invalid
MATCHING (P3)	Valid	10	0
	Invalid	0	8
	Inconclusive	0	0

DOMAIN:- 12

		REALITY	
		Valid	Invalid
MATCHING (P3)	Valid	10	0
	Invalid	0	1
	Inconclusive	0	2

Figure 9 3 x 2 Contingency Tables for the Accuracy of the Matching Method

ALL DOMAINS

		REALITY	
		Valid	Invalid
MATCHING (P3)	Valid	153	1
	Invalid	0	48
	Inconclusive	3	19

Figure 10 3 x 2 Contingency Table for the Accuracy of the Matching Method, for the Entire Collection of Test Items

3.2 Accuracy of the Rating and Matching Methods

So far, we have summarized by domain and overall, the results and correctness of judgments of the content validity of individual test items. One major question of concern to the present research is, are the methods used to elicit judgments of the content validity of test items accurate? To investigate this question, three sets of hypotheses were tested separately for the Rating and Matching methods:

(1) The null hypothesis that the population proportion of false-positive errors equals zero, against the alternative hypothesis that the population proportion of false-positive errors is greater than zero,

(2) The null hypothesis that the population proportion of false-negative errors equals zero, against the alternative hypothesis that the population proportion of false-negative errors is greater than zero, and

(3) The null hypothesis that the population correlation between judges' judgments and the state of "Reality" equals zero, against the alternative hypothesis that there is a positive correlation between judges' judgments and the state of "Reality."

For each test and for the entire collection of tests, the results derived from testing the accuracy of the Rating and Matching methods are tabulated in Tables 5, 6 and 7. Note that the number of test items in each domain has a bearing on the statistical significance of each proportion of false-positive or false-negative errors that is presented in Tables 5-7.

By testing the statistical significance of the worst-case and the best-case proportions of the false-positive and false-negative errors in Tables 5-7, the following results were obtained at a .01 Type I error level:

(1) For Domains 1, 2, 4-12 tests and for the entire collection of tests, the null hypothesis that the best-case population proportion of false-positive errors equals zero when individual test items were rated on the "Overall" dimension of judgment could not be rejected; the best-case proportion of false-positive errors for Domain 3 was statistically significant. The null hypothesis that the worst-case population proportion of false-positive errors equals zero could not be retained for Domains 2, 3,

6, 7, 9, 10 and 12 tests and for the entire collection of tests; the worst-case proportions of false-positive errors for Domains 1, 4, 5, 8 and 11 were not statistically significant.

(2) For each test and for the entire collection of tests, the data support the null hypothesis that the worst- or the best-case population proportion of false-negative errors equals zero when individual test items were rated on the "Overall" dimension of judgment.

(3) For Domains 1-12 tests and for the entire collection of tests, the data do not support the null hypothesis that the worst-case population proportion of false-positive errors equals zero when test items were rated individually on "format," "wording," "number" and "behavior" dimensions of judgment. For Domains 1-3, 6, 8, 10, 12 tests and for the entire collection of tests, the data suggest that the that best-case proportion of false-positive errors was greater than zero when test items were rated individually on "format," "wording," "number" and "behavior" dimensions of judgment; however, for Domains 4, 5, 7 9 and 11 tests, the data support the null hypothesis that the best-case population proportion equals zero.

(4) For each test and for the entire collection of tests, the data provide support for the null hypothesis that the worst- or the best-case proportion of false-negative

errors equals zero when test items were rated individually on "format," "wording," "number" and "behavior" dimensions of judgment.

(5) For each test and for the entire collection of tests, the null hypothesis that the best-case population proportion of false-positive errors equals zero when individual test items were matched to their corresponding domains could not be rejected. The null hypothesis that the worst-case population proportion of false-positive errors equals zero could not be retained for Domains 2, 3, 6-10 and 12 tests and for the entire collection of tests; the worst-case proportions of false-positive errors for Domains 1, 4, 5 and 11 were statistically significant.

(6) For each test and for the entire collection of tests, the data support the null hypothesis that the worst- or the best-case population proportion of false-negative errors equals zero when individual test items were matched to their corresponding domains.

An examination of the results above and the intervals between the worst- and best-case proportions of false-positive errors and between the worst- and best-case proportions of false-negative errors reveals two findings:

(1) At a .01 Type I error level of significance, it is unlikely that a statistically significant proportion of false-negative errors will result when either the Rating

method or the Matching method is used to elicit judgments of the content validity of mathematics achievement test items, and

(2) When mathematics achievement test items are rated on the basis of "Overall" quality or matched to their corresponding domains, the resulting proportion of false-positive errors is likely to be lower than the proportion of false-positive errors when the items are rated individually on "format," "wording," "number" and "behavior" dimensions of judgment.

The worst- and best-case values of Phi and the corresponding values of the Chi-square test statistic are presented in Tables 5, 6 and 7, for each test and for the entire collection of tests. Because the number of invalid items in "Reality" that were matched or rated as invalid was zero for certain domains, and division by zero is mathematically undefined, some of the worst case values of Phi and Chi-square are undefined in Tables 5-7. The value of a Chi-square random variable with one degree of freedom for which the right-tail probability is .01 is 6.64. Therefore, according to the worst- and best-case values of Phi and Chi-square, the following results were obtained at a .01 Type I error level:

(1) For each test and for the entire collection of tests, the best-case correlations between "Overall" rating

judgments and the state of "Reality" were statistically significant. For Domains 2, 10 and 12 the worst-case correlations between "Overall" rating judgments and the state of "Reality" were not statistically significant; the worst-case correlations between "Overall" rating judgments and the state of "Reality" for Domains 1, 3-8, 11 and for the entire collection of tests were statistically significant.

(2) The best-case correlations between rating judgments on individual dimensions and the state of "Reality" for Domains 1-11 and for the entire collection of tests were statistically significant; the correlation for Domain 12 was not statistically significant. Although the worst-case correlation between the rating of individual dimension judgments and the state of "Reality" was statistically significant for the entire collection of tests, no worst-case correlation that could be estimated for any domain was statistically significant.

(3) For each test and for the entire collection of tests, the best-case correlations between matching judgments and the state of "Reality" were statistically significant. For Domains 2, 3, 10 and 12 the worst-case correlations between matching judgments and the state of "Reality" were not statistically significant. However, correlations between matching judgments and the state of "Reality" for

Domains 1, 4-8, 11 and for the entire collection of tests were statistically significant.

A close examination of the widths of the intervals containing the range of estimated correlations (the intervals between the worst-case and the best-case correlations) for the entire collection of tests leads to the following conclusions:

(1) There tends to be a strong positive relationship between rating judgments on the "Overall" dimension and the state of "Reality." There tends to be a strong positive relationship between matching judgments and the state of "Reality."

(2) There tends to be at least a small positive relationship between rating judgments on all dimensions and the state of "Reality."

DOMAIN	WcPfp	BcPfp	WcPfn	BcPfn	WcPhi	WcChi-Sq	BcPhi	BcChi
1	.125	.00	.00	.00	.89	28.64	1.00	36.00
2	.600*	.00	.00	.00	.55	4.20	1.00	14.00
3	1.000*	.33*	.00	.00	#	#	.78	7.88
4	.000	.00	.00	.00	1.00	22.00	1.00	22.00
5	.000	.00	.00	.00	1.00	11.00	1.00	11.00
6	.375*	.00	.00	.00	.69	8.65	1.00	18.00
7	.167*	.00	.00	.00	.89	20.63	1.00	26.00
8	.000	.00	.00	.00	1.00	14.00	1.00	14.00
9	1.000*	.00	.00	.00	#	#	1.00	11.00
10	.400*	.100	.167	.00	.45	5.53	.92	23.87
11	.125	.00	.00	.00	.89	14.32	1.00	18.00
12	.667*	.00	.00	.00	.53	3.61	1.00	13.00
ALL	.294*	.029	.019	.000	.753	126.98	.979	214.66

SYMBOL	MEANING
WcPfp	Worst Case Proportion of False Positive Errors
BcPfp	Best Case Proportion of False Positive Errors
WcPfn	Worst Case Proportion of False Negative Errors
BcPfn	Best Case Proportion of False Negative Errors
WcPhi	Worst Case value of Phi
WcChi-Sq	Worst Case Chi-Square for the value of WcPhi
BcPhi	Best Case value of Phi
BcChi	Best Case Chi-Square for the value of BcPhi
*	Statistically significant at a .01 Type I error level
#	Undefined value

Table 5 Accuracy Results for the Rating Method P1

DOMAIN	WcPfp	BcPfp	WcPfn	BcPfn	WcPhi	WcChi-Sq	BcPhi	BcChi
1	.750*	.188*	.00	.00	.39	5.63	0.84	25.43
2	1.000*	.400*	.00	.00	#	#	0.70	6.87
3	1.000*	.333*	.00	.00	#	#	0.78	7.88
4	1.000*	.000	.00	.00	#	#	1.00	22.00
5	1.000*	.000	.00	.00	#	#	1.00	11.00
6	1.000*	.375*	.00	.00	#	#	0.69	8.65
7	.833*	.000	.00	.00	.37	3.47	1.00	26.00
8	.800*	.200*	.00	.00	.37	1.94	0.85	10.08
9	1.000*	.000	.00	.00	#	#	1.00	11.00
10	1.000*	.400*	.167	.00	.26	1.87	0.70	13.75
11	.875*	.000	.00	.00	.27	1.32	1.00	18.00
12	1.000*	.667*	.00	.00	#	#	0.53	3.61
ALL	.879*	.235*	.019	.000	.186	7.78	.833	155.36

SYMBOL	MEANING
WcPfp	Worst Case Proportion of False Positive Errors
BcPfp	Best Case Proportion of False Positive Errors
WcPfn	Worst Case Proportion of False Negative Errors
BcPfn	Best Case Proportion of False Negative Errors
WcPhi	Worst Case value of Phi
WcChi-Sq	Worst Case Chi-Square for the value of WcPhi
BcPhi	Best Case value of Phi
BcChi	Best Case Chi-Square for the value of BcPhi
*	Statistically significant at a .01 Type I error level
#	Undefined value

Table 6 Accuracy Results for the Rating Method P2

DOMAIN	WcPfp	BcPfp	WcPfn	BcPfn	WcPhi	WcChi-Sq	BcPhi	BcChi
1	.063	.063	.00	.00	.95	32.14	0.95	32.14
2	.800*	.000	.00	.00	.37	1.94	1.00	14.00
3	.667*	.000	.00	.00	.53	3.61	1.00	13.00
4	.000	.000	.00	.00	1.00	22.00	1.00	22.00
5	.000	.000	.00	.00	1.00	11.00	1.00	11.00
6	.375*	.000	.00	.00	.69	8.65	1.00	18.00
7	.167*	.000	.00	.00	.89	20.63	1.00	26.00
8	.400*	.000	.00	.00	.70	6.87	1.00	14.00
9	1.000*	.000	.00	.00	#	#	1.00	11.00
10	.400*	.000	.167	.00	.44	5.53	1.00	28.00
11	.000	.000	.00	.00	1.00	18.00	1.00	18.00
12	.667*	.000	.00	.00	.53	3.61	1.00	13.00
ALL	.294*	.015	.019	.000	.753	126.98	.989	219.30

SYMBOL	MEANING
WcPfp	Worst Case Proportion of False Positive Errors
BcPfp	Best Case Proportion of False Positive Errors
WcPfn	Worst Case Proportion of False Negative Errors
BcPfn	Best Case Proportion of False Negative Errors
WcPhi	Worst Case value of Phi
WcChi-Sq	Worst Case Chi-Square for the value of WcPhi
BcPhi	Best Case value of Phi
BcChi	Best Case Chi-Square for the value of BcPhi
*	Statistically significant at a .01 Type I error level
#	Undefined value

Table 7 Accuracy Results for the Matching Method P3

3.3 Accuracy of Rating Judgments on Individual Dimensions

Although the observed correlation between the state of "Reality" and rating judgments on all dimensions was statistically significant, this correlation tends to be low (see the worst- and best-case values of Phi for "ALL" domains in Table 6). Since the test items were rated separately on each of "format," "number," "wording" and "behavior" dimensions, one major question of interest to the current research is, are the content specialists accurate in their judgments of test items over the rating dimensions? This question is of particular concern since information derived from the rating judgments on individual dimensions will be used to rectify "invalid" test items. Instead of investigating this research question for each test, the entire collection of tests was used because the expected numbers of false-positive and false-negative errors were zeros for many of the individual tests, making it impossible to estimate (mathematically) the population correlations.

For the entire collection of tests, Figure 11 contains the results of rating judgments on individual dimensions. To illustrate the interpretations of the results, consider the results tabulated for rating judgments on the "Numbers" dimension. According to these results, we conclude with five percent chance of committing a Type I error that:

(a) One-hundred and sixty-eight test items for which the sizes of their numbers were valid in reality were rated correctly on the "Numbers" dimension of judgment as measures of magnitudes of numbers defined by their domains.

(b) One test item for which the numbers in the item were in accordance with the domain definition in reality was rated incorrectly on the "Numbers" dimension as invalid.

(c) Thirteen test items whose magnitudes of numbers do not correspond in reality to numbers defined by their domains were rated correctly as non-measures of the sizes of numbers specified by the domains.

(d) Six test items for which the magnitudes of their numbers were valid in reality resulted in inconclusive decisions with regard to the sizes of their numbers as measures of numbers defined by their domains.

(e) Eleven test items for which the sizes of their numbers were invalid in reality resulted in inconclusive evaluations of their numbers as measures of numbers defined by their domains.

For the entire collection of tests, Table 8 contains the accuracy results when individual rating dimensions were used to elicit judgments. By testing the statistical significance of the worst-case and the best-case proportions of the false-positive and false-negative errors in Table 8 at a .01 Type I error level, the following results were obtained:

(1) For the entire collection of tests, the null hypothesis that the worst-case population proportion of false-positive errors equals zero when individual test items were rated either on the "Format" or "Numbers" or "Wording" or "Behavior" dimensions of judgments could not be retained.

(2) For the entire collection of tests, the best-case proportion of false-positive errors which resulted when each of "Format," "Number" and "Behavior" dimensions was used to elicit judgments, was not statistically significant; the best-case proportion of false-positive errors which resulted from using the "Wording" dimension to elicit judgments was statistically significant.

(3) For the entire collection of tests, the null hypothesis that the population proportion of false-negative errors equals zero when individual test items were rated either on the "Format" or "Numbers" or "Wording" or "Behavior" dimensions of judgment could not be rejected.

(4) The data support the alternative hypothesis that there is a positive worst-case population correlation between the state of "Reality" and the rating judgments derived from each of the "Format," "Numbers" and "Behavior" dimensions; the correlation between rating judgments on the "Wording" dimension and the state of "Reality" was not

statistically significant. For each of the four rating dimensions, the data support the alternative hypothesis that the best-case population correlation between the state of "Reality" and rating judgments derived from each dimension is positive.

An examination of the widths of the intervals containing ranges of estimated correlations between the state of "Reality" and each of the "Format," "Numbers," "Wording" and "Behavior" rating dimensions reveals that: The accuracy of rating judgments tends to decrease as judges proceeded from (1) assessing the appropriateness of item "Format" to evaluating sizes of "Numbers" as measures of numbers defined by the domain, (2) evaluating the magnitudes of "Numbers" to assessing whether or not the "Wording" used to state the items was within the vocabulary recognition of Grade 4 students, and (3) assessing the "Wording" of an item to evaluating whether or not the item elicited knowledge measured by the domain.

		REALITY	
		Valid	Invalid
RATING OF FORMAT	Valid	182	1
	Invalid	0	21
	Inconclusive	9	11

		REALITY	
		Valid	Invalid
RATING OF NUMBERS	Valid	168	0
	Invalid	1	13
	Inconclusive	6	11

		REALITY	
		Valid	Invalid
RATING OF WORDING	Valid	22	1
	Invalid	0	1
	Inconclusive	0	2

		REALITY	
		Valid	Invalid
RATING OF BEHAVIOR	Valid	174	0
	Invalid	0	12
	Inconclusive	7	31

Figure 11 3 x 2 Contingency Tables for the Accuracy of Rating Judgments on Individual Dimensions

DIMENSION	WcPfp	BcPfp	WcPfn	BcPfn	WcPhi	WcChi	BcPhi	BcChi
Format	.364*	.030	.047	.000	.613	84.23	.982	216.08
Numbers	.458*	.000	.040	.006	.562	62.54	1.00	198.00
Wording	.750*	.25*	.000	.000	.469	5.72	.847	18.65
Behavior	.721*	.000	.039	.000	.340	25.87	1.00	224.00

SYMBOL	MEANING
WcPfp	Worst Case Proportion of False Positive Errors
BcPfp	Best Case Proportion of False Positive Errors
WcPfn	Worst Case Proportion of False Negative Errors
BcPfn	Best Case Proportion of False Negative Errors
WcPhi	Worst Case value of Phi
WcChi	Worst Case Chi-Square for the value of WcPhi
BcPhi	Best Case value of Phi
BcChi	Best Case Chi-Square for the value of BcPhi
*	Statistically significant at a .01 Type I error level
#	Undefined value

Table 8 Accuracy Results for the Rating Judgments on Individual Dimensions

3.4 Similarity of the Rating and Matching Methods

In Section 3.2 we found that rating test items on the basis of their "Overall" quality and matching test items to their corresponding domains tended to be accurate techniques for eliciting judgments of content validity. However, there was mixed feeling on the accuracy of judgments when test items were rated separately on the "Format," "Numbers, "Wording" and "Behavior" dimensions. Two questions naturally arise: (1) Is rating test items on the basis of "Overall" quality after rating the items separately on each of "Format," "Numbers," "Wording" and "Behavior" dimensions similar to matching the items to their corresponding domains? and (2) is the accuracy of rating judgments on all dimensions the same as the accuracy of matching the test items to their corresponding domains?

For each test and for the entire collection of tests, Figures 12 and 13 contain the respective data for investigating the similarity between rating judgments on the "Overall" dimension and matching judgments. For each test and for the entire collection of tests, the respective data for examining the similarity between rating judgments on all dimensions and matching judgments are presented in Figures 14 and 15. To illustrate the interpretations of the values in each of the tables in Figures 12-15, consider the results tabulated in Figure 15. According to numbers in the cells of this table we conclude with five percent risk of committing a Type I error for each item that:

(1) One hundred and fify-four items for which the Matching method resulted in "valid" judgments also resulted in "valid" judgments when rated on all dimensions,

(2) four items matched as invalid to their domains were rated on all dimensions as valid,

(3) ten items for which no consensus could be reached on match between the items and their domains, were rated on all dimensions as valid measures of their domains,

(4) eight items that were matched as invalid to their domains were also rated on all dimensions as invalid,

(5) thirty-seven test items for which the Matching method resulted in "invalid" judgments, resulted in "inconclusive" judgments when rated on all dimensions, and

(6) eleven test items for which no consensus could be reached on the correspondence between the items and their domains also resulted in "inconclusive" judgments when rated on all dimensions.

DOMAIN:- 1

		MATCHING		
		Valid	Inval	Incon
	Valid	20	0	0
RATING (Pl)	Invalid	0	14	0
	Inconclusive	1	1	0

DOMAIN:- 2

		MATCHING		
		Valid	Inval	Incon
	Valid	9	0	0
RATING (Pl)	Invalid	0	1	1
	Inconclusive	0	0	3

DOMAIN:- 3

		MATCHING		
		Valid	Inval	Incon
	Valid	10	0	1
RATING (Pl)	Invalid	0	0	0
	Inconclusive	0	1	1

DOMAIN:- 4

		MATCHING		
		Valid	Inval	Incon
	Valid	20	0	0
RATING (Pl)	Invalid	0	2	0
	Inconclusive	0	0	0

DOMAIN:- 5

		MATCHING		
		Valid	Inval	Incon
	Valid	10	0	0
RATING (Pl)	Invalid	0	1	0
	Inconclusive	0	0	0

DOMAIN:- 6

		MATCHING		
		Valid	Inval	Incon
	Valid	10	0	0
RATING (Pl)	Invalid	0	5	0
	Inconclusive	0	0	3

Figure 12 3 x 3 Contingency Tables for Comparing the Similarity of the Rating (Pl) and Matching Methods

Figure 12 continued--

DOMAIN:- 7

		MATCHING		
		Valid	Inval	Incon
RATING (P1)	Valid	20	0	0
	Invalid	0	5	0
	Inconclusive	0	0	1

DOMAIN:- 8

		MATCHING		
		Valid	Inval	Incon
RATING (P1)	Valid	9	0	0
	Invalid	0	3	2
	Inconclusive	0	0	0

DOMAIN:- 9

		MATCHING		
		Valid	Inval	Incon
RATING (P1)	Valid	10	0	0
	Invalid	0	0	0
	Inconclusive	0	0	1

DOMAIN:- 10

		MATCHING		
		Valid	Inval	Incon
RATING (P1)	Valid	15	0	1
	Invalid	0	6	0
	Inconclusive	0	0	6

DOMAIN:- 11

		MATCHING		
		Valid	Inval	Incon
RATING (P1)	Valid	10	0	0
	Invalid	0	7	0
	Inconclusive	0	1	1

DOMAIN:- 12

		MATCHING		
		Valid	Inval	Incon
RATING (P1)	Valid	10	0	0
	Invalid	0	1	0
	Inconclusive	0	2	0

Figure 12 3 x 3 Contingency Tables for Comparing the Similarity of the Rating (P1) and Matching Methods

ALL DOMAINS

		MATCHING		
		Valid	Invalid	Inconclusive
RATING (P1)	Valid	153	0	2
	Invalid	0	45	3
	Inconclusive	1	5	15

Figure 13 3 x 3 Contingency Table for Comparing the Similarity of the Rating (P1) and Matching Methods Using the Entire Collection of Test Items

DOMAIN:- 1

		MATCHING		
		Valid	Inval	Incon
RATING (P2)	Valid	21	2	0
	Invalid	0	4	0
	Inconclusive	0	9	0

DOMAIN:- 2

		MATCHING		
		Valid	Inval	Incon
RATING (P2)	Valid	9	0	1
	Invalid	0	1	0
	Inconclusive	0	1	2

DOMAIN:- 3

		MATCHING		
		Valid	Inval	Incon
RATING (P2)	Valid	10	0	1
	Invalid	0	0	0
	Inconclusive	0	1	1

DOMAIN:- 4

		MATCHING		
		Valid	Inval	Incon
RATING (P2)	Valid	20	0	0
	Invalid	0	2	0
	Inconclusive	0	0	0

Figure 14 3 x 3 Contingency Tables for Comparing the Similarity of the Rating (P2) and Matching Methods

Figure 14 continued--

DOMAIN:- 5

		MATCHING		
		Valid	Inval	Incon
RATING (P2)	Valid	10	0	0
	Invalid	0	0	0
	Inconclusive	0	1	0

DOMAIN:- 6

		MATCHING		
		Valid	Inval	Incon
RATING (P2)	Valid	10	1	2
	Invalid	0	0	0
	Inconclusive	0	4	1

DOMAIN:- 7

		MATCHING		
		Valid	Inval	Incon
RATING (P2)	Valid	20	0	0
	Invalid	0	1	0
	Inconclusive	0	4	1

DOMAIN:- 8

		MATCHING		
		Valid	Inval	Incon
RATING (P2)	Valid	9	0	1
	Invalid	0	1	0
	Inconclusive	0	2	1

DOMAIN:- 9

		MATCHING		
		Valid	Inval	Incon
RATING (P2)	Valid	10	0	0
	Invalid	0	0	0
	Inconclusive	0	0	1

DOMAIN:- 10

		MATCHING		
		Valid	Inval	Incon
RATING (P2)	Valid	15	1	3
	Invalid	0	0	0
	Inconclusive	0	5	4

Figure 14 3 x 3 Contingency Tables for Comparing the Similarity of the Rating (P2) and Matching Methods.

Figure 14 continued--

DOMAIN:- 11

		MATCHING		
		Valid	Inval	Incon
RATING (P2)	Valid	10	0	0
	Invalid	0	1	0
	Inconclusive	0	7	0

DOMAIN:- 12

		MATCHING		
		Valid	Inval	Incon
RATING (P2)	Valid	10	0	2
	Invalid	0	0	0
	Inconclusive	0	1	0

Figure 14 3 x 3 Contingency Tables for Comparing the Similarity of the Rating (P2) and Matching Methods

ALL DOMAINS

		MATCHING		
		Valid	Invalid	Inconclusive
RATING (P2)	Valid	154	4	10
	Invalid	0	8	0
	Inconclusive	0	37	11

Figure 15 3 x 3 Contingency Table for Comparing the Similarity of the Rating (P2) and Matching Methods Using the Entire Collection of Test Items

In Tables 5, 6 and 7 we presented the worst- and best-case proportions of false-positive and false-negative errors which resulted from rating and matching judgments. In order to illuminate the similarity between the Rating and the Matching methods, the bivariate plots for pairs of worst-case proportions and for pairs of best-case proportions across tests are presented in Figure 16, separately for the false-positive and false-negative errors. From the bivariate plots in Figure 16, the followings observations were made:

(1) Since the points in Figure 16a appear to be scattered unsystematically in the plane, there tended to be little or no linear relationship between the worst-case proportions of false-positive errors which resulted from rating judgments on all dimensions and rating judgments on the "Overall" dimension.

(2) The points in Figure 16b appear to be distributed unsystematically in the plane. Thus, there tended to be little or no linear relationship between the worst-case proportions of false-positive errors which resulted from rating judgments on all dimensions and matching judgments.

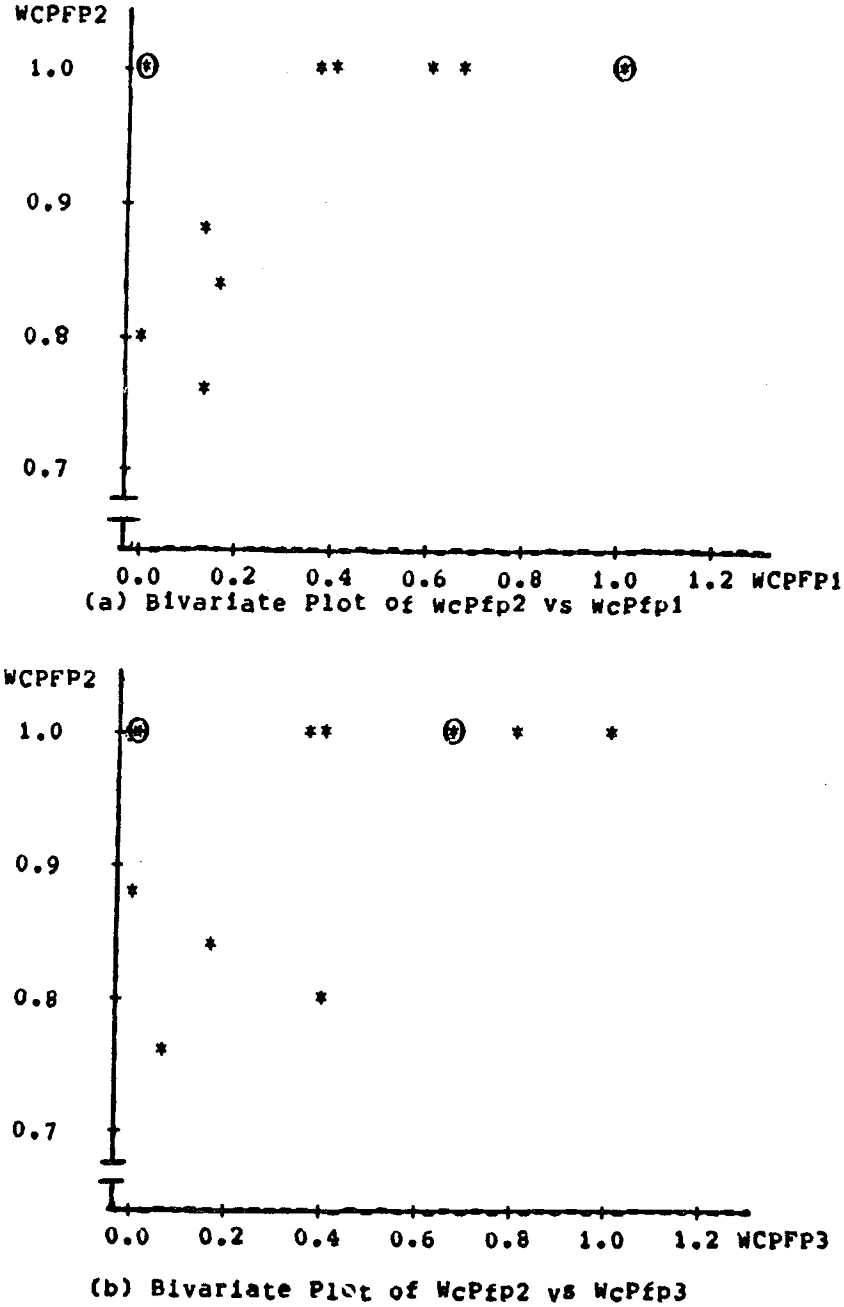
(3) Since the points in Figure 16c appear to be distributed in a straight line pattern, there appears to be a strong direct linear relationship between the worst-case proportions of false-positive errors which resulted from rating judgments on the "Overall" dimension and matching judgments.

(4) Figures 16d and 16e suggest that as the best-case proportions of false-positive errors tended to remain constant at zero with rating judgments on the "Overall" dimension and with matching judgments, the best-case

proportions of false-positive errors which resulted from rating judgments on all dimensions tended to increase.

(5) Figure 16f suggests that the best-case proportions of false-positive errors which resulted from rating judgments on the "Overall" dimension tended to remain constant at zero as the best-case proportions of false-positive errors which resulted from matching judgments remained constant at zero.

(6) Figures 16g-16l reveal that as the worst- and best-case proportions of false-negative errors which resulted from rating judgments remained constant at zero, the worst- and best-case proportions of false-negative errors which resulted from matching judgments remained constant at zero.



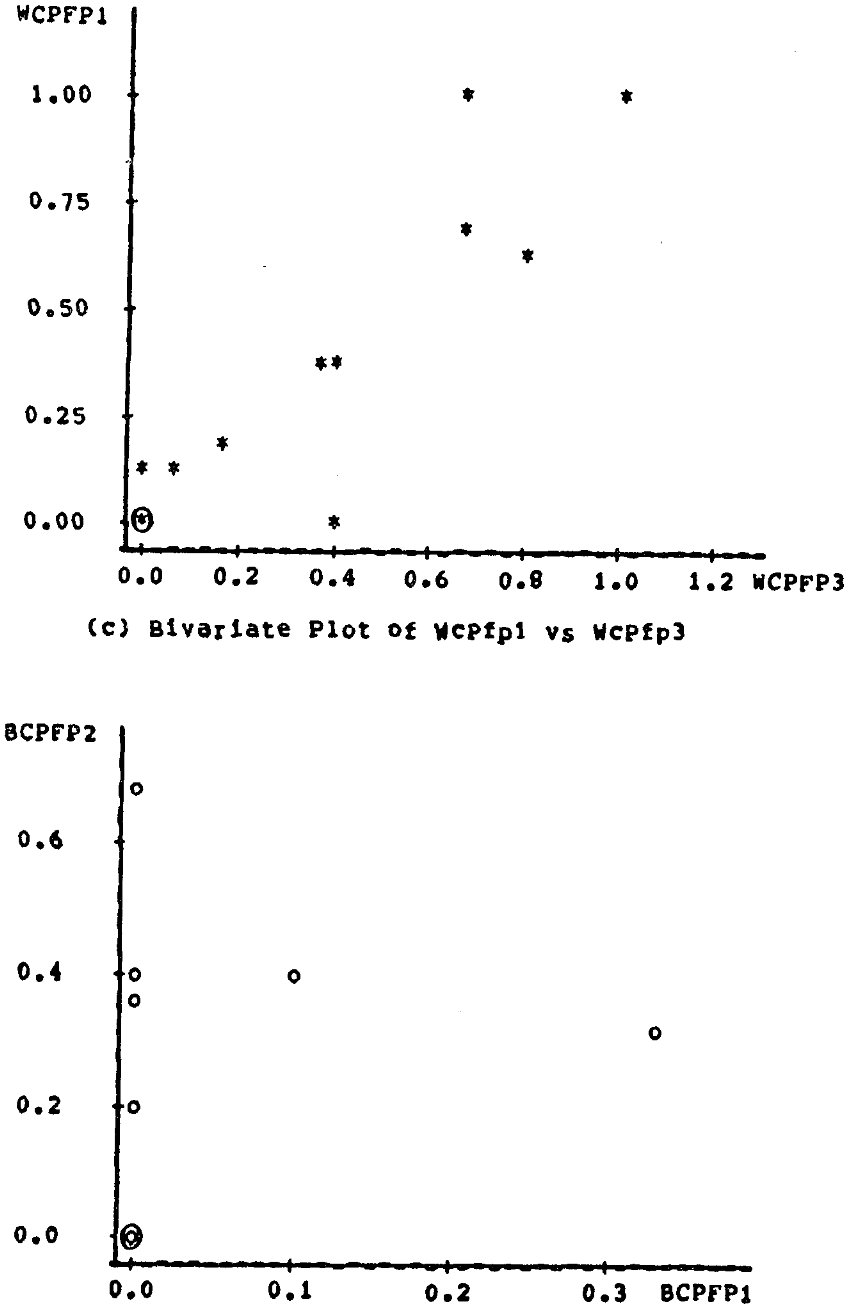
(a) Bivariate Plot of WcPfp2 vs WcPfp1

(b) Bivariate Plot of WcPfp2 vs WcPfp3

⊛ Signifies one or more hidden observations

Figure 16 Bivariate Plots of the Worst- and Best-Case Proportions of False-Positive and False-Negative Errors

Figure 16 continued--

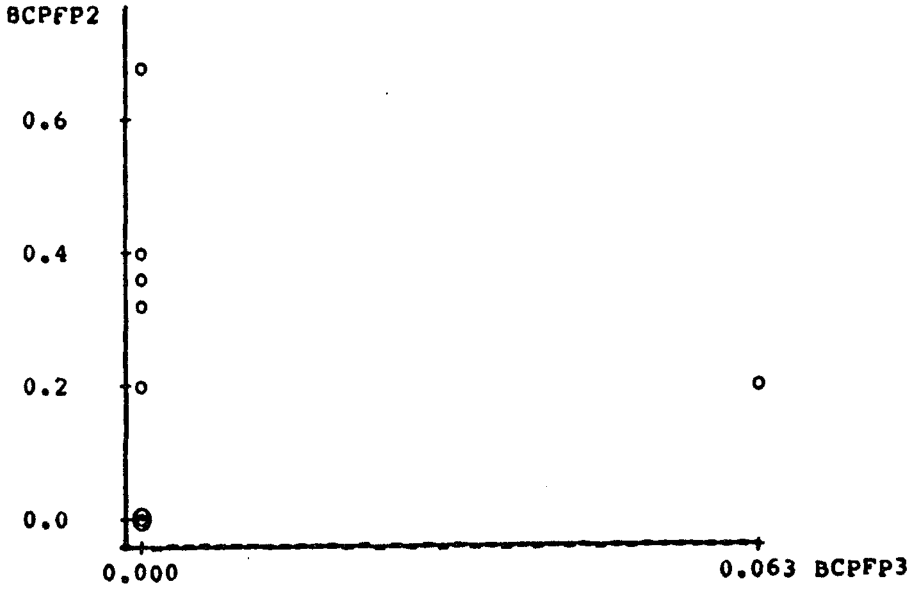


(c) Bivariate Plot of WCPfp1 vs WCPfp3

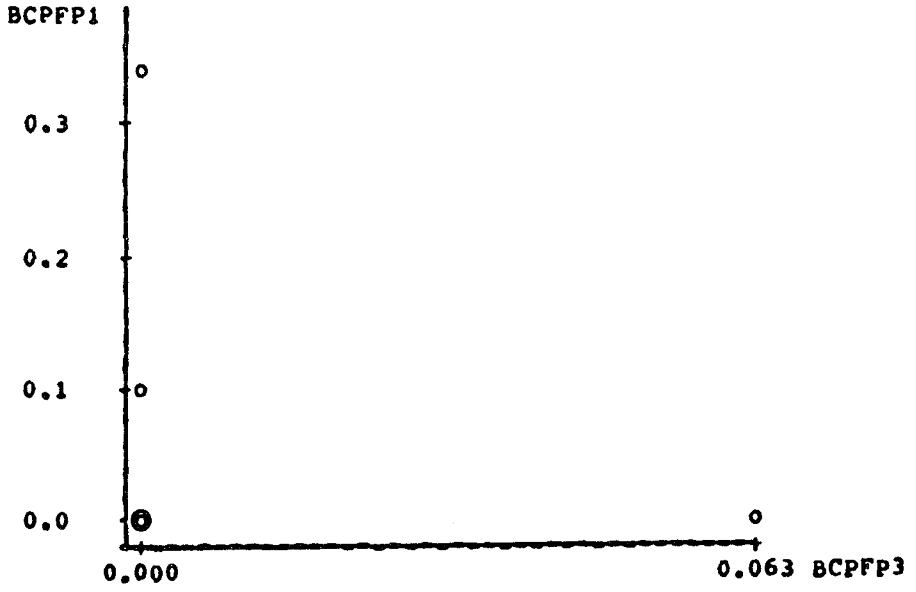
(d) Bivariate Plot of BCPfp2 vs BCPfp1

⊙ ⊛ Signifies one or more hidden observations

Figure 16 Bivariate Plots of the Worst- and Best-Case Proportions of False-Positive and False-Negative Errors



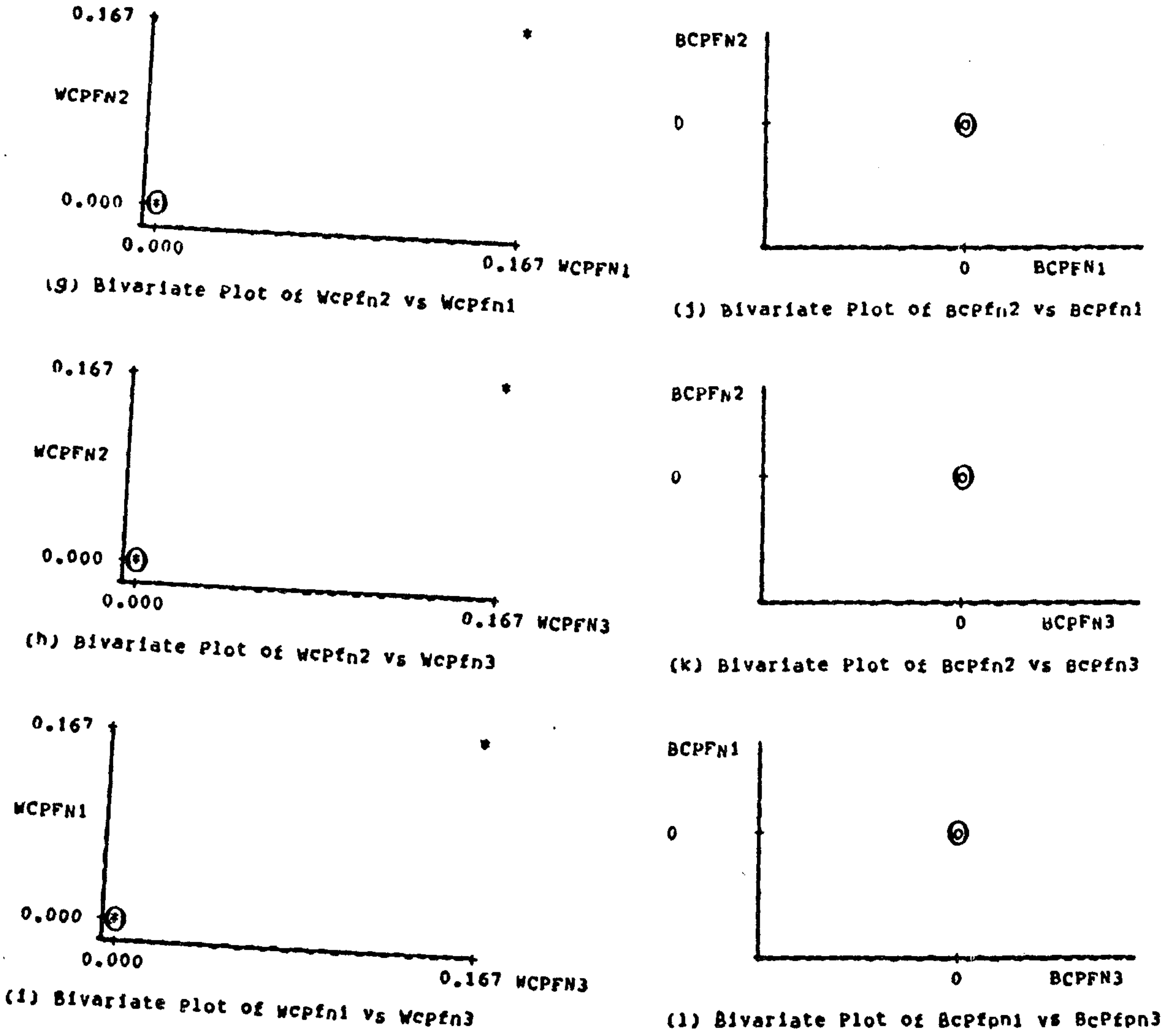
(e) Bivariate Plot of BCPfp2 vs BCPfp3



(f) Bivariate Plot of BCPfp1 vs BCPfpf3

O ⊖ Signifies one or more hidden observations

Figure 16 Bivariate Plots of the Worst- and Best-Case Proportions of False-Positive and False-Negative Errors



(g) Bivariate Plot of Wcpfn2 vs Wcpfn1

(j) Bivariate Plot of Bcpfn2 vs Bcpfn1

(h) Bivariate Plot of Wcpfn2 vs Wcpfn3

(k) Bivariate Plot of Bcpfn2 vs Bcpfn3

(i) Bivariate Plot of Wcpfn1 vs Wcpfn3

(l) Bivariate Plot of Bcpfpn1 vs Bcpfpn3

O ⊖ Signifies two or more hidden observations

Figure 16 Bivariate Plots of the Worst- and Best-Case Proportions of False-Positive and False-Negative Errors

Table 9 contains the results of binomial tests of equality of pairs of worst- and best-case proportions of false-positive and false-negative errors. These results were used to examine (at a .01 level of statistical significance) whether or not the Rating and Matching methods were similar. Note that each number in Table 9 represents the number of tests for which comparison of pairs of worst- or best-case proportions of false-positive or false-negative errors were or were not significantly different. For example, the numbers 3 and 9 in the table mean that the difference in each of three pairs of the worst-case proportions of false-positive errors derived from rating judgments on the "Overall" dimension and rating judgments on all dimensions was statistically significant; the difference in each of nine pairs of the worst-case proportions of false-positive errors was not statistically significant. In other words, for three tests the worst-case proportions of false-positive errors derived from rating judgments on the "Overall" dimension tended to be unequal to the worst-case proportions of false-positive errors derived from rating judgments on all dimensions; for nine tests, the worst-case proportions of false-positive errors tend to be equal. According to the data in Table 9, it is reasonable to infer that matching judgments tended to be more similar to rating judgments on the "Overall" dimension than they were to rating judgments on all dimensions.

		RATING P2				MATCHING P3			
		WcPfp	BcPfp	WcPfn	BcPfn	WcPfp	BcPfp	WcPfn	BcPfn
RATING P1	Sig	3*	5*	0	0	2+	3o	0	0
	Not Sig	9	7	12	12	10	9	12	12
MATCHING P3	Sig	4*	6*	0	0				
	Not Sig	8	6	12	12				

SYMBOL	MEANING
Sig	Significantly different at a .01 level.
*	The WcPfp for the Rating method (P2) was significantly larger in each of these tests.
+	The WcPfp for the Matching method (P3) was significantly larger in one test but smaller in the other.
o	The BcPfp for the Matching method (P3) was significantly larger in one test but smaller in the other two tests.

Table 9 Results of Binomial Tests on Pairs of Worst- and Best-Case Proportions of False-Positive and False-Negative Errors.

For the entire collection of test items, the worst- and best-case numbers and proportions of false-positive and false-negative errors derived from rating and matching judgments are presented in Table 10. At a .01 Type I error level, the results of the normal approximation to the binomial test of equality of proportions led to the following conclusions:

(1) The worst-case proportions of false-positive errors which resulted from matching judgments and rating judgments on the "Overall" dimension tended to be equal.

(2) The worst-case proportion of false-positive errors which resulted from rating judgments on all dimensions was larger than the worst-case proportion of false-positive errors which resulted from rating judgments on the "Overall" dimension or matching judgments.

(3) The best-case proportion of false-positive errors which resulted from rating judgments on all dimensions was larger than the best-case proportion of false-positive errors which resulted from rating judgments on the "Overall" dimension; the best-case proportion of false-positive errors which resulted from rating judgments on all dimensions was larger than the best-case proportion of false-positive errors which resulted from matching judgments.

(4) The worst- and best-case proportions of false-negative errors derived from rating judgments were equal to those derived from matching judgments.

METHOD	Number	WcPfp	No.	BcPfpf	No.	WcPfn	No.	BcPfpn
RATING P1	20	.294	2	.029	3	.019	0	0
RATING P2	61	.897	16	.235	3	.019	0	0
MATCHING	20	.294	1	.015	3	.019	0	0

Table 10 Worst- and Best-Case Number and Proportion of False-Positive and False-Negative Errors for the Entire Collection of Tests

For each test and for the entire collection of tests, the null hypothesis that there is no correlation between rating judgments and matching judgments was tested against the alternative hypothesis that there is a positive correlation between rating judgments and matching judgments. The values of the Cramér measure of association between matching judgments and rating judgments, and their corresponding Chi-square values are shown in Table 11.

Notice that some of the Cramér contingency coefficients in Table 11 were derived by inspection. To illustrate the derivation of the Cramér contingency coefficient by inspection, consider the results tabulated for Domain 4 in Figure 14. Note that the values in the row and column for the "inconclusive" judgments are zeros, thereby propagating zeros as their expected values. As a part of the calculation of the Cramér statistic, the expected value of each cell is subtracted from the observed frequency, the difference is squared and then divided by the expected frequency. Since division by zero is mathematically undefined, the Cramér statistic was derived by inspection whenever it was possible. Note that the results tabulated for Domain 4 in Figure 14 clearly showed a perfect agreement between matching and rating judgments (20 test items matched as valid were also rated as valid, and 2 test items matched as invalid were also rated as invalid). In that case, the Cramér coefficient was derived as 1.0.

Though all values of Cramér measure in Table 11 were statistically significant at the .01 level, note that the average expected values per cell (of a Chi-square random variable) for half of the domains were less than 2.0. When the average expected value per cell in a contingency table is at least 2.0, the approximation of the sampling distribution of the Chi-square has been found to produce reliable results (see Kendall and Yule, 1950). With five percent probability of committing a Type I error, we conclude that:

(1) There tended to be a positive correlation between rating judgments on the "Overall" dimension and matching judgments.

(2) There tended to be a positive correlation between rating judgments on all dimensions and matching judgments.

An examination of Cramér values for the entire collection of items revealed that matching judgments tended to be more correlated with rating judgments on the "Overall" dimension than they did with rating judgments on all dimensions.

DOMAIN		MATCHING vs RATING P1		MATCHING vs RATING P2	
	Avge-Ex	Cramér	Chi-sq	Cramér	Chi-sq
1	4.0	#	#	#	#
2	1.6	0.829	19.3	0.723	14.6
3	1.4	#	#	#	#
4	2.4	1.0(*)	44.0(*)	1.0(*)	44.0(*)
5	1.2	1.0(*)	22.0(*)	1.0(*)	22.0(*)
6	2.0	1.000	36.0	#	#
7	2.9	1.000	52.0	0.721	27.0
8	1.6	#	#	0.669	12.5
9	1.2	1.0(*)	20.0(*)	1.0(*)	22.0
10	3.1	0.950	50.5	#	#
11	2.0	#	#	#	#
12	1.4	#	#	#	#
ALL	24.9	0.835	312.5	0.637	182.00

SYMBOL	MEANING
Avge-Ex	Average expected value per cell
Cramér	Cramér contingency coefficient
Chi-sq	Chi-square value
#	Undefined value

(*) Not calculated mathematically but derived by inspection

Table 11 Correlations Between Rating and Matching Judgments

3.5 Domain Representativeness

The numbers and percents of teachers who expressed judgments on whether or not collections of self-rated or self-matched content valid items adequately covered the scopes of their domains, are tabulated in Appendix C.2. The numbers and proportions of teachers who expressed correct judgments on whether or not collections of self-rated or self-matched content valid items adequately covered the scopes of their domains, are presented in Appendix C.4. The question before us is, does each collection of self-rated or self-matched content valid items adequately cover the scope of its domain?

For each domain, the null hypothesis that no decision can be made on whether or not the collection of content valid items adequately covers the scope of its domain was tested against the alternative hypothesis that the collection of content valid items adequately covers (or does not cover) the scope of its domain. Since a perfect relationship was not found between the Rating and the Matching methods, the set of hypothesis above was tested using rating and matching judgments separately, and also by using combined rating and matching judgments. The results derived from investigating domain representativeness using self-rated and self-matched content valid items are tabulated in Figure 17. To illustrate the interpretation of the values in the figure, consider the results tabulated in (a). According to the numbers in (a) we conclude with five percent risk of committing a Type I error for each domain that:

(1) By making use of self-rated content valid items, seven domains for which the valid items adequately covered the scope of the domains in reality were judged correctly as representative.

(2) Using the collections of self-rated content valid items two domains for which the valid items were representative in reality resulted in "inconclusive" evaluation of the representativeness of the self-rated valid items.

(3) Three domains for which the valid items did not adequately cover the scope of the domains in reality resulted in "inconclusive" judgments on the representativeness of the self-rated valid items.

Of particular interest to content validation of test items are two questions: (1) Is using the self-rated or self-matched valid items to assess domain representativess an accurate technique? and (2) Is using the self-rated valid items to estimate domain representativeness similar to using the self-matched valid items?

To examine the first question above, the null hypothesis that the worst- or best-case population proportions of false-positive errors or false-negative

		REALITY	
		Representative	Not Representative
RATING	Representative	7	0
	Not Representative	0	0
	Inconclusive	2	3

(a)

		REALITY	
		Representative	Not Representative
MATCHING	Representative	8	0
	Not Representative	0	0
	Inconclusive	1	3

(b)

		REALITY	
		Representative	Not Representative
RATING & MATCHING	Representative	9	0
	Not Representative	0	0
	Inconclusive	0	3

(c)

Figure 17 Accuracy of Domain Representativeness Judgments

errors equal zero were tested against the alternative that the worst- or best-case population proportions of false-positive or false-negative errors were greater than zero. Also, the hypothesis that the population correlation between the state of "Reality" and judgments which resulted from the self-rated or self-matched valid items equals zero was tested against the alternative hypothesis that the population correlation was positive. The results of these statistical tests are presented in Table 12. The best-case proportions of false-positive errors were not statistically significant; however, the worst-case proportions of false-positive errors were statistically significant when self-rated or self-matched or the combined self-rated and self-matched valid items were used to establish decisions on domain respresentativeness. The worst-case proportions of false-negative errors were statistically significant when the collection of either self-rated or self-matched valid items were used to validate domain representativeness. Since the intervals containing the range of estimated correlations between the state of "Reality" and judgments which resulted from the self-rated or self-matched items was wide and the worst-case correlation was not significantly different from zero, we conclude that: Using the collections of either self-rated or self-matched valid items to judge domain representativeness tended not to be an accurate technique.

METHOD	WcPfp	BcPfp	WcPfn	BcPfn	WcPhi	WcChi	BcPhi	BcChi
Rating	1.00*	0	.222*	0	0.258	0.80	1.0	12
Matching	1.00*	0	.111*	0	0.174	0.36	1.0	12
Rating & Matching	1.00*	0	0	0	#	#	1.0	12

SYMBOL	MEANING
#	Undefined
*	Statistically significant at the .01 level

Table 12 Accuracy Test Results of Domain Representativeness Derived from Rating and Matching Judgments

Figure 18 contains pertinent data for investigating the second question above. According to the numbers in Figure 18, we conclude that:

(a) Six domains for which self-matched valid items resulted in judgments of "Representative" were also judged as representative when self-rated valid items were used

(b) Two domains for which self-matched valid items resulted in judgments of "Representative" resulted in "inconclusive" judgments when self-rated valid items were used.

(c) One domain for which self-matched valid items resulted in an "inconclusive" judgment was judged to be "Representative" when self-rated valid items were used.

(d) Three domains for which self-matched valid items resulted in "inconclusive" judgments also resulted in "inconclusive" judgments when self-rated valid items were used.

		MATCHING	
		Representative	Inconclusive
RATING	Representative	6	1
	Inconclusive	2	3

Figure 18 2 x 2 Contingency Table for Investigating Similarity Between the Usefulness of Self-rated and Self-matched Valid Items in Assessing Domain Respresentativeness

The observed correlation (Phi) between domain representativeness judgments which resulted from using self-rated valid items and domain representativeness judgments which resulted from using self-matched valid items was 0.48. The corresponding Chi-square test statistic for this correlation was 2.74. Thus, at the .01 error level this correlation was not statistically significant. Since neither the use of self-rated valid items nor the use of self-matched valid items tended to be an accurate technique for eliciting judgments of domain representativeness, the valid test items in "Reality" that were matched or rated correctly as valid (Figures 5, 7 and 9) were used, along with the domain specifications in Appendix A.2 and the information in Appendix C.3, to determine domain representativeness. Specifically, for each domain, we went through the requirements in the specifications and determined whether the collection of items judged correctly as valid adequately covered the scope of the domain. In the process, we paid particular attention to the number of items that must be sampled to measure each tested behavior. For the Rating and the Matching method, it was found that the correct consensus of valid judgments for the test items was useful in the determination of domain representativeness --all nine domains that were representative and three non-representative domains in "Reality" were identified correctly from the correct consensus of valid items.

3.6 Effects of the Proportion of Bad Items on Accuracy of Judgments and Content Validity Indices

The worst-case numbers of false-positive errors across tests for which a Chi-square test of association could be performed are shown for the Rating and Matching methods in Figure 19. The null hypothesis that the worst-case population proportions of false-positive errors and the methods of eliciting judgments were independent was tested against the alternative hypothesis that the worst-case population proportions of false-positive errors and the methods were dependent. The computed value of the Chi-square test statistic for the contingency table in Figure 19 was 5.36, with 6 degrees of freedom. The right-tail probability of a Chi-square random variable whose value is 5.36 with 6 degrees of freedom is between 0.30 and 0.50, and hence the asymptotic approximate p-value is given as p > .30. Thus, the data do support the null hypothesis of independence between worst-case population proportions of false-positive errors and methods of eliciting judgments.

		Mixtures of Bad/Good Items			
		9-91%	23-77%	36-64%	45-55%
	Rating (P1)	1	6	7	6
Method	Rating (P2)	4	11	19	27
	Matching (P3)	1	5	10	4

Figure 19 3 x 4 Contingency Table for Estimating Association Between Methods of Eliciting Judgments and Mixtures of "bad" and "good" Items

The question before us now is, are the worst-case proportions of false-positive errors the same or different? To investigate this question, the Chi-square test for equality of proportions was used to test the null hypothesis that the four worst-case population proportions of false-positive errors were equal against the alternative hypothesis that at least two worst-case population proportions of false-positive errors were not equal. The Chi-square test was also used to perform an equality test for each pair of worst-case proportions of false-positive errors. The results of the Chi-square tests for equality of proportions are presented in Figure 20. From these results we conclude at a .05 level of significance that:

(1) When the test items were rated on the "Overall" dimension or rated on all dimensions or matched to their corresponding domains, the data supported the null hypothesis that the worst-case population proportions of false-positive errors which resulted from the four mixtures of "bad" and "good" items were equal.

(2) When the test items were rated on on all dimensions, the data supported the null hypothesis that the best-case population proportions of false-positive errors which resulted from the four mixtures of "bad" and "good" items were equal.

Note that in Figure 20(d) only the worst-case proportions of false-positive errors which resulted from the 36% bad-64% good and the 45% bad-55% good mixtures of items were significantly different. This difference we attributed to random fluctuation and concluded accordingly that, there was no statistically significant effect of the mixtures of "bad" and "good" items on the worst-case proportions of false-positive errors.

	Mixtures of Bad/Good Items			
	9-91%	23-77%	36-64%	45-55%
Observed Number of Errors	1	6	7	6
Total Number of Items	4	12	20	32

Chi-square = 3.21 with 3 df.

	9-91%	23-77%	36-64%	45-55%
9-91%		0.43	0.10	0.07
23-77%			0.42	3.13
36-64%				1.30

(a) 2 x 4 Contingency Table and Chi-square Values for Testing Equality among Worst-Case Proportions of False-Positive Errors Derived from Rating Judgments on the "Overall" Dimension

Figure 20 Contingency Tables and Chi-square Values for Testing Equality among Proportions of False-Positive Errors Derived from Rating and Matching Judgments

Figure 20 continued--

	Mixtures of Bad/Good Items			
	9-91%	23-77%	36-64%	45-55%
Observed Number of Errors	4	11	19	27
Total Number of Items	4	12	20	32

Chi-square = 0.22 with 3 df.

	9-91%	23-77%	36-64%	45-55%
9-91%		0.02	0.01	0.10
23-77%			0.01	0.05
36-64%				0.16

(b) 2 x 4 Contingency Table and Chi-square Values for Testing Equality among Worst-Case Proportions of False-Positive Errors Derived from Rating Judgments on all Dimensions

Figure 20 Contingency Tables and Chi-square Values for Testing Equality among Proportions of False-Positive Errors Derived from Rating and Matching Judgments

Figure 20 continued--

Mixtures of Bad/Good Items

	9-91%	23-77%	36-64%	45-55%
Observed Number of Errors	0	3	7	6
Total Number of Items	4	12	20	32

Chi-square = 2.38 with 3 df.

	9-91%	23-77%	36-64%	45-55%
9-91%		1.00	1.40	0.75
23-77%			0.24	0.17
36-64%				1.30

(c) 2 x 4 Contingency Table and Chi-square Values for Testing Equality among Best-Case Proportions of False-Positive Errors Derived from Rating Judgments on all Dimensions

Figure 20 Contingency Tables and Chi-square Values for Testing Equality among Proportions of False-Positive Errors Derived from Rating and Matching Judgments

Figure 20 continued--

	Mixtures of Bad/Good Items			
	9-91%	23-77%	36-64%	45-55%
Observed Number of Errors	1	5	10	4
Total Number of Items	4	12	20	32

Chi-square = 6.63 with 3 df.

	9-91%	23-77%	36-64%	45-55%
9-91%		0.22	0.45	0.40
23-77%			0.11	3.63
36-64%				6.43*

* statistically significant at a .05 level

(d) 2 x 4 Contingency Table and Chi-square Values for Testing Equality among Worst-Case Proportions of False-Positive Errors Derived from Matching Judgments

Figure 20 Contingency Tables and Chi-square Values for Testing Equality among Proportions of False-Positive Errors Derived from Rating and Matching Judgments

The nature of the relationship between the four levels of mixtures of "bad" and "good" items and the proportions of false-positive errors is illuminated graphically in Figure 21. From this graph, the following observations were made:

(1) Regardless of the level of mixture of "bad" and "good" test items, the worst-case proportion of false-positive errors derived from rating judgments on all dimensions was larger than the worst-case proportion of false-positive errors derived from either rating judgments on the "Overall" dimension or matching judgments.

(2) When test items were rated on all dimensions, the worst-case proportion of false-positive errors decreased as the percentage of "bad" items in the mixtures of "bad" and "good" items increased from 9% to 23% , increased as the percentage rose to 36%, and decreased as the percentage of "bad" items increased to 45%; that is, the functional relationship tended to be nonlinear. The relationship between the best-case proportions of false-positive errors and proportions of "bad" items also tended to be nonlinear.

(2) When test items were rated on the "Overall" dimension or matched to their corresponding domains, the relationship between the worst-case proportions of false-positive errors and proportions of "bad" items tended to be nonlinear.

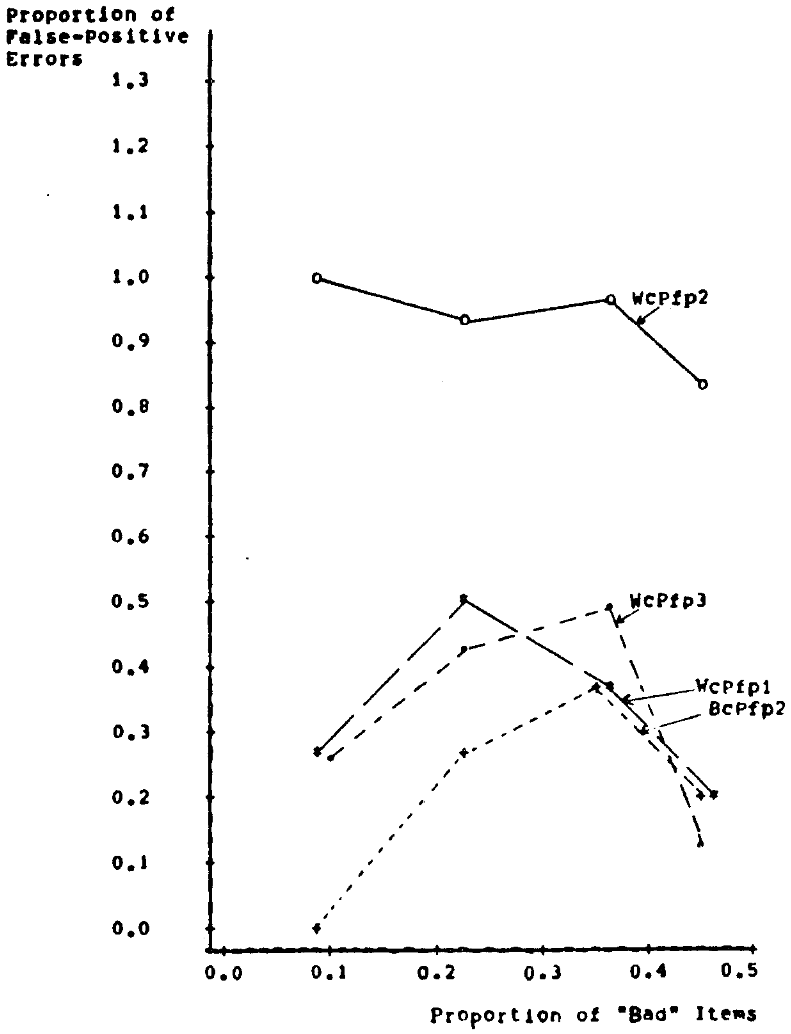


Figure 21 Graph of Proportions of False-Positive Errors versus Proportions of "Bad" Items

The numbers and proportions of "correct," "incorrect" and "inconclusive" decisions across tests are shown shown for P1 (index of content validity of individual items using rating judgments on the "Overall" dimension), P2 (index of content validity of individual items using rating judgments on all dimensions) and P3 (index of content validity of individual items using matching judgments) in Figure 22. For each category of decision which resulted from using each index to determine the content validity of individual test items, the null hypothesis that the four population proportions of decisions were equal against the alternative that at least two population proportions of decisions were not equal. The Chi-square test was also used to perform an equality test for each pair of proportions in each category of decision. The results of the Chi-square tests for equality of proportions are shown in Figure 22. From these results we conclude at a .05 level of significance that:

(1) When the content validity of test items was determined using the index P1 or the index P2 or the index P3, the data supported the null hypothesis that the population proportions of "correct" or "incorrect" decisions which resulted from the four proportions of "bad" items in the mixtures of "good" and "bad" items were equal.

(2) When the content validity of test items was established using the index P1 or P2, the data supported

the null hypothesis that the population proportions of "inconclusive" decisions which resulted from the four proportions of "bad" items in the mixtures of "good" and "bad" items were equal; when the content validity of test items was determined using the index P3, the proportions of the "inconclusive" decisions which resulted from the four proportions of "bad" items in the mixtures of "good" and "bad" items were not equal.

In Figure 22, the Chi-square test statistic values are provided for the pairs of proportions of "incorrect" or "inconclusive" decisions in which there were statistically significant differences. For example, it was found at a .05 level of significance that: When the content validity of test items was determined using the index P2, the data supported the alternative hypothesis that the population proportions of "incorrect" decisions which resulted from the 0.09 and 0.36 proportions of "bad" items were not equal.

The nature of the functional relationship between the decisions ("correct," "incorrect" and "inconclusive") which resulted from using each of the P1, P2 and P3 indices to establish the content validity of test items, and the proportions of "bad" items is illuminated in Figure 23. From the graphs (a), (b) and (c) in Figure 23, the following observations were made:

(1) When the content validity of test items was determined using the index P1 or P2 or P3, the proportion of "correct" decisions decreased as the the proportion of "bad" items increased from 0.09 to 0.23 and from 0.23 to 0.36, and then increased as the proportion of "bad" items increased from 0.36 to 0.45; that is, there tended to be a somewhat inverse linear component in the functional relationship between the proportions of "correct" decisions and the proportions of "bad" items.

(2) When the content validity of test items was established using the index P1 or P3, the proportions of "incorrect" decisions tended to remain constant at zero as the the proportion of "bad" items increased. With the index P2, the proportion of "incorrect" decisions increased as the proportion of "bad" items increased from 0.09 to 0.23 and from 0.23 to 0.36, and then decreased as the proportion of "bad" items increased from 0.36 to 0.45; that is, there tended to be a somewhat direct linear component in the functional relationship between the proportions of "incorrect" decisions and the proportions of "bad" items.

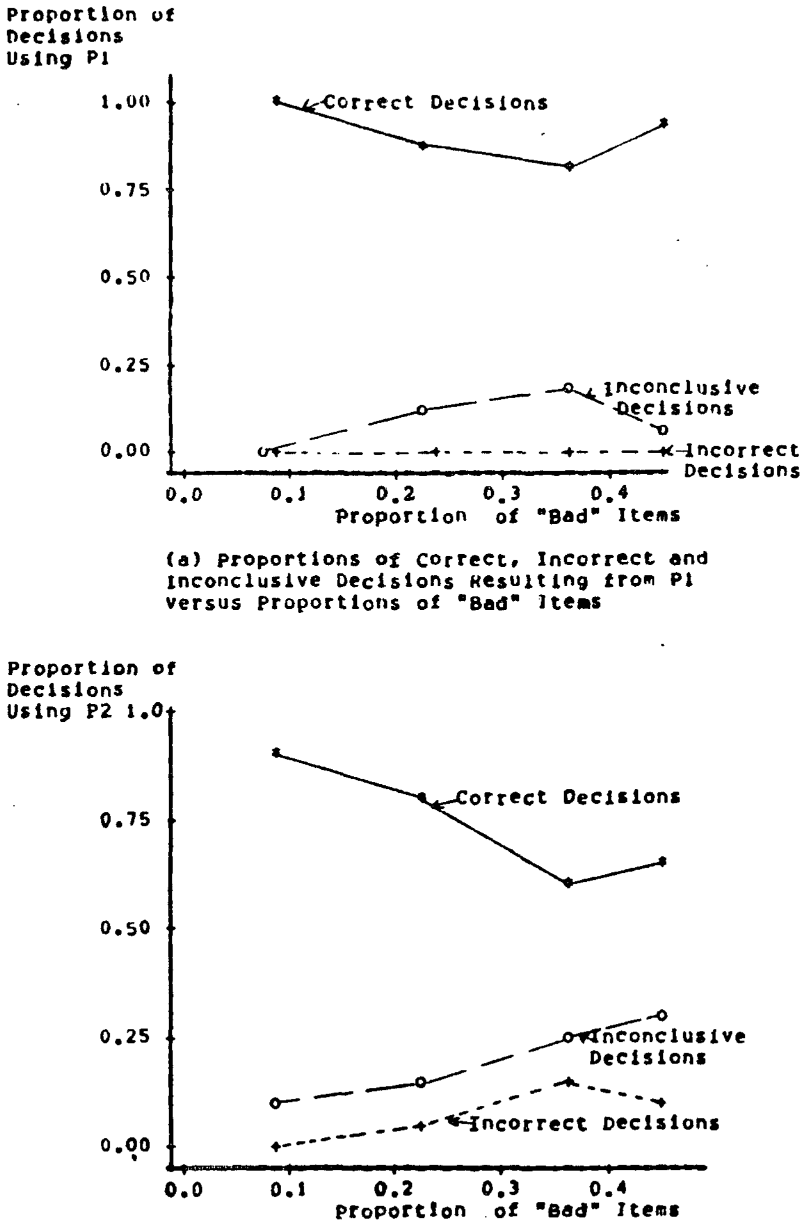
(3) When the content validity of test items was determined using the index P2, the proportions of "inconclusive" decisions increased as the the proportion of "bad" items increased. With the index P1 or P3, the proportion of

"inconclusive" decisions increased as the proportion of "bad" items increased from 0.09 to 0.23 and from 0.23 to 0.36, and then decreased as the proportion of "bad" items increased from 0.36 to 0.45; that is, there tended to be a somewhat direct linear component in the functional relationship between the proportions of "inconclusive" decisions and the proportions of "bad" items.

Decision		Index P1				Index P2				Index P3			
		Proportion of Bad Items				Proportion of Bad Items				Proportion of Bad Items			
		.09	.23	.36	.45	.09	.23	.36	.45	.09	.23	.36	.45
Correct	#	43	46	46	66	40	41	34	45	43	47	43	68
	p	.98	.89	.82	.92	.91	.79	.61	.63	.98	.90	.77	.94
		Chi-square=0.71				Chi-square=4.44				Chi-square=1.54			
Incor-rect	#	0	1	1	0	0	3	7	6	0	0	0	1
	p	.00	.02	.02	.00	.00	.06	.13	.08	.00	.00	.00	.01
						Q=5.5							
		Chi-square=2.15				Chi-square=5.67				Chi-square=2.11			
Incon-clusive	#	1	5	9	6	4	8	15	21	1	5	13	3
	p	.02	.10	.16	.08	.09	.15	.27	.29	.02	.10	.23	.04
		Q=4.69				Q=4.1		Q=5.1		Q=7.7		Q=9	
		Chi-square=5.13				Chi-square=6.77				Chi-square=15.1			

SYMBOL	MEANING
#	Number of Test Items
p	Proportion of Test Items
Q	Chi-Square Test Statistic for Comparing two Population Proportions

Figure 22 Numbers, Proportions and Chi-square Values for Testing Equality among Proportions of Correct, Incorrect and Inconclusive Decisions Derived from P1, P2 and P3

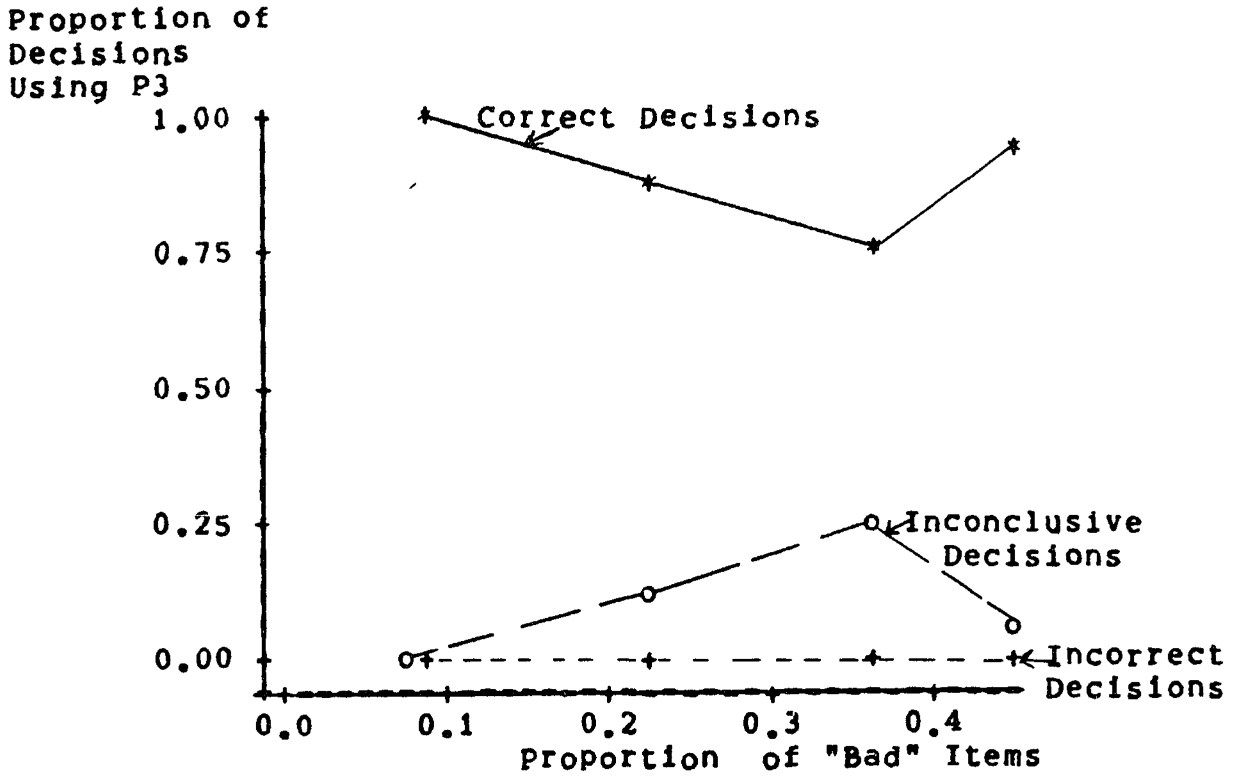


(a) Proportions of Correct, Incorrect and Inconclusive Decisions Resulting from P1 versus Proportions of "Bad" Items

(b) Proportions of Correct, Incorrect and Inconclusive Decisions Resulting from P2 versus Proportions of "Bad" Items

Figure 23 Graph of Proportions of Correct, Incorrect and Inconclusive Decisions Resulting from P1, P2 and P3 versus Proportions of "Bad" Items

Figure 23 continued--



(c) Proportions of Correct, Incorrect and Inconclusive Decisions Resulting from P3 versus Proportions of "Bad" Items

Figure 23 Graph of Proportions of Correct, Incorrect and Inconclusive Decisions Resulting from P1, P2 and P3 versus Proportions of "Bad" Items

3.7 Consistency of the Content Specialists

The results of the degree to which the content specialists were consistent in their judgments over the domains are presented in Figure 24. These reliability estimates reflect the degree to which the content specialists were similar in their judgments. When individual test items were rated on the "Overall" dimension, the interjudge reliability varied from 0.605 to 0.969, over the domains. The index of consistency of rating judgments on all dimensions over the domains, ranged from 0.882 to 0.998. The interjudge reliability varied from 0.652 to 0.940 over the domains, when test items were matched to their corresponding domains. As might be observed from Figure 24, the reliability estimate of rating judgments on all dimensions was consistently larger than the reliability estimate derived from either rating judgments on the "Overall" dimension or matching judgments, regardless of the domain.

To examine the possible effects of the proportions of "bad" test items on the degree of consistency of the judges, consider the graphs (a), (b) and (c) in Figure 25. The points connected in each of these graphs are the averages of the estimated reliabilities at the four levels of mixtures of "bad" and "good" items. Note that these averages are provided in Figure 24. From the graphs (a), (b) and (c), we conclude that:

(1) When test items were rated on all dimensions of judgment, judges tended to be more similar in their judgments as the proportion of "bad" items increased.

(2) When test items were rated on the "Overall" dimension or matched to their corresponding domains, judges tended to be more similar in their judgments as the proportion of "bad" items increased from .09 to .23 and from .23 to .36. The similarity in their judgments became smaller as the proportion of "bad" items increased from .36 to .45.

To examine the nature of similarity between the reliability estimates derived from rating and matching judgments, consider the bivariate reliability graphs (d), (e) and (f) in Figure 25. By inspection, these graphs reveal that:

(1) As the reliability estimates derived from rating judgments on all dimensions tended to increase, the reliability estimates derived from rating judgments on the "Overall" dimension also tended to increase. That is, there tended to be a direct linear relationship between the reliability estimates derived from rating judgments on all dimensions and the reliability estimates calculated from rating judgments on the "Overall" dimension.

(2) There tended to be a somewhat weaker direct linear relationship between the reliability estimates derived from the rating judgments on all dimensions and the reliability estimates derived from matching judgments.

(3) There tended to be a direct linear relationship between the reliability estimates derived from the rating judgments on the "Overall" dimension and the reliability estimates derived from matching judgments.

Domain	Rating (P1)	Rating (P2)	Matching (P3)
1	0.815	0.961	0.885
2	0.904	0.952	0.906
3	0.922	0.953	0.904
4	0.605	0.963	0.727
5	0.969	0.998	0.940
6	0.912	0.960	0.911
7	0.814	0.910	0.878
8	0.881	0.955	0.885
9	0.764	0.822	0.763
10	0.884	0.950	0.883
11	0.864	0.947	0.652
12	0.903	0.958	0.869

(a) Interjudge Reliability Estimates for the Rating and the Matching Methods

Figure 24 Interjudge and Average Interjudge Reliability Estimates for the Rating and the Matching Methods

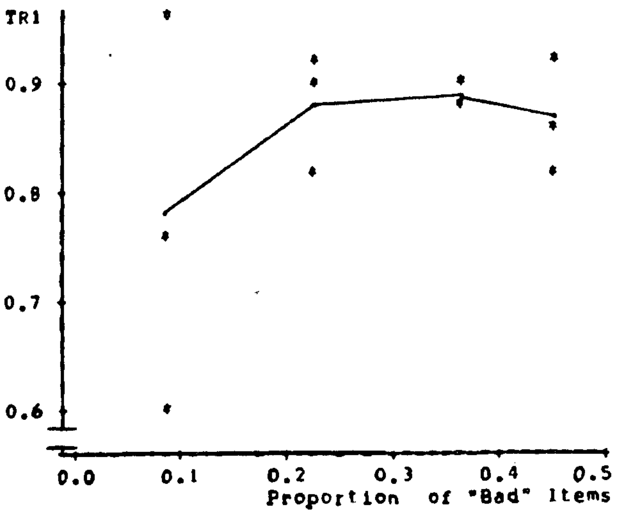
Figure 24 continued--

Mixtures of Bad/Good Items

		9-91%	23-77%	36-64%	45-55%
	Rating (P1)	0.779	0.880	0.890	0.864
Method	Rating (P2)	0.928	0.940	0.952	0.956
	Matching (P3)	0.810	0.884	0.891	0.816

(b) Average Interjudge Reliability Estimates Across Mixtures of "Bad" and "Good" Items, for the Rating and the Matching Methods

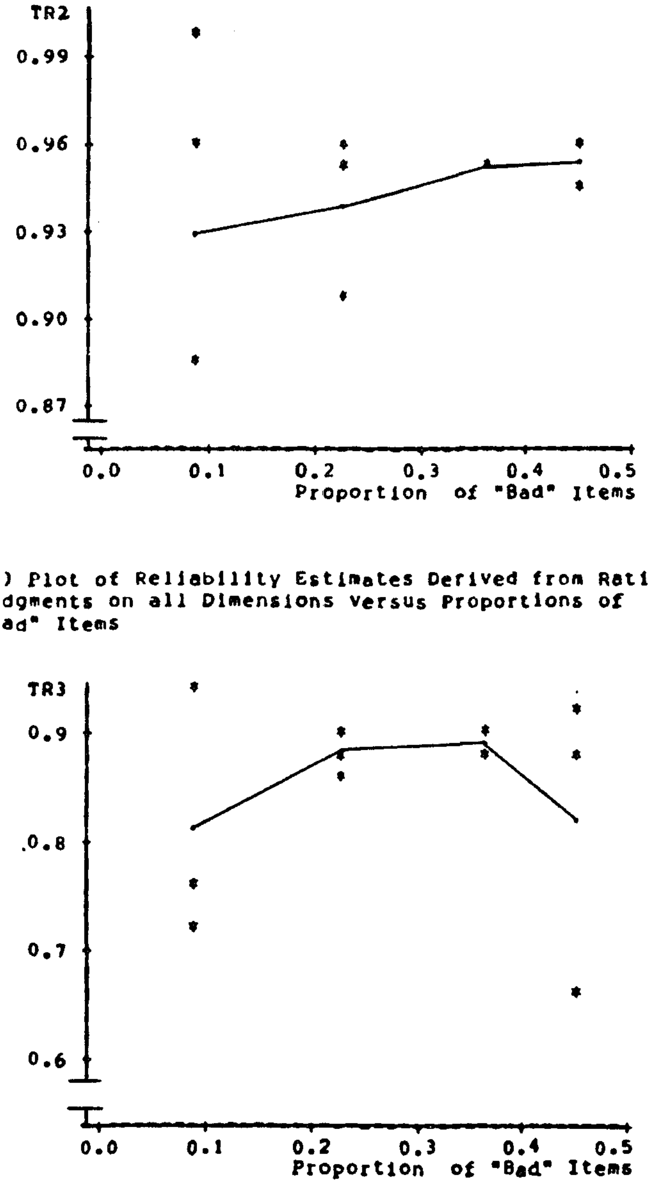
Figure 24 Interjudge and Average Interjudge Reliability Estimates for the Rating and the Matching Methods



(a) Plot of Reliability Estimates Derived from Rating Judgments on the Overall Dimension versus Proportions of "Bad" Items

Figure 25 Plots of Reliability Estimates versus Proportions of "Bad" Items and Bivariate Plots of Reliability Estimates Derived from Rating and Matching Judgments

Figure 25 continued--

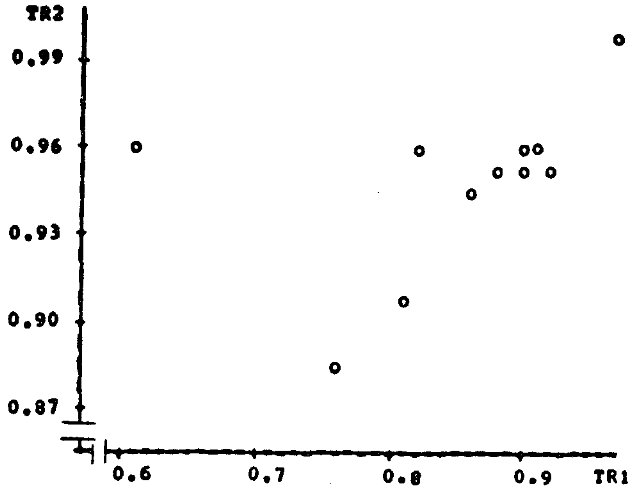


(b) Plot of Reliability Estimates Derived from Rating Judgments on all Dimensions versus Proportions of "Bad" Items

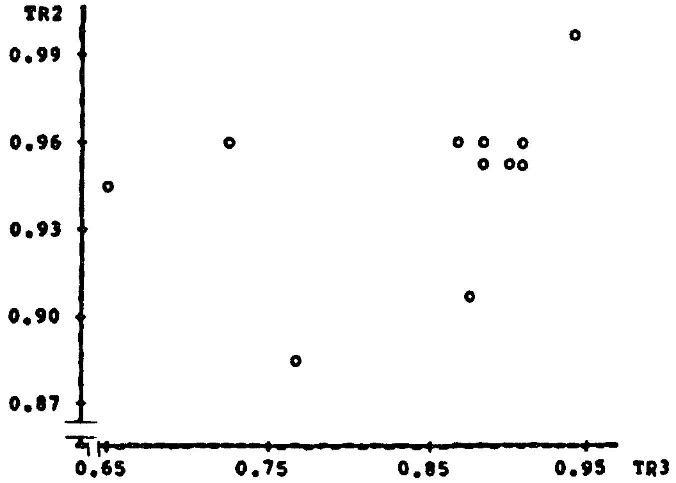
(c) Plot of Reliability Estimates Derived from Matching Judgments versus Proportions of "Bad" Items

Figure 25 Plots of Reliability Estimates versus Proportions of "Bad" Items and Bivariate Plots of Reliability Estimates Derived from Rating and Matching Judgments

Figure 25 continued--



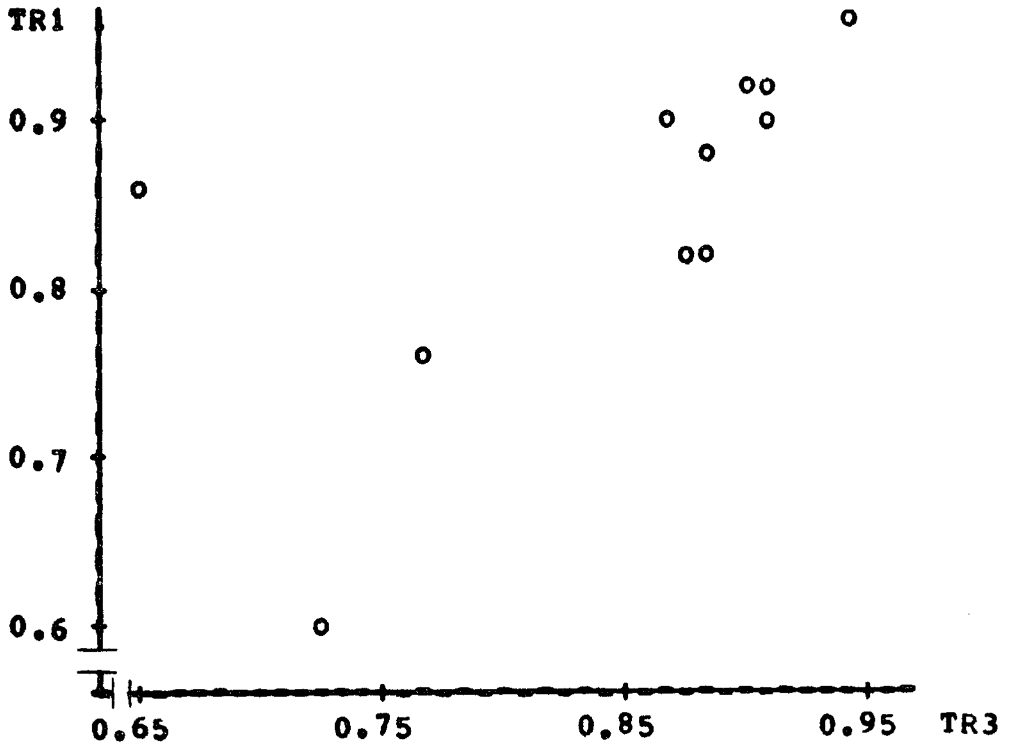
(d) Plot of Reliability Estimates Derived from Rating Judgments on all Dimensions versus Reliability Estimates Derived from Rating Judgments on the Overall Dimension.



(e) Plot of Reliability Estimates Derived from Rating Judgments on all Dimensions versus Reliability Estimates Derived from Matching Judgments.

Figure 25 Plots of Reliability Estimates versus Proportions of "Bad" Items and Bivariate Plots of Reliability Estimates Derived from Rating and Matching Judgments

Figure 25 continued--



(f) Plot of Reliability Estimates Derived from Rating Judgments on the Overall Dimension versus Reliability Estimates Derived from Matching Judgments.

Figure 25 Plots of Reliability Estimates versus Proportions of "Bad" Items and Bivariate Plots of Reliability Estimates Derived from Rating and Matching Judgments

In Chapter 4, we examine the implications of the results of various applications of content validity indices that have been discussed in this chapter.

CHAPTER 4

CONCLUSIONS

The results of content validation evidently depend to some extent on the procedures used for eliciting judgments. When the Rating method or the Matching method is used to elicit judgments of the content validity of mathematics achievement test items, it is unlikely that valid test items will be judged invalid. However, the number of invalid test items that are judged as valid is likely to be greater when the items are rated individually on "Format," "Numbers," "Wording" and "Behavior" dimensions than when the items are either matched to their corresponding domains or rated on the basis of overall quality. The current research showed that the correlations between the state of "Reality" and rating judgments on the "Overall" dimension or matching judgments were larger than the correlation between the state of "Reality" and rating judgments derived from all dimensions. Moreover, the results of a comparison between the Rating and the Matching methods indicated that matching judgments were more closely related to summative rating judgments on the "Overall" dimension than they were to rating judgments on the "Format," "Numbers," "Wording" and "Behavior" dimensions.

The finding that the Matching method tends to be more accurate than does the Rating method using all dimensions of judgment was somewhat disappointing. Given that judgments must be accurate on each of the "Format," "Numbers," "Wording" and "Behavior" dimensions for judgment of an item to be correct, one might expect judges to be more prone to commit errors with the Rating method using all dimensions than they would with the Matching method or the summative Rating method just using an "Overall" dimension. However, since judges were content specialists in the area of elementary school mathematics and they made judgments on the most noticeable characteristics of test items, it might be reasonable to expect greater accuracy from rating judgments on all dimensions than from matching judgments.

The finding that summative rating judgments were more accurate than were rating judgments on all dimensions was not surprising, in view of the fact that the summative rating method considers all dimensions of rating in addition to external criteria other than the given dimensions of judgment. Since the the rating method using the "Overall" dimension and matching method both require summative types of judgments, the similarity between the two methods was not surprising.

Although the matching and summative rating methods were more accurate than was the rating method using all dimensions, with the latter method judges were more consistent in their judgments of the content validity of mathematics test items over the domains.

The results of the present research revealed that using judges to assess domain representativeness was not an accurate method. However, by manually examining the number of sample items defined for each class of assessed behavior in each domain specification and the collection of test items that were matched or rated correctly as measures of the domain, domain representativeness was successfully established. Thus, if the number of items to be constructed for measuring each type of knowledge defined for a domain is clearly specified, domain representativeness could be determined from the results of content validation of individual test items.

Although the matching method and rating method using the "Overall" dimension were found to be accurate procedures for eliciting judgments, the procedures themselves do not provide information for rectifying the defects in invalid test items. Thus, the rating method using all dimensions is more useful if one is interested not only in identifying invalid test items, but also in identifying the specific details of what may be wrong with each item. The present research results showed that rating judgments on the "Format" and "Numbers" dimensions were more accurate than were rating judgments on the "Wording" and "Behavior" dimensions.

The proportion of false-negative errors which resulted from rating and matching judgments remained constant at zero as the proportion of "bad" items provided to judges increased. The worst-case proportions of false-positive errors derived from rating judgments using the "Overall" dimension and those derived from matching judgments were not significantly different. The research results indicated that the worst-case proportion of false-positive errors derived from rating judgments using all dimensions was significantly larger than the worst-case proportion of false-positive errors derived from either rating judgments using the "Overall" dimension or matching judgments.

As evidenced from analyses of the worst- and best-case proportions of false-positive errors, when test items were rated on all dimensions, the accuracy of judges varied with the proportion of "bad/good" items presented to them. At the worst case, the accuracy of judges increased as the proportion of "bad" items increased from .09 to .23, the accuracy decreased as the proportion of "bad" items increased from .23 to .36 and then increased as the proportion of "bad" items increased from .36 to .45; at the best case, the accuracy of judges decreased as the proportion of "bad" items increased from .09 to .23 and from .23 to .36, and then increased as the proportion of "bad" items increased from .36 to .45.

When test items were matched to their corresponding domains or rated using the "Overall" dimension, the accuracy of judges decreased as the proportion of "bad" items increased to a certain level and then the accuracy increased. The results derived from matching judgments indicated that the worst-case proportion of false-positive errors increased as the proportion of "bad" items increased from .09 to .23 and from .23 to .36 and then decreased as the proportion of "bad" items increased from .36 to .45. The accuracy results derived from rating judgments using the "Overall" dimension showed that the worst-case proportion of false-positive errors increased as the proportion of "bad" items increased from .09 to .23 and then decreased as the proportion of "bad" items increased from .23 to .36 and from .36 to .45.

The three test item validity indices are to some extent "good." The accuracy of the three test item content validity indices established using rating and matching judgments tended to decrease as the proportion of "bad" items increased from from .09 to .23 and from .23 to increased from .36 to .45. Ideally, a "good" index should decrease as the proportion of "bad" items increases.

The reliability estimates derived from rating judgments on all dimensions tended to increase as the proportion of "bad" items increased. This means that judges were more similar in their judgments over the domains as the proportion of "bad" items provided to them increased. The research results showed that when test items were matched to their corresponding domains or rated on the basis of "Overall" quality, judges were more similar in their judgments as the proportion of "bad" items increased to a certain level (0.36) and then were less similar beyond that level.

When content specialists make judgments of the content validity of test items, it is doubtful that the most desirable sample size will be available. This raises questions about the probability of rejecting a false null hypothesis; that is, the power of a statistical test in the content validity study. In the present research, we realized that the power of the test for each null hypothesis dealing with an individual test item would have increased had more content specialists been available; however, to compensate for the small sample size, we increased the probability of rejecting a true null hypothesis (α) from .01 to .05, to increase the likelihood of achieving reasonable power for each hypothesis tested.

Although content validity evidence is insufficient for validation of the many different interpretations of criterion-referenced test scores, a good criterion-referenced test must be content valid. As derived from the current research results, the following guidelines are recommended for validating the content of criterion-referenced mathematics achievement tests:

(1) Prepare and validate domain specifications. The intended content and behaviors to be measured by a criterion-referenced test must be stated clearly in the domain specifications. Because they provide information on item format, size of numbers, boundaries on words for stating a word problem, and number of sampled items for each behavior elicited, domain specifications can be referred to not only in the process of content validation of mathematics achievement tests but also in the process of correcting invalid test items. If domain specifications themselves are invalid, incorrect judgments of the content validity of tests will always result. In order to avoid this problem, a number of qualified content domain specialists must reach consensus on the validity of domain specifications.

(2) Provide training for judges. When classroom teachers are used as expert judges, they probably will not have been exposed to the concepts, terms and issues surrounding validity in general, and content validity in particular. If this is the case, the importance of training judges on methods of eliciting judgments and on content validation procedures cannot be overemphasized.

(3) Assess the content validity of all test items. Evidence of the content validity of all criterion-referenced tests must be amassed, if the aims of criterion-referenced testing programs are to be achieved.

One possible approach to content validation would be to use the rating method alone. Unfortunately, the rating method is very time consuming in addition to being less accurate than the matching method. A second approach would be to use the matching method alone to validate individual test items. Although this approach is more accurate and less time consuming than the rating method, the matching method will not provide information for rectifying invalid test items. A third approach would be to validate all test items using a two-stage process in which test items are initially validated using the matching method, and then all invalid test items are revalidated using the rating method. If a significant proportion of invalid test items exists in a large collection of test items, this approach would be burdensome, particularly if the same judges were used in the two-stage process. A possible solution to this problem would be to use two different groups of judges, one group to validate the entire collection of test items using the matching method, and the other group to use the rating method in evaluating the test items which were judged invalid by the first group using the matching method. This solution might produce confounded results due to group differences. Although attrition would be a threat if the same group of judges was to be used in the two-stage process, we recommend that the same group of judges be used to validate the test items, with the stages separated by a period of time --e.g., the judges could revalidate the test items they initially judged to be invalid, one or two weeks after the initial validity study.

(4) Validate domain representativeness. Unless a collection of content-valid test items covers the scope of the domain the test is designed to measure, the test as a whole is invalid in content. According to the present research results, using judges to

assess domain representativeness is not an accurate technique. However, by making use of a collection of self-matched or self-rated content-valid items and the number of sample items defined for each class of assessed behavior in each domain specification, domain representativeness was successfully determined. This manual process is cumbersome. Although it was not investigated in this research, it might be the case that, if test item writers are asked to specify the category of behavior that each item measures, a computer could be programmed to determine domain representativeness using the collection of valid test items.

In the present research we developed two alternative methods of eliciting judgments about the content validity of mathematics test items. The usefulness of the methods and several newly-developed content validity indices has been demonstrated. The research work has shed some light on the theoretical and practical nature of quantitative determination of the content validity of mathematics achievement test items. Statistical determination of content validity in content areas such as English is worth studying. Perhaps there are other dimensions of judgments or even rating scales that may be of interest in other areas of education; this subject is worth investigating.

REFERENCES

Aiken L. R. (1980). Content validity and reliability of single items or questionnaires. Educational and Psychological Measurement, 40: 955-959.

American Psychological Association (1954). Technical recommendations for psychological tests and diagnostic techniques. Psychological Bulletin Supplement, 51, 2: 1-38.

Conover, W. J. (1980). Practical Nonparametric Statistics. New York: John Wiley and Sons, 180-181.

Cramér, H. (1946). Mathematical Methods of Statistics. New Jersey: Princeton Press.

Crocker, L. M., and Algina, J. (1986). Introduction to Classical and Modern Test Theory. New York: CBS College Publishing.

Cronbach, L. J. (1971). Test validity. In R. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D. C.: American Council on Education, (1971), 443-507.

Distefano, M. K., Pryer M. W., and Craig S. H. (1980). Job relatedness of a posttraining job knowledge criterion used to assess validity and test fairness. Personnel Psychology, 33: 785-793.

Distefano, M. K., Pryer M. W., and Erffmeyer R. C. (1983). Application of content validity methods to the development of a job-related performance rating criterion. Personnel Psychology, 36: 621-631.

Gagné, R. M. (1974). Task analysis - its relation to content analysis. Educational Psychologist, 11: 11-18.

Gibbons, J. D. (1976). Nonparametric Methods for Quantitative Analysis. New York: Holt, Rinehart and Winston, 330-333.

Glass, G. V. and Hopkins, K. D. (1984). Statistical Methods in Education and Psychology. Englewood Cliffs, N.J: Prentice Hall.

Glaser, R., and Klaus, D. J. (1962). Proficiency measurement: Assessing human performance. In R. M. Gagne (Ed.), Psychological Principles in Systems Development New York: Holt, Rinehart and Winston. Pp. 419-74.

Glaser, R., and Nitko, A. J. (1971). Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D. C.: American Council on Education.

Greensboro Public Schools (GPS (1982). Revised Administrative Regulation IHE-R. Greensboro Public Schools publication.

Guion, R. M. (1977). Content validity: The source of my discontent. Applied Psychological Measurement 1: 1-10.

Hambleton, R. K., Swaminathan, H., Algina, J., and Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research 48: 1-47.

Hambleton, R. K. (1980). Test score validity and standard setting methods. In R. A. Berk (Ed.), Criterion-Referenced Measurement: The State of the Art. Baltimore: Johns Hopkins University Press.

Hively, W., Patterson, H. L., and Page, S. A. (1968). A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement 5: 275-90.

Jones, D. H., and Szatrowski, T. H. (1983). On the statistical determination of content validity. Educational and Psychological Measurement 43: 995-04.

Kane M. T. (1982). A sampling model for validity. Applied Psychological Measurement, 2, 6: 125-160.

Kendall, N. G., and Yule, G. U. (1950). An Introduction to Theory of Statistics. New York: Haffner, p. 53.

Lawshe, C. H. (1975). A quantitative approach to content validity. Personnel Psychology 28: 563-75.

Linn, R. L. (1979). Issues of validity in measurement for competency-based programs. In M. A. Bunda and J. R. Sanders (Ed.), Practices and Problems in Competency-Based Measurement. Washington, D. C.: National Council on Measurement in Education. Pp 108-23.

Linn, R. L. (1980). Issues of validity for criterion-referenced measures. Applied Psychological Measurement, 4: 547-561.

Loevinger, J. (1965). Person and population as psychometric concepts. Psychological Review, 72: 143-155.

Mayo, S. T. (1970). Mastery learning and mastery testing. Measurement in Education 1: 1-4.

Messick, S. A. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist 30: 955-66.

Millman, J. (1974). Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in Education: Current Applications. Berkeley, Calif.: McCutchan. Pp. 311-97.

Nichols, E. D. et. al. (1981). Holt Mathematics, Student's Edition. New York: Holt, Rinehart and Winston, Publishers.

Nichols, E. D. et. al. (1981). Holt Mathematics, Teacher's Edition. New York: Holt, Rinehart and Winston, Publishers.

Popham, W. J. (1974). Selecting objectives and generating test items for objectives-based tests. In C. W. Harris, M. C. Alkin, and W. J. Popham (Ed.), Problems in Criterion-Referenced Measurement. CSE Monograph Series in Evaluation, no. 3. Los Angeles.

Popham, W. J. (1975). Educational Evaluation. Englewood Cliffs, N.J.: Prentice-Hall.

Popham, W. J. (1978). Criterion-Referenced Measurement. Englewood Cliffs, N.J.:Prentice-Hall.

Popham, W. J. (1980). Domain specification strategies. In R. A. Berk (Ed.), Criterion-Referenced Measurement. The John Hopkins University Press. Pp 15-31.

Rogers T. B. (1973). Ratings of content as means of assessing personality items. Educational and Psychological Measurement, 53: 845-858.

Rovinelli, R. J., and Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. Dutch Journal of Educational Research 2: 49-60.

APPENDIX A.1

Greensboro Public Schools' Mathematics Promotion Standards for Grade 4

OBJECTIVE 4.1
Solves one-step word problems.

OBJECTIVE 4.2
Names and writes numbers to 1,000,000.

OBJECTIVE 4.3
States the place value of each digit in any six-digit number.

OBJECTIVE 4.4
Adds numbers up to four digits with regrouping.

OBJECTIVE 4.5
Writes words for numbers 1 to 100.

OBJECTIVE 4.6
Writes the value of a collection of coins and bills.

OBJECTIVE 4.7
Subtracts numbers up to four digits with regrouping.

OBJECTIVE 4.8
Expresses a measurement in centimeters, meters, and kilometers.

OBJECTIVE 4.9
Tells and writes time to the minute.

OBJECTIVE 4.10
Expresses a measurement in inches, feet, yards, and miles.

OBJECTIVE 4.11
Identifies the numerator and denominator of a fraction.

OBJECTIVE 4.12
Identifies points, lines, and line segments.

APPENDIX A.2

Grade 4 Domain Specifications

DOMAIN 1

Adding 2, 3, or 4 numbers, each with 4 or fewer digits, with or without regrouping.

CONDITIONS FOR ITEM CONSTRUCTION AND TESTING

(1) The student is given 2, 3, or 4 addends, each with 4 or fewer digits, and is asked to provide the sum. The addition may require no regrouping or regrouping from ones to tens, from tens to hundreds, from hundreds to thousands, or any combination of such regrouping.

Number of Sample Items:

(1) No more than two items may test addition operations requiring no regrouping.

(2) No more than eight items may test addition operations requiring single regrouping; however, each of 2, 3 and 4-digit addends must be represented at least once by items requiring single regrouping.

(3) At least ten items must test addition operations requiring multiple regrouping; at least 3 items containing three 4-digit addends must be represented.

Format

Vertical format with plus sign written: The addends may or may not be written in increasing or decreasing order of their magnitudes. The addends must be written vertically according to their place values. A plus sign must be written in front of the last addend, and must be vertically aligned with a blank space to the left of the most significant digit of the largest addend. The last addend and the plus sign must be underlined.

BOUNDARY FOR ITEM CONSTRUCTION

(1) The number of addends for an addition operation must be four or less.

EXAMPLE:- (1.)

$$\begin{array}{r} 357 \\ \underline{+\ 46} \end{array}$$

DOMAIN 2

Identifying the best customary units of linear measurement for objects and distances, and changing units among inches, feet, yards, and miles.

CONDITIONS FOR ITEM CONSTRUCTION AND TESTING

(1) The student is given the name an object and is asked to provide the best customary unit from among inches, feet, yards, or mile, for measuring the length, or width, or height of the object.

(2) The student is given a measurement in one of the customary units (in., ft., yd., or mi) and is asked to convert it into another customary unit. For examples, in. to yd. and yd. to ft.

Number of Sample Items:

(1) Five items must test identification of customary units for measuring objects or distances. The knowledge of each unit of measurement (in., ft., yd., mi.) must be tested at least once.

(2) Five items must test common conversions of measurements among the customary units (in., ft., yd., or mi.). The knowledge of conversions from inches to feet, from inches to yards, from feet to yards, from feet to miles, and from yards to miles, or vice versa, must each be represented.

Format

Free-response: On a separate line, each object's measurement noun (length, or width, or height) must be written, followed by "of a " followed by the name of the object, followed by the space for the student's response.

Fill-in-the-blank: For each conversion item, on a separate line, the value to be converted must be written, followed by its unit, " =______" and the unit for which customary conversion is required.

BOUNDARIES FOR ITEM CONSTRUCTION

(1) The object in a test item must be familiar to grade 4 students.

(2) Only a whole number value can be used to state a customary conversion test item. The number to be converted must be less than 10.

EXAMPLES:- (1.) Width of a table ________________
(2.) 6 mi. = _______ ft.

DOMAIN 3

Identifying and naming points, lines, and line segments.

CONDITIONS FOR ITEM CONSTRUCTION AND TESTING

(1) The student is given a drawing of a curve, or a straight-line passing through or connecting two points, and is asked to state whether the drawing is, or is not a line segment.

(2) The student is given the picture of a point, or a line, or a line segment with appropriate points identified with letters, and is asked to name the object in the picture.

Number of Sample Items:

(1) Five items must test knowledge of line segments; at least one and at most two of the pictures given to the student must be line segments.

(2) Five items must test ability to name points, lines, and line segments; a point, a line, and a line segment must each be represented at least once.

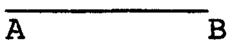
Format:

For each item, the picture of a point, or a line, or a line segment is drawn and the space for the student's response is provided either underneath or to the right-hand side of the drawing.

BOUNDARIES FOR ITEM CONSTRUCTION

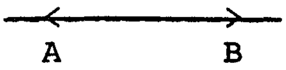
(1) Each line segment must have an arrow at each end. Each line or line segment or curve must contain two different labelled points. A single point must be labelled.

EXAMPLES:- (1.)



____

(2.)



____

DOMAIN 4

Stating the place and value of an underlined digit in a standard numeral of 6 or less digits.

CONDITIONS FOR ITEM CONSTRUCTION AND TESTING

(1) The student is given a number with 6 or less digits and with one of the digits underlined, and is asked to provide the word name for the underlined digit.

(2) The student is given a number with 6 or less digits and with one of the digits underlined, and is asked to provide the value for the underlined digit.

Number of Sample Items:

(a) Ten items must test the knowledge of word names for the places of underlined digits; each place must be represented at least once.
(b) Ten items must test the knowledge of the values for underlined digits; each place must be represented at least once.

Format

Free-response: Each standard numeral and the space for the student to provide the word names or the value of an underlined digit, must occupy a separate horizontal line. The numeral of an item must come before the space for the student's response. Only one digit of a numeral must be underlined.

BOUNDARY FOR ITEM CONSTRUCTION

(1) Each standard numeral must represent a whole number from 1 to 999,999.

EXAMPLE:- (1.) $\underline{3}45$ ______________________________

DOMAIN 5

Identifying the numerators and denominators of fractions.

CONDITIONS FOR ITEM CONSTRUCTION AND TESTING

(1) The student is given a fraction and is asked to name the numerator.

(2) The student is given a fraction and is asked to name the denominator.

Number of Sample Items:

(1) Five items must test identification of numerators; only one item must contain a numerator larger than the denominator.

(2) Five items must test identification of denominators; only one item must contain a numerator larger than the denominator.

Format:

The denominator must be written directly underneath the numerator; with a small line in between the two numbers. Space must be provided for the student's response on the same horizontal line containing the denominator.

BOUNDARIES FOR ITEM CONSTRUCTION

(1) The numerator and the denominator of a fraction must be whole numbers from 1 through 9. The numerator may be larger or smaller than the denominator.

EXAMPLE:- (1.) $\frac{9}{5}$ ____________

DOMAIN 6

Writing dollars and cents value of a collection of coins and bills.

CONDITIONS FOR ITEM CONSTRUCTION AND TESTING

(1) The student is given the word name for a collection of coins, and is asked to write the corresponding dollar value using the '$' sign.

(2) The student is given the word name for a collection of bills, and is asked to write the corresponding dollar value using the '$' sign.

(3) The student is given the word name for a collection of coins and bills, and is asked to write its dollar value using the '$' sign.

Number of Sample Items:

(a) No more than two items may test knowledge of word names for collection of coins.
(b) No more than two items may test knowledge of word names for collection of bills.
(c) At least six items must test knowledge of word names for collection of coins and bills.

Format:

For each item, the word names for the collection of coins, or bills, or coins and bills, and the space for the student's must occupy a single horizontal line. The word names must be written before the space provided for the student's response. In the case where a collection of coins and bills are given to the student, the word name for the number of bills must go before the word name for the number of coins.

BOUNDARIES FOR ITEM CONSTRUCTION

(1) The value of the collection of coins must be less than one dollar.
(2) The value of the collection of bills or coins and bills must be less than one hundred dollars.

EXAMPLE:- (1.) Ninety-nine dollars and one cent ___________

DOMAIN 7

Subtracting a standard numeral from a minuend with or without regrouping.

CONDITIONS FOR ITEM CONSTRUCTION AND TESTING

(1) The student is given a 2- to 4-digit minuend and a subtrahend up to 4 digits, and is asked to provide the difference. The subtraction may require no regrouping or regrouping from thousands to hundreds, from hundreds to tens, from tens to ones, or any combination of such regrouping.

Number of Sample Items:

(1) No more than four items may test subtraction operations requiring no regrouping; each of 2, 3, and 4-digit minuends may be represented.

(2) No more than eight items may test subtraction operations requiring single regrouping; however, each of 3, and 4-digit minuends and subtrahends must be represented at least once by items requiring single regrouping.

(3) At least eight items must test subtraction operations requiring multiple regrouping; each of 3, and 4-digit minuends and subtrahends must be represented.

Format

Vertical, with minus sign written: The subtrahend must be written underneath the minuend, in the order of the place values of the digits of the minuend. A minus sign must be written in front of the subtrahend, and must be vertically aligned with a blank space to the left of the most significant digit of the minuend. The minus sign and the subtrahend must be underlined.

BOUNDARY FOR ITEM CONSTRUCTION

(1) The result from a subtraction operation must be positive or zero. That is, the minuend must be greater than or equal to the subtrahend.

EXAMPLES:- (1) $\begin{array}{r} 245 \\ \underline{-\ 98} \end{array}$ (2) $\begin{array}{r} 49 \\ \underline{-\ 3} \end{array}$

DOMAIN 8

Identifying the best metric unit for measuring a given linear attribute of an object and distances, and changing units between centimeters and meters, and between meters and kilometers.

CONDITIONS FOR ITEM CONSTRUCTION AND TESTING

(1) The student is given the name of an object and is asked to provide the best metric unit (cm, m, or km) for measuring the length, or height, or width of the object.

(2) The student is given the names of two locations and is asked to provide the best metric unit for measuring the distance between the two locations.

(3) The student is given a measurement in centimeters and is asked to convert it into meters, and vice versa.

(4) The student is given a measurement in meters and is asked to convert it into kilometers, and vice versa.

Number of Sample Items:

(1) Five items must test identification of metric units for measuring objects, or distances. The knowledge of each unit of measurement (cm, m, km) must be tested at least once.

(2) Five items must test metric conversions of measurements in centimeters to measurements in meters, or conversions of measurements in meters to measurements in kilometers, and vice versa. The knowledge of each metric conversion must be tested at least once.

Format

Open-ended: On a separate line, each object's measurement noun (length, or width, or height) must be written, followed by "of a " followed by the name of the object and the space for the student's response.

Open-ended: On a separate line, each location test item must be written starting with "distance between " followed by the name of a location, "and" the name of another location, followed by the space for the student's answer.

Format

Fill-in the blank: For each conversion item, on a separate line, the value to be converted must be written, followed by its unit, " =______" and the unit for which metric conversion is required.

BOUNDARIES FOR ITEM CONSTRUCTION

(1) The object or locations of an item must be familiar to Grade 4 students.

(2) The value for a conversion problem must be a whole positive number and must be no larger than 10 kilometers.

EXAMPLES:- (1.) Length of a book ____________

(2.) 9m = ________ cm

DOMAIN 9

Telling the time shown on a round analog clock to the hour, half-hour, quarter-hour, and minute.

CONDITIONS FOR ITEM CONSTRUCTION AND TESTING

(1) The student is given a time to the hour shown on a round anaolg clock and is asked to write the time.

(2) The student is given a time to the half-hour shown on a round analog clock and is asked to write the time.

(3) The student is given a time to the quarter-hour shown on a round analog clock and is asked to write the time.

(4) The student is given a time to the minute shown on a round analog clock and is asked to write the time.

Number of Sample Items:

(1) No more than one item may test knowledge of time to the hour.

(2) No more than two items may test knowledge of time to the half-hour.

(3) No more than three items may test knowledge of time to the quarter-hour.

(4) At least four items must test knowledge of time to the minute;

Format:

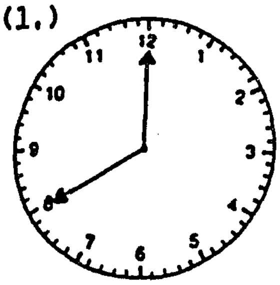
The clock for each item must be drawn with the hour and the minute hands. Underneath the clock, space must be provided for the student to write the time shown on the clock.

BOUNDARY FOR ITEM CONSTRUCTION

(1) The hour and the minute hands shown on a clock must correspond exactly to the time in a question.

EXAMPLE:- Write the time to the minute.

(1.)



____

DOMAIN 10

Matching word names with standard numerals and writing standard numerals for word names to one million.

CONDITIONS FOR ITEM CONSTRUCTION AND TESTING

(1) The student is given a randomly arranged column of numerals and a randomly arranged column of matching word names for standard numerals, and is asked to match the word names with the numerals.

(2) The student is given a word name for a standard numeral and is asked to provide the standard numeral.

Number of Sample Items:

(a) Ten test items must consist of a collection of standard numerals for which the student is to provide corresponding word names. Each of 2 to 6-digit numerals must be represented.

(b) Ten test items must be word names for which student is asked to provide numerals. Each of 2 to 7-digit numeral must be represented.

Format

Matching: A single horizontal line must contain a word name, a standard numeral that does not match the word name, and the space for the student's response. The space for the student's answer must be presented first, followed by the word name, followed by the standard numeral. The columns of spaces for a student's answers, the word names, and the standard numerals must be vertically aligned to the left.

Open-ended response: The word names and the space for the student to write the standard numeral, must be on the same line. For longer word names, the second line of the word name must contain the space for the student's answer.

BOUNDARY FOR ITEM CONSTRUCTION

(1) Each standard numeral must be no larger than 1,000,000

EXAMPLES:- _____ 1. Fourteen a. 73

_____ 2. Seventy-three b. 14

(3.) Two hundred thousand, forty-five (3.) _____

DOMAIN 11

Writing word names for whole numbers from one through one hundred.

CONDITION FOR ITEM CONSTRUCTION AND TESTING

(1) The student is given a natural number and is asked to provide the word name.

Number of Sample Items:

(a) No more than one item must test word name for a single-digit number.

(b) At least eight items must test word names for 2-digit numbers.

(c) One item must test the word name for 100.

Format

Free-response: Each whole number and the space for the student to provide the word name must occupy a single horizontal line. The whole number must be written before the space provided for the student's response. Each problem must be written on a separate line.

BOUNDARIES FOR ITEM CONSTRUCTION

Each whole number must be from 1 through 100.

EXAMPLE:- (1.) 83 ____

DOMAIN 12

Choosing the correct operation (addition or subtraction) to solve one-step word problems and solving the problems.

CONDITIONS FOR ITEM CONSTRUCTION AND TESTING

(1) The student is given a single-step, addition word problem and is asked to solve the problem.

(2) The student is given is a single-step, subtraction word problem, and asked to solve the problem.

Number of Sample Items:

(a) Five items must test knowledge of solving one-step, addition word problems.
(b) Five items must test knowledge of solving one-step, subtraction word problems.
(c) No more than two word problems may be stated by using dollars and cents to express facts.

Format

Open-ended free-response: The word facts for each one-step word problem can be presented on a single horizontal line, or span two or more lines. The statement of each word fact can also occupy a single horizontal line. A question can start on a new line or start immediately after the second word fact, and can extend to the next line. Lines for word facts and question lines when they begin on a new line, must be vertically aligned to the left.

BOUNDARIES FOR ITEM CONSTRUCTION

(1) The words used for defining an addition or subtraction word problem must be within the recognition vocabulary of grade 4 students.
(2) A word problem must make use of whole numbers, or dollars and cents, to express the facts.
(3) The magnitudes of the whole numbers associated with the word facts must be limited to 3-digits. The dollars and cents associated with a word problem must less be than $10.00.
(4) The sum or the difference must be limited to a 3-digit number or less than $10.00. The difference must be positive or zero.
(5) No item must contain any irrelevant information.

EXAMPLE:- (1.) 11 small oranges, 23 big oranges. How
many oranges in all? ________________

APPENDIX A.3

Grade 4 Test Items

Domain 1
Adding 2, 3, or 4 numbers, each with 4 or fewer digits, with or without regrouping.

(1.) $\begin{array}{r} 3,749 \\ \underline{+2,134} \end{array}$

(2.) $\begin{array}{l} 3,456 \\ \underline{+97\quad} \end{array}$

(3.) $\begin{array}{r} 198 \\ 35 \\ \underline{+\ \ 5} \end{array}$

(4.) $\begin{array}{r} 3,567 \\ 879 \\ 734 \\ 87 \\ \underline{+\ \ \ 37} \end{array}$

(5.) $\begin{array}{r} 0 \\ \underline{+0} \end{array}$

(6.) $\begin{array}{r} 12 \\ \underline{+\ 6} \end{array}$

(7.) $\begin{array}{r} 23,457 \\ \underline{4,568} \end{array}$

(8.) $\begin{array}{r} 3,928 \\ \underline{+5,048} \end{array}$

(9.) $\begin{array}{r} 789 \\ \underline{+486} \end{array}$

(10.) $\begin{array}{r} 56 \\ 28 \\ \underline{+35} \end{array}$

(11.) $\begin{array}{r} 3,456 \\ \underline{+2,678} \end{array}$

(12.) $\begin{array}{r} 2,604 \\ 1,365 \\ \underline{2,969} \end{array}$

(13.) $\begin{array}{r} 2,968 \\ 1,309 \\ \underline{+2,348} \end{array}$

(14.) $\begin{array}{r} 417 \\ 94 \\ 649 \\ \underline{+\ \ 7} \end{array}$

(15.) 77 + 99 = ________

(16.) $\begin{array}{r} 5,918 \\ \underline{-\ \ 859} \end{array}$

(17.) $\begin{array}{r} 324,123 \\ \underline{+\ 57,978} \end{array}$

(18.) $\begin{array}{r} 4,281 \\ 3,039 \\ \underline{+1,945} \end{array}$

(19.) $\begin{array}{r} 11,157 \\ \underline{+\ 5,628} \end{array}$

(20.) $\begin{array}{r} 345 \\ 20 \\ 409 \\ \underline{+\ 27} \end{array}$

(21.) $\begin{array}{r} 68 \\ \underline{+\ 7} \end{array}$

(22.) $\begin{array}{r} 56 \\ 67 \\ 98 \\ 235 \\ \underline{+579} \end{array}$

(23.) $\begin{array}{r} 1,157 \\ \underline{+5,628} \end{array}$

(24.) $\begin{array}{l} 896 \\ \ 79 \\ \underline{+8\quad} \end{array}$

(25.) $\begin{array}{r} 726 \\ 478 \\ \underline{+137} \end{array}$

(26.) $\begin{array}{r} 238,789 \\ 578,668 \\ \underline{+\ 98,500} \end{array}$

(27.) $\begin{array}{r} 372 \\ \underline{-678} \end{array}$

(28.) $\begin{array}{r} 4,963 \\ \underline{+4,732} \end{array}$

(29.) $\begin{array}{r} 90 \\ \underline{+\ 8} \end{array}$

(30.) 19+ 2,789
=________

(31.) $\begin{array}{r} 1,237,950 \\ \underline{+\ \ 950,673} \end{array}$

(32.) $\begin{array}{r} 328 \\ \underline{-\ \ 9} \end{array}$

(33.) $\begin{array}{r} 2,154 \\ \underline{+1,367} \end{array}$

(34.) $\begin{array}{r} 4,169 \\ \underline{+7,896} \end{array}$

(35.) $\begin{array}{r} 976 \\ 248 \\ \underline{+\ 97} \end{array}$

(36.) $\begin{array}{r} 78 \\ \underline{67} \end{array}$

Domain 2

Identifying the best customary units of linear measurement for objects and distances, and changing units between inches, feet, yards, and miles.

A. Which unit (inches, feet, yards, or miles) would you use to measure each object or distance between two locations?

1. length of a room ____________________

2. width of this paper _________________

3. height of Mount Kilimanjaro ______________

4. length of a highway _________________

5. height of a tree ____________________

6. a book __________________

B. Complete:

7. 1 mi. = _______ ft.

8. __yds. = 9 ft.

9. 4 yds. = _______ in.

10. 2 ft. = _______ in.

11. 1760 yds. = _____ mi.

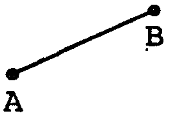
12. 4 ft. = _______ in.

13. -2 mi. = _______ yds.

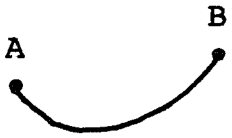
14. 3 yds. = _______ ft.

Domain 3
Identifying and naming points, lines, and line segments

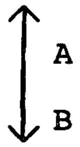
A. Which are line segments? (Write yes or no.)

1.) 

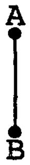
____

2.) 

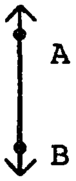
____

3.) 

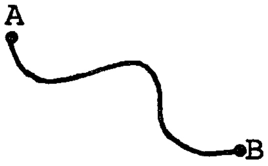
____

4.) 

____

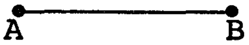
5.) 

____

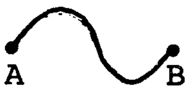
6.) 

____

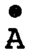
B. Name these as a point, a line, or a line segment.

7.) 

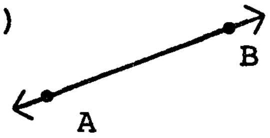
____

8.) 

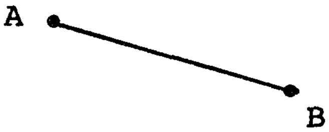
____

9.) 

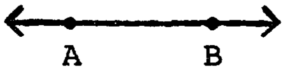
____

10.) 

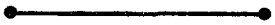
____

11.) 

____

12.) 

____

13.) 
____

Domain 4

Stating the place and value of an underlined digit in a standard numeral of 6 or less digits.

A. In which place is each underlined digit (word)?

1.) $33,\underline{3}33$ ________________________________

2.) $\underline{2},583$ ________________________________

3.) $\underline{7}47,477$ ________________________________

4.) $51\underline{9}$ ________________________________

5.) $6\underline{7}1,043$ ________________________________

6.) $3\underline{8}7,495$ ________________________________

7.) $\underline{6}74$ ________________________________

8.) $\underline{3},456,670$ ________________________________

9.) $\underline{2}3,448$ ________________________________

10.) $7\underline{0}2$ ________________________________

11.) $\underline{8}79,486$ ________________________________

B. What is the value of each underlined digit?

12.) $\underline{7}4,679$ ________________________________

13.) $9,4\underline{7}6$ ________________________________

14.) $\underline{6},5\underline{6}7$ ________________________________

15.) $\underline{8},760$ ________________________________

16.) $6\underline{8}2,947$ ________________________________

17.) $97,2\underline{8}8$ ________________________________

18.) $6\underline{5},007$ ________________________________

19.) $17\underline{7}$ ________________________________

20.) $\underline{6}71,059$ ________________________________

21.) $\underline{8}61$ ________________________________

22.) $\underline{8}64,543$ ________________________________

Domain 5

Identifying the numerators and denominators of fractions.

A. Name the numerators:

1.) $\frac{5}{3}$ ____________

2.) $\frac{2}{3}$ ____________

3.) 4/5 ____________

4.) $\frac{1}{5}$ ____________

5.) $\frac{7}{8}$ ____________

6.) $\frac{3}{6}$ ____________

B. Name the denominators:

7.) $\frac{1}{2}$ ____________

8.) $\frac{3}{5}$ ____________

9.) $\frac{2}{4}$ ____________

10.) $\frac{7}{3}$ ____________

11.) $\frac{3}{4}$ ____________

Domain 6

Writing dollars and cents value of a collection of coins and bills.

Write using the "$" sign:

1.) Four dollars and 99 cents ________________

2.) Two dollars and fifty cents ________________

3.) ________________________ = $50.60

4.) Thirteen dollars and thirteen cents __________

5.) Two cents and seven dollars ________________

6.) Eighteen dollars ____________

7.) One hundred and fifty cents ________________

8.) 99 dollars and 5 cents _______________

9.) Thirty-eight dollars and sixty-five cents ____________

10.) Five cents _________________

11.) Twelve dollars and fifteen cents ________________

12.) One thousand dollars __________________________

13.) Ninety dollars and four cents ___________________

14.) Four dollars ________________

15.) Two hundred and sixty-six cents ________________

16.) Ninety-nine dollars and eighteen cents __________

17.) One hundred dollars __________________

18.) Seventeen cents _____________

<u>DOMAIN 7</u>

Subtracting a standard numeral from a minuend with or without regrouping.

(1.) $\begin{array}{r}756\\ \underline{-\ 92}\end{array}$ (2.) $\begin{array}{r}5,637\\ \underline{-2,393}\end{array}$ (3.) $\begin{array}{r}726\\ \underline{-724}\end{array}$ (4.) 83-47=___

(5.) $\begin{array}{r}5,302\\ \underline{-2,748}\end{array}$ (6.) $\begin{array}{r}97\\ \underline{+37}\end{array}$ (7.) $\begin{array}{r}57\\ \underline{-\ 8}\end{array}$ (8.) $\begin{array}{r}5,241\\ \underline{-\ \ \ 56}\end{array}$

(9.) $\begin{array}{r}4,000\\ \underline{-1,987}\end{array}$ (10.) $\begin{array}{r}720\\ \underline{-659}\end{array}$ (11.) $\begin{array}{r}42\\ \underline{-13}\end{array}$ (12.) $\begin{array}{r}561\\ \underline{-9\ \ }\end{array}$

(13.) $\begin{array}{r}86\\ \underline{-25}\end{array}$ (14.) $\begin{array}{r}10,151\\ \underline{-\ 9,374}\end{array}$ (15.) $\begin{array}{r}340\\ \underline{-192}\end{array}$ (16.) $\begin{array}{r}2,724\\ \underline{-1,897}\end{array}$

(17.) $\begin{array}{r}90\\ \underline{-14}\end{array}$ (18.) $\begin{array}{r}380\\ \underline{-389}\end{array}$ (19.) $\begin{array}{r}3,076\\ \underline{-1,906}\end{array}$ (20.) $\begin{array}{r}8,625\\ \underline{-3,879}\end{array}$

(21.) $\begin{array}{r}8,353\\ \underline{-2,063}\end{array}$ (22.) $\begin{array}{r}7,498\\ \underline{-3,128}\end{array}$ (23.) $\begin{array}{r}1,437\\ \underline{-\ \ \ 848}\end{array}$ (24.) $\begin{array}{r}234,675\\ \underline{-\ \ 6,897}\end{array}$

(25.) $\begin{array}{r}391\\ \underline{-\ \ \ 8}\end{array}$ (26.) $\begin{array}{r}97\\ \underline{-\ 7}\end{array}$

Domain 8

Identifying the best metric unit for measuring a given linear attribute of an object and distances, and changing units between centimeters and meters, and between meters and kilometers.

A. Which unit (centimeters, meters, or kilometers) would you use to measure each object or distance between two locations?

1. distance between New York and Miami ________________

2. Length of a floppy diskette ________________________

3. width of a television ______________________

4. length of a pencil _________________________

5. length of a bird _______________________

6. Size of a cup ________________________

7. height of a house _________________________

B. Complete:

8. 8m = ____________ cm

9. 1000cm
 = ____________m

10. 2000m = _________ km

11. 5cm = ___________ cm

12. 5.5km = _________ m

13. 200cm = _________ m

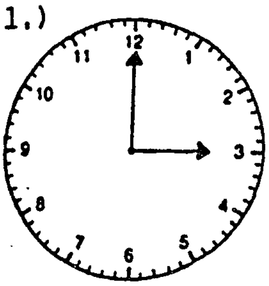
14. 6km = ___________ m

Domain 9

Telling the time shown on a round analog clock to the hour, half-hour, quarter-hour, and minute.

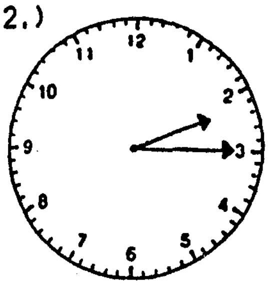
Write the time to the minute:

(1.)



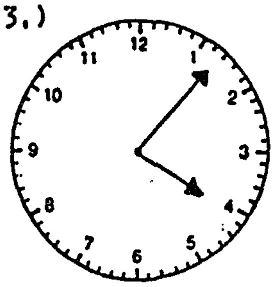
____

(2.)



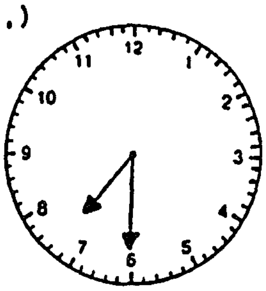
____

(3.)



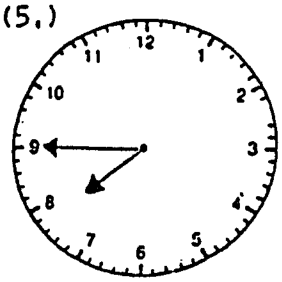
____

(4.)



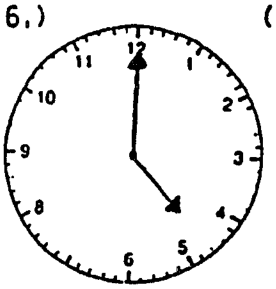
____

(5.)



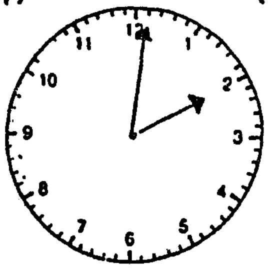
____

(6.)



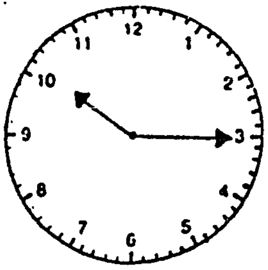
____

(7.)



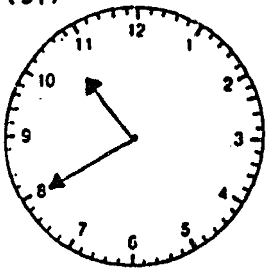
____

(8.)



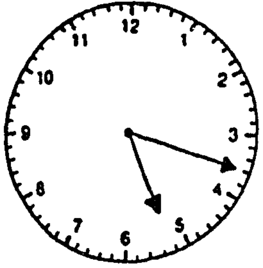
____

(9.)



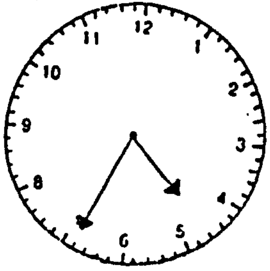
____

(10.)



____

(11.)



____

DOMAIN 10
Matching word names with standard numerals and writing standard numerals for word names to one million.

A. Match the word names with the numerals. Write the letter in the space provided.

_____	1	One million, five hundred	a. 45,000
_____	2	Twenty thousand, three hundred seventy-five.	b. 705,670
_____	3	Forty thousand, two hundred two	c. 1,000,500
_____	4	Ninety	d. 1,304
_____	5	Nine hundred, seventy-four.	e. 28
_____	6	62 thousand, 2 ones	f. 5
_____	7	Six thousand, four hundred.	g. 600,004
_____	8	Two thousand, two hundred thirty-four.	h. 6,400
_____	9	Four hundred fifty-three	i. 20,375
_____	10	Twenty-eight	j. 90
_____	11	62,002	k. five
_____	12	Seven hundred and five thousand, six hundred seventy.	l. 2,234
_____	13	One thousand, three hundred four	m. 974
_____	14	Six hundred thousand, four.	n. 453

B. WRITE STANDARD NUMERALS.

15. Negative eighty-eight (15) ____________

16. One million, one hundred (16) ____________

17. Ninety-three (17) ____________

18. 6 tens and 5 ones (18) ____________

19. Sixteen thousand, one hundred
eighty-six (19) ____________

20. Six thousands, four hundreds,
zero tens, and three ones (20) ____________

21. One million, six hundred thirty-five (21)____________

22. Five hundred twenty-four (22) ____________

23. Two hundred thousand, three hundred
seventy-five (23) ____________

24. 8 hundreds, 3 tens and 2 ones (24) ____________

25. Three hundred seventy-five thousand (25) ____________

26. 356,456 (26) ____________

27. Seventeen ones (27) ____________

28. Three thousand , six hundred
thirty-four (28) ____________

Domain 11

Writing word names for whole numbers from one through one hundred.

Write word name:

1. 75 ____________________

2. 9 ____________________

3. 101 ____________________

4. 17 ____________________

5. -40 ____________________

6. 28 ____________________

7. 43 ____________________

8. Ten ____________________

9. 36 ____________________

10. 71 ____________________

11. 10 + 12 = ____________________

12. 62 ____________________

13. 0 ____________________

14. 9 tens 3 ones ____________________

15. 99 ____________________

16. ____________________ = 56

17. 100 ____________________

18. XXIV ____________________

Domain 12

Choosing the correct operation (addition and subtraction) to solve one-step word problems and solving the problems.

Test Items

(1.) Nick paid $3.49 for a hat.
Tony paid $2.95 for his hat.
How much more did Nick pay
than Tony? ________________

(2.) Kathy bought 6 apples and
8 pears. How many pieces of
fruit did she buy? ________________

(3.) Amos has access to 7 kilobytes
of core memory. He needs a total of
128 kilobytes to run his program.
How many more kilobytes does
Amos need? ____________________

(4.) Andy had 145 baseball cards.
He gave 76 to Jeffrey. How
many did Andy have left? ________________

(5.) There are 27 people in the
room. There are 19 chairs
in the room. How many more
chairs are needed so
everyone can sit down? ________________

(6.) Mary walks 9 blocks to school
each day. Tommy walks 6 blocks
to school each day. How many
more blocks does Mary walk? ______________

(7.) 8 dozen pine trees. 30 oak tress.
How many trees in all? ______________________

(8.) Mr. Smith is a postal worker.
He delivered 189 letters in the
morning. He delivered 261 letters
in the afternoon. How many letters
did he deliver in all? ____________________

(9.) Maria went to school 21 days in
May and 11 days in June. How
many days did she go in all?___________________

(10.) 10 cities, 6 towns
How many more cities? ______________________

(11.) 8 small cars, 9 large cars.
How many in all? __________________________

(12.) 1000 tests to mark. 80 were
marked yesterday. How
many more to mark? ________________________

(13.) Mrs Davis had a birthday party.
She bought 39 hats. She bought
67 balloons. How many more
balloons than hats were there? _______________

APPENDIX B.1

Item-Domain Rating Instrument

INSTRUCTIONS

The purpose of this instrument is to decide if individual items fit within the domain specification. A domain specification includes the details describing an area of Grade 4 mathematics for which a competency test is to be constructed. A domain specification includes a statement that defines an area of Grade 4 mathematics, test item formats, and conditions under which test items will be constructed (see Grade 4 Domain Specifications). A test item measures a particular domain if that item asks the student to demonstrate knowledge defined by that domain. For example, the item, 77 + 98 = requires that a student demonstrate his/her knowledge of how to add two 2-digit numbers. An item which violates any domain specification is not a measure of that particular domain.

First, read each domain specification and test item carefully. Next, indicate whether or not you feel each item satisfies the requirements of the domain it has been written to measure. Please rate each item solely on the basis of the correspondence between its characteristics and the content specified in the domain that the test item was written to measure. Use the rating scale below. Please check (✓) the column corresponding to your rating ("YES" or "NO") beside EACH dimension of judgment for each test item. If a dimension of judgment does not apply to an item write "N/A" to indicate that the dimension is not applicable.

Dimension of Judgments

FORMAT (Is the way the facts are arranged for this item appropriate for measuring this domain? e.g., vertical or horizontal arrangement of items)

WORDING (Are the words used to state the problem for this item simple enough and within the recognition vocabulary of Grade 4 students?)

NUMBERS (Do the numbers in this item agree with the range of the numbers required for this domain?)

BEHAVIOR (Does this item elicit the behavior or knowledge to be measured by the domain?)

OVERALL (Overall, do you feel this item is a measure of the domain for which it has been written?)

Item-Domain Rating Instrument

DOMAIN	Test Item	Item Ratings									
		FORMAT		WORDING		NUMBERS		BEHAVIOR		OVERALL	
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO
1	1										
	2										
	3										
	4										
	5										
	6										
	7										
	8										
	9										
	10										
	11										
	12										
	13										
	14										
	15										
	16										
	17										
	18										
	19										
	20										
	21										

Item-Domain Rating Instrument

DOMAIN	Test Item	Item Ratings									
		FORMAT		WORDING		NUMBERS		BEHAVIOR		OVERALL	
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO
1	22										
	23										
	24										
	25										
	26										
	27										
	28										
	29										
	30										
	31										
	32										
	33										
	34										
	35										
	36										
2	1										
	2										
	3										
	4										
	5										
	6										

Item-Domain Rating Instrument

DOMAIN	Test Item	Item Ratings									
		FORMAT		WORDING		NUMBERS		BEHAVIOR		OVERALL	
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO
2	7										
	8										
	9										
	10										
	11										
	12										
	13										
	14										
3	1										
	2										
	3										
	4										
	5										
	6										
	7										
	8										
	9										
	10										
	11										
	12										
	13										

Item-Domain Rating Instrument

DOMAIN	Test Item	Item Ratings									
		FORMAT		WORDING		NUMBERS		BEHAVIOR		OVERALL	
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO
4	1										
	2										
	3										
	4										
	5										
	6										
	7										
	8										
	9										
	10										
	11										
	12										
	13										
	14										
	15										
	16										
	17										
	18										
	19										
	20										
	21										

Item-Domain Rating Instrument

DOMAIN	Test Item	Item Ratings									
		FORMAT		WORDING		NUMBERS		BEHAVIOR		OVERALL	
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO
4	22										
5	1										
	2										
	3										
	4										
	5										
	6										
	7										
	8										
	9										
	10										
	11										
6	1										
	2										
	3										
	4										
	5										
	6										
	7										
	8										
	9										

Item-Domain Rating Instrument

DOMAIN	Test Item	Item Ratings									
		FORMAT		WORDING		NUMBERS		BEHAVIOR		OVERALL	
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO
6	10										
	11										
	12										
	13										
	14										
	15										
	16										
	17										
	18										
7	1										
	2										
	3										
	4										
	5										
	6										
	7										
	8										
	9										
	10										
	11										
	12										

Item-Domain Rating Instrument

DOMAIN	Test Item	Item Ratings									
		FORMAT		WORDING		NUMBERS		BEHAVIOR		OVERALL	
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO
7	13										
	14										
	15										
	16										
	17										
	18										
	19										
	20										
	21										
	22										
	23										
	24										
	25										
	26										
8	1										
	2										
	3										
	4										
	5										
	6										
	7										

Item-Domain Rating Instrument

DOMAIN	Test Item	Item Ratings									
		FORMAT		WORDING		NUMBERS		BEHAVIOR		OVERALL	
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO
8	8										
	9										
	10										
	11										
	12										
	13										
	14										
9	1										
	2										
	3										
	4										
	5										
	6										
	7										
	8										
	9										
	10										
	11										
10	1										
	2										
	3										

Item-Domain Rating Instrument

DOMAIN	Test Item	Item Ratings									
		FORMAT		WORDING		NUMBERS		BEHAVIOR		OVERALL	
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO
10	4										
	5										
	6										
	7										
	8										
	9										
	10										
	11										
	12										
	13										
	14										
	15										
	16										
	17										
	18										
	19										
	20										
	21										
	22										
	23										
	24										

Item-Domain Rating Instrument

DOMAIN	Test Item	Item Ratings									
		FORMAT		WORDING		NUMBERS		BEHAVIOR		OVERALL	
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO
10	25										
	26										
	27										
	28										
11	1										
	2										
	3										
	4										
	5										
	6										
	7										
	8										
	9										
	10										
	11										
	12										
	13										
	14										
	15										
	16										
	17										

Item-Domain Rating Instrument

DOMAIN	Test Item	Item Ratings									
		FORMAT		WORDING		NUMBERS		BEHAVIOR		OVERALL	
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO
11	18										
12	1										
	2										
	3										
	4										
	5										
	6										
	7										
	8										
	9										
	10										
	11										
	12										
	13										

APPENDIX B.2:

Item/Domain Matching Instrument

INSTRUCTIONS

The purpose of this instrument is to decide if individual test items fit within the domain specification. A domain specification includes the details describing an area of Grade 4 mathematics for which a competency test is to be constructed. A domain specification includes a statement that defines an area of Grade 4 mathematics, item formats, and conditions under which test items will be constructed (see Grade 4 Domain Specifications). A test item measures a particular domain if that item asks the student to demonstrate knowledge defined by that domain. For example, the item, 77 + 98 = requires that a student demonstrate his/her knowledge of how to add two 2-digit numbers. An item which violates any domain specification is not a measure of that particular domain.

Read the lists of domain specifications carefully. Take each domain specification and find test items within the domain specification that you feel measure that domain. Beside each domain, write the item numbers corresponding to the test items that you feel measure that domain. In some instances, you may feel that an item does not measure any of the available domain specifications. In that case write these test item numbers in the space provided at the end of that particular domain, under the heading "Items that Do Not Measure the Domain."

Item/Domain Matching Instrument

Domain Number	Items that Measure the Domain						
1							
	Items that Do Not Measure the Domain						
2	Items that Measure the Domain						
	Items that Do Not Measure the Domain						

Item/Domain Matching Instrument

Domain Number	Items that Measure the Domain						
3							
	Items that Do Not Measure the Domain						
4	Items that Measure the Domain						
	Items that Do Not Measure the Domain						

Item/Domain Matching Instrument

Domain Number	Items that Measure the Domain						
5							
	Items that Do Not Measure the Domain						
6	Items that Measure the Domain						
	Items that Do Not Measure the Domain						

Item/Domain Matching Instrument

Domain Number	Items that Measure the Domain						
7							
	Items that Do Not Measure the Domain						
8	Items that Measure the Domain						
	Items that Do Not Measure the Domain						

Item/Domain Matching Instrument

Domain Number	Items that Measure the Domain						
9							
	Items that Do Not Measure the Domain						
10	Items that Measure the Domain						
	Items that Do Not Measure the Domain						

Item/Domain Matching Instrument

Domain Number	Items that Measure the Domain						
11							
	Items that Do Not Measure the Domain						
12	Items that Measure the Domain						
	Items that Do Not Measure the Domain						

APPENDIX B.3

Domain Representativeness Instrument

INSTRUCTIONS

The purpose of this instrument is to decide if an entire collection of items measuring a domain is representative of the domain. Complete this instrument only after you have completed rating or matching all the individual fourth grade mathematics items of a domain. The scope of a domain defines a body of mathematics knowledge for Grade 4 students. For example, the scope of a domain might cover the knowledge of addition of three two-digit numbers with and without regrouping. Since addition with and without regrouping defines two types of knowledge, test items must be constructed for each type of knowledge. The number of items constructed to measure each type of knowledge defined for a domain should correspond to the relative amount of time spent on teaching that knowledge. Therefore, a test constructed to determine mastery of the knowledge defined by a domain should contain a sample of items that test each area of knowledge (in proportion to the time spent teaching each area of knowledge). An entire collection of items is REPRESENTATIVE of a domain if: (1) that collection of items covers the scope of the domain they have been written to measure, and (2) the numbers of items that assess mastery of each area of knowledge are proportionate to the instructional time devoted to teaching that area.

Your task is to indicate whether or not each collection of items you rated as measures of a domain (on the Item-Domain Rating Instrument) is or is not representative of that domain. Read through each Grade 4 domain specification, paying close attention to the number of items which must be constructed for each condition of testing of that domain.

(1.) For each domain, list the item numbers of all items you (a) rated as "YES" on the OVERALL dimension of the Item-Domain Rating Instrument, or (b) assigned as measures of the domain on the Item/Domain Matching Instrument. Use the spaces headed "Collection of Items" on the instrument below.

(2.) Look through the entire collection of items you rated as measures of each domain (items you rated as "YES" on the OVERALL dimension of the Item-Domain Rating Instrument or assigned as measures of the domain on the Item /Domain Matching Instrument). Then, in the "Rating of Representativeness" columns below, please check (✓) "YES" if you feel a collection of items is representative of the corresponding domain, check "NO" otherwise.

Domain Representativeness Instrument

DOMAIN	Collection of Items	Rating of Representativeness	
		YES	NO
1			
2			
3			
4			
5			
6			

Domain Representativeness Instrument

DOMAIN	Collection of Items	Rating of Representativeness YES	NO
7			
8			
9			
10			
11			
12			

APPENDIX B.4

Inservice Training Instruments

A. Introduction to Achievement Testing

I. Background Definitions:

(1) Measurement -

(2) Evaluation -

(3) Precision -

II. Types of Achievement Tests

(1) Teacher Made vs. Published Tests -

(2) Criterion-Referenced vs. Norm Referenced Tests -

(3) Objective vs. Subjective Tests -

a. fixed-response items -

b. short answer items -

c. essay questions -

III. The Type of Test and Bloom's Taxonomy

(1) Knowledge -

(2) Comprehension -

(3) Application -

(4) Analysis -

(5) Synthesis -

(6) Evaluation -

IV. Validity

(1) Content -

(2) Construct -

(3) Criterion Related -

a. concurrent -

b. predictive -

V. Reliability and Validity

B. Greensboro Public Schools' Mathematics Promotion Standards: Two Objectives of Grade 3 Tests

Objective 1.
Masters addition and subtraction facts to 18 and solves simple word problems using these facts, either orally or in writing.

Objective 2.
Tells time to the quarter hour.

C. Domain Specifications for Two Objectives of Grade 3 Tests

<u>DOMAIN 1a</u>

Adding two numbers and solving simple word problems orally or in writing using two word facts.

<u>CONDITIONS FOR ITEM CONSTRUCTION AND TESTING</u>

(1) The student is given two whole numbers and is asked to provide the sum.

(2) The student is given two simple word facts and is asked to provide the sum.

<u>Number of Sample Items:</u>

(a) Twenty items must test addition operations using whole numbers.
(b) Five items must test addition operations using word facts.

<u>Format:</u>

(1) The second addend of the number facts must be written underneath the first addend. A plus sign must be written in front of the second addend, and must be vertically aligned with a blank space to the left of the first addend. The second addend and the plus sign must be underlined.

(2) Standard numerals must be used for expressing the word facts. Each number of the word problem and each question must start on a new line. A question line can span two lines and must contain the space for the student's response. The word facts and the question lines must be vertically aligned to the left.

<u>BOUNDARIES FOR ITEM CONSTRUCTION</u>

(1) Each addend must be between 3 and 9.
(2) The sum must be non-negative, and must be less than or equal to 18.
(3) The words used for defining an addition problem must be within the recognition vocabulary of Grade 3 students.

<u>EXAMPLES</u>:- (1.) $\begin{array}{r}4\\ \underline{+5}\end{array}$ (2.) 6 big pencils
1 small pencils
How many pencils in all?____

DOMAIN 1b

Subtracting two numbers and solving simple word problems orally or in writing using word facts.

CONDITIONS FOR ITEM CONSTRUCTION AND TESTING

(1) The student is given two whole numbers and is asked to provide the difference.

(2) The student is given two simple word facts and is asked to provide the difference.

Number of Sample Items:

(a) Twenty items must test subtraction operations using whole numbers.

(b) Five items must test subtraction operations using word facts.

Format:

(1) The subtrahend must be written underneath the minuend, in the order of the place values of the digits of the minuend. A minus sign must be written in front of the subtrahend, and must be vertically aligned with a blank space to the left of the minuend. The minuend and the minus sign must be underlined.

(2) Standard numerals must be used for expressing the word facts. Each number of the word problem and each question must start on a new line. A question line can span two lines and must contain the space for the student's response. The word facts and the question lines must be vertically aligned to the left.

BOUNDARIES FOR ITEM CONSTRUCTION

(1) The two numbers and their difference must be non-negative, and each number and the difference must be less than or equal to 18.

(2) The words used for defining a subtraction problem must be within the recognition vocabulary of Grade 3 students.

EXAMPLES:- (1.)

$$\begin{array}{r} 8 \\ \underline{-4} \end{array}$$

(2.) 6 large eggs
2 eggs broken
How many are left?__________

DOMAIN 2

Telling time on a clock to the hour, half-hour, and quarter hour.

CONDITIONS FOR ITEM CONSTRUCTION AND TESTING

(1) The student is given a time to the hour shown on a clock and is asked to write the time.

(2) The student is given a time to the half-hour shown on a clock and is asked to write the time.

(3) The student is given a time to the quarter-hour shown on a clock, either before or after the hour and is asked to write the time.

Number of Sample Items:

(a) One test item must test knowledge of time to the hour.

(b) Three test items must test knowledge of time to the half-hour

(c) Three test items must test knowledge of time to the quarter-hour before the hour, and three test items must test knowledge of time to the quarter-hour after the hour.

Format:

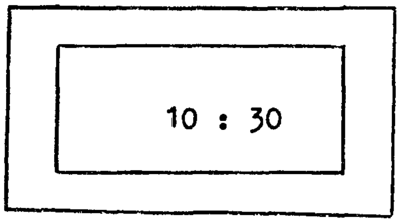
The hour and the minute hands shown on a round analog (with numbers and hands) clock must correspond to the time in a question.

BOUNDARIES FOR ITEM CONSTRUCTION

(1) The clock must be a digital rectangular clock or a round analog clock.

(2) The time on a digital rectangular clock must be displayed showing only the hour and the minute.

EXAMPLE:-



D. Check Lists for Three Domains of Grade 3 Tests

Check List 1a
Task 1: Add 2 numbers

	Check One	
	YES	NO
Is each addend between 3 and 9?	[ ]	[ ]
Does this item test an addition operation?	[ ]	[ ]
Are the 2 addends vertically arranged and aligned?	[ ]	[ ]

Task 2: Solve addition word problem

	Check One	
	YES	NO
Is each number of the word problem between 3 and 9?	[ ]	[ ]
Does this item test an addition operation?	[ ]	[ ]
Are the words used to state the problem within the recognition vocabulary of Grade 3 students?	[ ]	[ ]
Are the word facts and question lines vertically aligned to the left?	[ ]	[ ]

Check List 1b
Task 1: Subtract 2 numbers

	Check One	
	YES	NO
Is each number between 0 and 18?	[ ]	[ ]
Is the minuend larger than the subtrahend?	[ ]	[ ]
Does this item test a subtraction operation?	[ ]	[ ]
Are the minuend and subtrahend vertically arranged and aligned to the right?	[ ]	[ ]

Task 2: Solve subtraction word problem

	Check One	
	YES	NO
Is each number of the word problem between 0 and 18?	[ ]	[ ]
Is the minuend larger than the subtrahend?	[ ]	[ ]
Does this item test a subtraction operation?	[ ]	[ ]
Are the words used to state the problem within the recognition vocabulary of Grade 3 students?	[ ]	[ ]
Are the word facts and question lines vertically aligned to the left?	[ ]	[ ]

Check List 2.
Task: Tells Time on a clock.

	Check One	
	YES	NO
Is the time displayed showing the hour and the minute on a rectangular digital clock or showing the hour and the minute hands on a round analog clock?	[ ]	[ ]
Is the time shown correctly to the hour, or half-hour, or quarter-hour?	[ ]	[ ]

E. Test Items for Item-Domain Rating and Item/Domain Matching

DOMAIN 1a
Adding two numbers and solving simple word problems orally or in writing using word facts.

1) 5 basketballs
8 baseballs
How many balls
in all? ________________

2) 3
+8

3) 0
+1

4) 9
-4

5) 9 + 7=____

6) 6
+6

7) 8
+8

8) 8
+9

9) 6 red birds
9 blue birds
How many in all? _______________

10) 7
+9

11) 10
+ 7

12) 7
+4

13) 14 slippers
7 boots
How many shoes
in all? __________________

14) 5
+4

15) 6
+4

16) 7 small microchips
2 big microchips
How many microchips
in all? __________________

17) 8
+7

18) 5
+2

19) 9 -6 = ___

20) 5
+7

21) 5 green grapes
7 purple grapes
How many grapes
in all? ______

22) 3
+9

23) 6
+7

24) 7 small dogs
9 big dogs
How many dogs
in all? ________

25) 9
+4

26) 7
+7

27) 8
+6

28) 5
+8

29) 9
+9

30) -8 + 7 =_______

DOMAIN 1b

Subtracting two numbers and solving simple word problems orally or in writing using word facts.

1) $\begin{array}{r}10\\ \underline{-\ 2}\end{array}$

2) 13 cups
6 saucers
How many more
cups than saucers? ____

3) $\begin{array}{r}8\\ \underline{-9}\end{array}$

4) $\begin{array}{r}15\\ \underline{-\ 7}\end{array}$

5) $\begin{array}{r}18\\ \underline{-\ 9}\end{array}$

6) $\begin{array}{r}14\\ \underline{-\ 6}\end{array}$

7) $\begin{array}{r}9\\ \underline{-5}\end{array}$

8) $\begin{array}{r}13\\ \underline{-\ 7}\end{array}$

9) $\begin{array}{r}11\\ \underline{-\ 8}\end{array}$

10) $\begin{array}{r}12\\ \underline{-\ 3}\end{array}$

11) 9 eggs
All broken
How many are left?____

12) $\begin{array}{r}16\\ \underline{-\ 7}\end{array}$

13) $\begin{array}{r}10\\ \underline{-\ 6}\end{array}$

14) $\begin{array}{r}19\\ \underline{-\ 8}\end{array}$

15) $\begin{array}{r}13\\ \underline{-\ 8}\end{array}$

16) $\begin{array}{r}12\\ \underline{-\ 6}\end{array}$

17) $\begin{array}{r}16\\ \underline{-\ 8}\end{array}$

18) $\begin{array}{r}13\\ \underline{-\ 4}\end{array}$

19) $\begin{array}{r}17\\ \underline{-\ 9}\end{array}$

20) $\begin{array}{r}12\\ \underline{-\ 7}\end{array}$

21) 17 grey kittens
8 black kittens
How many more
grey kittens?_____

22) $\begin{array}{r}9\\ \underline{+8}\end{array}$

23) $\begin{array}{r}15\\ \underline{-\ 6}\end{array}$

24) $\begin{array}{r}11\\ \underline{-\ 4}\end{array}$

25) 15 apples
7 apples eaten
How many apples
are left? _______

26) $\begin{array}{r}14\\ \underline{-\ 7}\end{array}$

27) $\begin{array}{r}14\\ \underline{-\ 9}\end{array}$

28) $\begin{array}{r}13\\ \underline{-14}\end{array}$

29) 8 small cars
6 cars drive away
How many cars
are left? _______

30) 18 - 14 = ______

31) $\begin{array}{r}18\\ \underline{-19}\end{array}$

32) 14 shoes
7 boots
How many more shoes
than boots? ________

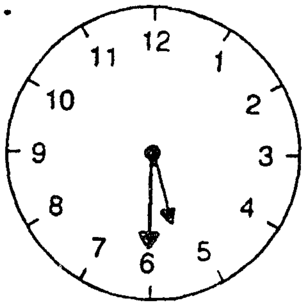
33) 7 microprocessors
9 macroprocessors
How many more microprocessors
than macroprocessors? ______

34) 18 chairs
9 desks
How many more chairs
than desks ? ______

DOMAIN 2

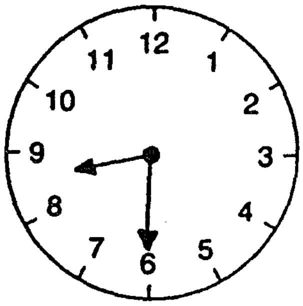
Telling time on a clock to the hour, half-hour, and quarter hour.

1.



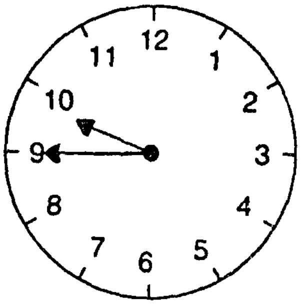
______

2.



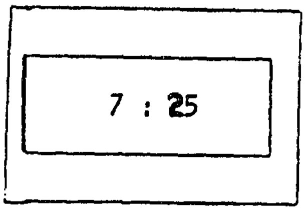
______

3.



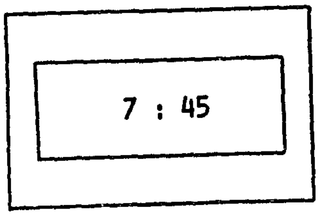
______

4.



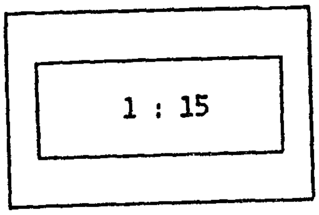
______

5.



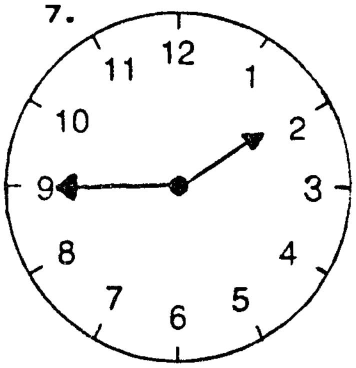
______

6.

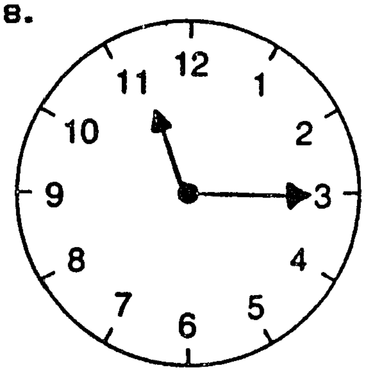


______

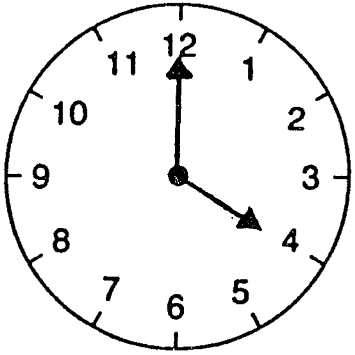
7.


12
1
2
3
4
5
6
7
8
9
10
11

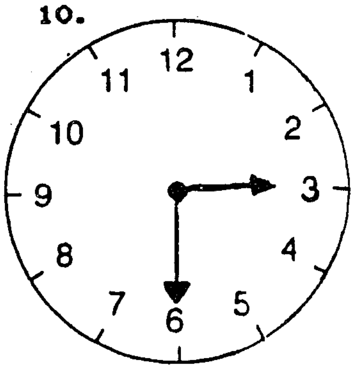
8.


12
1
2
3
4
5
6
7
8
9
10
11

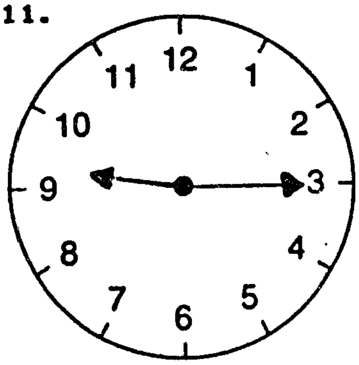
9.


12
1
2
3
4
5
6
7
8
9
10
11

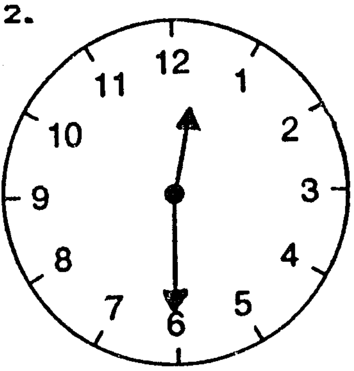
10.


12
1
2
3
4
5
6
7
8
9
10
11

11.


12
1
2
3
4
5
6
7
8
9
10
11

12.


12
1
2
3
4
5
6
7
8
9
10
11

F. Item-Domain Rating Instrument for Grade 3 Tests

INSTRUCTIONS

The purpose of this instrument is to decide if individual items fit within the domain specification. A domain specification includes the details describing an area of Grade 3 mathematics for which a competency test is to be constructed. A domain specification includes a statement that defines an area of Grade 3 mathematics, test item formats, and conditions under which test items will be constructed (see Grade 3 Domain Specifications). A test item measures a particular domain if that item asks the student to demonstrate knowledge defined by that domain. For example, the item, 7 + 8 = requires that a student demonstrate his/her knowledge of how to add two 1-digit numbers. An item which violates any domain specification is not a measure of that particular domain.

First, read each domain specification and test item carefully. Next, indicate whether or not you feel each item satisfies the requirements of the domain it has been written to measure. Please rate each item solely on the basis of the correspondence between its characteristics and the content specified in the domain that the test item was written to measure. Use the rating scale below. Please check (✓) the column corresponding to your rating ("YES" or "NO") beside EACH dimension of judgment for each test item. If a dimension of judgment does not apply to an item write "N/A" to indicate that the dimension is not applicable.

Dimension of Judgments

FORMAT (Is the way the facts are arranged for this item appropriate for measuring this domain? e.g., vertical or horizontal arrangement of items)

WORDING (Are the words used to state the problem for this item simple enough and within the recognition vocabulary of Grade 3 students?)

NUMBERS (Do the numbers in this item agree with the range of the numbers required for this domain?)

BEHAVIOR (Does this item elicit the behavior or knowledge to be measured by the domain?)

OVERALL (Overall, do you feel this item is a measure of the domain for which it has been written?)

Item-Domain Rating Instrument

DOMAIN	Test Item	Item Ratings									
		FORMAT		WORDING		NUMBERS		BEHAVIOR		OVERALL	
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO
1a	1										
	2										
	3										
	4										
	5										
	6										
	7										
	8										
	9										
	10										
	11										
	12										
	13										
	14										
	15										
	16										
	17										
	18										
	19										
	20										
	21										

Item-Domain Rating Instrument

DOMAIN	Test Item	Item Ratings									
		FORMAT		WORDING		NUMBERS		BEHAVIOR		OVERALL	
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO
1a	22										
	23										
	24										
	25										
	26										
	27										
	28										
	29										
	30										
1b	1										
	2										
	3										
	4										
	5										
	6										
	7										
	8										
	9										
	10										
	11										
	12										

Item-Domain Rating Instrument

DOMAIN	Test Item	Item Ratings									
		FORMAT		WORDING		NUMBERS		BEHAVIOR		OVERALL	
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO
1b	13										
	14										
	15										
	16										
	17										
	18										
	19										
	20										
	21										
	22										
	23										
	24										
	25										
	26										
	27										
	28										
	29										
	30										
	31										
	32										
	33										

Item-Domain Rating Instrument

DOMAIN	Test Item	Item Ratings									
		FORMAT		WORDING		NUMBERS		BEHAVIOR		OVERALL	
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO
1b	34										
	35										
	36										
2	1										
	2										
	3										
	4										
	5										
	6										
	7										
	8										
	9										
	10										
	11										
	12										

G. Item/Domain Matching Instrument

INSTRUCTIONS

The purpose of this instrument is to decide if individual test items fit within the domain specification. A domain specification includes the details describing an area of Grade 3 mathematics for which a competency test is to be constructed. A domain specification includes a statement that defines an area of Grade 3 mathematics, item formats, and conditions under which test items will be constructed (see Grade 3 Domain Specifications). A test item measures a particular domain if that item asks the student to demonstrate knowledge defined by that domain. For example, the item, 7 + 8 = requires that a student demonstrate his/her knowledge of how to add two 1-digit numbers. An item which violates any domain specification is not a measure of that particular domain.

Read the lists of domain specifications carefully. Take each domain specification and find test items within the domain specification that you feel measure that domain. Beside each domain, write the item numbers corresponding to the test items that you feel measure that domain. In some instances, you may feel that an item does not measure any of the available domain specifications. In that case write these test item numbers in the space provided at the end of that particular domain, under the heading "Items that Do Not Measure the Domain."

Item/Domain Matching Instrument

Domain Number	Items that Measure the Domain						
1a							
	Items that Do Not Measure the Domain						
1b	Items that Measure the Domain						
	Items that Do Not Measure the Domain						

Item/Domain Matching Instrument

Domain Number	Items that Measure the Domain						
2							
	Items that Do Not Measure the Domain						

H. Domain Representativeness Instrument

INSTRUCTIONS

The purpose of this instrument is to decide if an entire collection of items measuring a domain is representative of the domain. Complete this instrument only after you have completed rating or matching all the individual third grade mathematics items of a domain. The scope of a domain defines a body of mathematics knowledge for Grade 3 students. For example, the scope of a domain might cover the knowledge of addition of two two-digit numbers with and without regrouping. Since addition with and without regrouping defines two types of knowledge, test items must be constructed for each type of knowledge. The number of items constructed to measure each type of knowledge defined for a domain should correspond to the relative amount of time spent on teaching that knowledge. Therefore, a test constructed to determine mastery of the knowledge defined by a domain should contain a sample of items that test each area of knowledge (in proportion to the time spent teaching each area of knowledge). An entire collection of items is REPRESENTATIVE of a domain if: (1) that collection of items covers the scope of the domain they have been written to measure, and (2) the numbers of items that assess mastery of each area of knowledge are proportionate to the instructional time devoted to teaching that area.

Your task is to indicate whether or not each collection of items you rated as measures of a domain (on the Item-Domain Rating Instrument) is or is not representative of that domain. Read through each Grade 3 domain specification, paying close attention to the number of items which must be constructed for each condition of testing of that domain.

(1.) For each domain, list the item numbers of all items you (a) rated as "YES" on the OVERALL dimension of the Item-Domain Rating Instrument, or (b) assigned as measures of the domain on the Item/Domain Matching Instrument. Use the spaces headed "Collection of Items" on the instrument below.

(2.) Look through the entire collection of items you rated as measures of each domain (items you rated as "YES" on the OVERALL dimension of the Item-Domain Rating Instrument or assigned as measures of the domain on the Item /Domain Matching Instrument). Then, in the "Rating of Representativeness" columns below, please check (✓) "YES" if you feel a collection of items is representative of the corresponding domain, check "NO" otherwise.

Domain Representativeness Instrument

DOMAIN	Collection of Items	Rating of Representativeness	
		YES	NO
1a			
1b			
2			

APPENDIX C.1

Number and Percent of Teachers Expressing Indicated Judgments on Whether Test Items are Measures of the Domains.

DOMAIN	Test Item	Item Ratings										Item Matching	
		FORMAT		NUMBERS		WORDING		BEHAVIOR		OVERALL			
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	MATCH	NO MATCH
1	1 #	27	1	28	0	(*)	(*)	28	0	27	1	28	0
	%	96.4	3.6	100	0	(*)	(*)	100	0	96.4	3.6	100	0
	2 #	1	27	23	5	(*)	(*)	21	7	2	26	0	28
	%	3.6	96.4	82.1	17.9	(*)	(*)	75.0	25.0	7.1	92.9	0	100
	3 #	28	0	28	0	(*)	(*)	26	2	26	2	27	1
	%	100	0	100	0	(*)	(*)	92.9	7.1	92.9	7.1	96.4	3.6
	4 #	22	6	16	12	(*)	(*)	15	13	10	18	6	22
	%	78.6	21.4	57.1	42.9	(*)	(*)	53.6	46.4	35.7	64.3	21.4	78.6
	5 #	22	6	21	7	(*)	(*)	17	11	16	12	19	9
	%	78.6	21.4	75.0	25.0	(*)	(*)	60.7	39.3	57.1	42.9	67.9	32.1
	6 #	27	1	27	1	(*)	(*)	25	3	24	4	25	3
	%	96.4	3.6	96.4	3.6	(*)	(*)	89.3	10.7	85.7	14.3	89.3	10.7
	7 #	20	8	5	23	(*)	(*)	14	14	6	22	5	23
	%	71.4	28.6	17.9	82.1	(*)	(*)	50.0	50.0	21.4	78.6	17.9	82.1
	8 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	9 #	28	0	28	0	(*)	(*)	28	0	28	0	27	1
	%	100	0	100	0	(*)	(*)	100	0	100	0	96.4	3.6
	10 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	11 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	12 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	13 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0

NOTE
(*) indicates that the dimension of judgment did not apply to this item

DOMAIN	Test Item	Item Ratings										Item Matching	
		FORMAT		NUMBERS		WORDING		BEHAVIOR		OVERALL			
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	MATCH	NO MATCH
1	14 #	27	1	27	1	(*)	(*)	26	2	25	3	26	2
	%	96.4	3.6	96.4	3.6	(*)	(*)	92.9	7.1	89.3	10.7	92.9	7.1
	15 #	2	26	25	3	(*)	(*)	18	10	2	26	0	28
	%	7.1	92.9	89.3	10.7	(*)	(*)	64.3	35.7	7.1	92.9	0	100
	16 #	16	12	23	5	(*)	(*)	7	21	4	24	4	24
	%	57.1	42.9	82.1	17.9	(*)	(*)	25.0	75.0	14.3	85.7	14.3	85.7
	17 #	18	10	3	25	(*)	(*)	13	15	3	25	1	27
	%	64.3	35.7	10.7	89.3	(*)	(*)	46.4	53.6	10.7	89.3	3.6	96.4
	18 #	28	0	28	0	(*)	(*)	28	0	27	1	27	1
	%	100	0	100	0	(*)	(*)	100	0	96.4	3.6	96.4	3.6
	19 #	22	6	6	22	(*)	(*)	19	9	5	23	6	22
	%	78.6	21.4	21.4	78.6	(*)	(*)	67.9	32.1	17.9	82.1	21.4	78.6
	20 #	27	1	28	0	(*)	(*)	27	1	26	2	27	1
	%	96.4	3.6	100	0	(*)	(*)	96.4	3.6	92.9	7.1	96.4	3.6
	21 #	28	0	28	0	(*)	(*)	27	1	27	1	26	2
	%	100	0	100	0	(*)	(*)	96.4	3.6	96.4	3.6	92.9	7.1
	22 #	21	7	8	20	(*)	(*)	14	14	7	21	5	23
	%	75.0	25.0	28.6	71.4	(*)	(*)	50.0	50.0	25.0	75.0	17.9	82.1
	23 #	28	0	28	0	(*)	(*)	28	0	28	0	27	1
	%	100	0	100	0	(*)	(*)	100	0	100	0	96.4	3.6
	24 #	1	27	25	3	(*)	(*)	20	8	1	27	0	28
	%	3.6	96.4	89.3	10.7	(*)	(*)	71.4	28.6	3.6	96.4	0	100
	25 #	23	5	27	1	(*)	(*)	27	1	26	2	28	0
	%	82.1	17.9	96.4	3.6	(*)	(*)	96.4	3.6	92.9	7.1	100	0
	26 #	20	8	2	26	(*)	(*)	16	12	3	25	2	26
	%	71.4	28.6	7.1	92.9	(*)	(*)	57.1	42.9	10.7	89.3	7.1	92.9
	27 #	23	5	25	3	(*)	(*)	3	25	2	26	1	27
	%	82.1	17.9	89.3	10.7	(*)	(*)	10.7	89.3	7.1	92.9	3.6	96.4
	28 #	28	0	28	0	(*)	(*)	28	0	27	1	27	1
	%	100	0	100	0	(*)	(*)	100	0	96.4	3.6	96.4	3.6

DOMAIN	Test Item	Item Ratings										Item Matching	
		FORMAT		NUMBERS		WORDING		BEHAVIOR		OVERALL			NO
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	MATCH	MATCH
1	29 #	28	0	26	2	(*)	(*)	26	2	25	3	25	3
	%	100	0	92.9	7.1	(*)	(*)	92.9	7.1	89.3	10.7	89.3	10.7
	30 #	0	28	26	2	(*)	(*)	15	13	0	28	0	28
	%	0	100	92.9	7.1	(*)	(*)	53.6	46.4	0	100	0	100
	31 #	23	5	1	27	(*)	(*)	14	14	1	27	1	27
	%	82.1	17.9	3.6	96.6	(*)	(*)	50.0	50.0	3.6	96.4	3.6	96.4
	32 #	16	12	24	4	(*)	(*)	3	25	2	26	2	26
	%	57.1	42.9	85.7	14.3	(*)	(*)	10.7	89.3	7.1	92.9	7.1	92.9
	33 #	28	0	28	0	(*)	(*)	27	1	27	1	28	0
	%	100	0	100	0	(*)	(*)	96.4	3.6	96.4	3.6	100	0
1	34 #	26	2	26	2	(*)	(*)	25	3	25	3	27	1
	%	92.9	7.1	92.9	7.1	(*)	(*)	89.3	10.7	89.3	10.7	96.4	3.6
	35 #	27	1	27	1	(*)	(*)	27	1	27	1	28	0
	%	96.4	3.6	96.4	3.6	(*)	(*)	96.4	3.6	96.4	3.6	100	0
	36 #	3	25	24	4	(*)	(*)	11	17	2	26	2	26
	%	10.7	89.3	85.7	14.3	(*)	(*)	39.3	60.7	7.1	92.9	7.1	92.9
2	1 #	28	0	(*)	(*)	28	0	28	0	28	0	28	0
	%	100	0	(*)	(*)	100	0	100	0	100	0	100	0
	2 #	24	4	(*)	(*)	27	1	27	1	25	3	21	7
	%	85.7	14.3	(*)	(*)	96.4	3.6	96.4	3.6	89.3	10.7	75.0	25.0
	3 #	18	10	(*)	(*)	11	17	22	6	10	18	11	17
	%	64.3	35.7	(*)	(*)	39.3	60.7	78.6	21.4	35.7	64.3	39.3	60.7
	4 #	27	1	(*)	(*)	27	1	24	4	24	4	23	5
	%	96.4	3.6	(*)	(*)	96.4	3.6	85.7	14.3	85.7	14.3	82.1	17.9
	5 #	27	1	(*)	(*)	28	0	26	2	26	2	23	5
	%	96.4	3.6	(*)	(*)	100	0	92.9	7.1	92.9	7.1	82.1	17.9
	6 #	0	28	(*)	(*)	21	7	9	19	6	22	12	16
	%	0	100	(*)	(*)	75.0	25.0	32.1	67.9	21.4	78.6	42.9	57.1
	7 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0

DOMAIN	Test Item	Item Ratings										Item Matching	
		FORMAT		NUMBERS		WORDING		BEHAVIOR		OVERALL			NO
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	MATCH	MATCH
2	8 #	9	19	27	1	(*)	(*)	23	5	12	16	12	16
	%	32.1	67.9	96.4	3.6	(*)	(*)	82.1	17.9	42.9	57.1	42.9	57.1
	9 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	10 #	28	0	28	0	(*)	(*)	28	0	28	0	27	1
	%	100	0	100	0	(*)	(*)	100	0	100	0	96.4	3.6
	11 #	24	4	15	13	(*)	(*)	20	8	14	14	14	14
	%	85.7	14.3	53.6	46.4	(*)	(*)	71.4	28.6	50.0	50.0	50.0	50.0
	12 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	13 #	13	15	4	24	(*)	(*)	11	17	1	27	0	28
	%	46.4	53.6	14.3	85.7	(*)	(*)	39.3	60.7	3.6	96.4	0	100
	14 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
3	1 #	26	2	(*)	(*)	(*)	(*)	27	1	26	2	28	0
	%	92.9	7.1	(*)	(*)	(*)	(*)	96.4	3.6	92.9	7.1	100	0
	2 #	26	2	(*)	(*)	(*)	(*)	27	1	26	2	25	3
	%	92.9	7.1	(*)	(*)	(*)	(*)	96.4	3.6	92.9	7.1	89.3	10.7
	3 #	17	11	(*)	(*)	(*)	(*)	24	4	15	13	14	14
	%	60.7	39.3	(*)	(*)	(*)	(*)	85.7	14.3	53.6	46.4	50.0	50.0
	4 #	25	3	(*)	(*)	(*)	(*)	26	2	25	3	25	3
	%	89.3	10.7	(*)	(*)	(*)	(*)	92.9	7.1	89.3	10.7	89.3	10.7
	5 #	26	2	(*)	(*)	(*)	(*)	26	2	24	4	23	5
	%	92.9	7.1	(*)	(*)	(*)	(*)	92.9	7.1	85.7	14.3	82.1	17.9
	6 #	25	3	(*)	(*)	(*)	(*)	23	5	23	5	20	8
	%	89.3	10.7	(*)	(*)	(*)	(*)	82.1	17.9	82.1	17.9	71.4	28.6
	7 #	27	1	(*)	(*)	(*)	(*)	26	2	27	1	27	1
	%	96.4	3.6	(*)	(*)	(*)	(*)	92.9	7.1	96.4	3.6	96.4	3.6
	8 #	21	7	(*)	(*)	(*)	(*)	20	8	19	9	16	12
	%	75.0	25.0	(*)	(*)	(*)	(*)	71.4	28.6	67.9	32.1	57.1	42.9

DOMAIN	Test Item	Item Ratings										Item Matching	
		FORMAT		NUMBERS		WORDING		BEHAVIOR		OVERALL			NO
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	MATCH	MATCH
3	9 #	26	2	(*)	(*)	(*)	(*)	25	3	25	3	25	3
	%	92.9	7.1	(*)	(*)	(*)	(*)	89.3	10.7	89.3	10.7	89.3	10.7
	10 #	27	1	(*)	(*)	(*)	(*)	27	1	27	1	28	0
	%	96.4	3.6	(*)	(*)	(*)	(*)	96.4	3.6	96.4	3.6	100	0
	11 #	27	1	(*)	(*)	(*)	(*)	26	2	27	1	27	1
	%	96.4	3.6	(*)	(*)	(*)	(*)	92.9	7.1	96.4	3.6	96.4	3.6
	12 #	28	0	(*)	(*)	(*)	(*)	28	0	28	0	28	0
	%	100	0	(*)	(*)	(*)	(*)	100	0	100	0	100	0
	13 #	10	18	(*)	(*)	(*)	(*)	14	14	11	17	7	21
	%	35.7	64.3	(*)	(*)	(*)	(*)	50.0	50.0	39.3	60.7	25.0	75.0
4	1 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	2 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	3 #	27	1	27	1	(*)	(*)	27	1	27	1	27	1
	%	96.4	3.6	96.4	3.6	(*)	(*)	96.4	3.6	96.4	3.6	96.4	3.6
	4 #	28	0	28	0	(*)	(*)	28	0	28	0	27	1
	%	100	0	100	0	(*)	(*)	100	0	100	0	96.4	3.6
	5 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	6 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	7 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	8 #	11	17	2	26	(*)	(*)	7	21	1	27	1	27
	%	39.3	60.7	7.1	92.9	(*)	(*)	25.0	75.0	3.6	96.4	3.6	96.4
	9 #	28	0	28	0	(*)	(*)	28	0	27	1	28	0
	%	100	0	100	0	(*)	(*)	100	0	96.4	3.6	100	0
	10 #	28	0	28	0	(*)	(*)	28	0	28	0	27	1
	%	100	0	100	0	(*)	(*)	100	0	100	0	96.4	3.6

DOMAIN	Test Item		Item Ratings										Item Matching	
			FORMAT		NUMBERS		WORDING		BEHAVIOR		OVERALL			NO
			YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	MATCH	MATCH
4	11	#	28	0	28	0	(*)	(*)	28	0	28	0	28	0
		%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	12	#	28	0	28	0	(*)	(*)	28	0	28	0	28	0
		%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	13	#	28	0	27	1	(*)	(*)	27	1	27	1	27	1
		%	100	0	96.4	3.6	(*)	(*)	96.4	3.6	96.4	3.6	96.4	3.6
	14	#	4	24	21	7	(*)	(*)	11	17	4	24	6	22
		%	14.3	85.7	75.0	25.0	(*)	(*)	39.3	60.7	14.3	85.7	21.4	78.6
	15	#	28	0	28	0	(*)	(*)	28	0	28	0	27	1
		%	100	0	100	0	(*)	(*)	100	0	100	0	96.4	3.6
	16	#	28	0	28	0	(*)	(*)	28	0	28	0	28	0
		%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	17	#	28	0	28	0	(*)	(*)	28	0	28	0	27	1
		%	100	0	100	0	(*)	(*)	100	0	100	0	96.4	3.6
	18	#	28	0	28	0	(*)	(*)	28	0	28	0	28	0
		%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	19	#	28	0	28	0	(*)	(*)	28	0	28	0	26	2
		%	100	0	100	0	(*)	(*)	100	0	100	0	92.9	7.1
	20	#	28	0	28	0	(*)	(*)	28	0	28	0	27	1
		%	100	0	100	0	(*)	(*)	100	0	100	0	96.4	3.6
	21	#	28	0	28	0	(*)	(*)	28	0	28	0	26	2
		%	100	0	100	0	(*)	(*)	100	0	100	0	92.9	7.1
	22	#	28	0	28	0	(*)	(*)	28	0	28	0	26	2
		%	100	0	100	0	(*)	(*)	100	0	100	0	92.9	7.1
5	1	#	28	0	28	0	(*)	(*)	28	0	28	0	27	1
		%	100	0	100	0	(*)	(*)	100	0	100	0	96.4	3.6
	2	#	28	0	28	0	(*)	(*)	28	0	28	0	28	0
		%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	3	#	1	27	22	6	(*)	(*)	14	14	2	26	0	28
		%	3.6	96.4	78.6	21.4	(*)	(*)	50.0	50.0	7.1	92.9	0	100

DOMAIN	Test Item	Item Ratings										Item Matching	
		FORMAT		NUMBERS		WORDING		BEHAVIOR		OVERALL			
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	MATCH	NO MATCH
5	4 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	5 #	28	0	28	0	(*)	(*)	28	0	28	0	27	1
	%	100	0	100	0	(*)	(*)	100	0	100	0	96.4	3.6
	6 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	7 #	28	0	28	0	(*)	(*)	28	0	28	0	27	1
	%	100	0	100	0	(*)	(*)	100	0	100	0	96.4	3.6
	8 #	28	0	28	0	(*)	(*)	28	0	28	0	27	1
	%	100	0	100	0	(*)	(*)	100	0	100	0	96.4	3.6
	9 #	28	0	28	0	(*)	(*)	27	1	27	1	28	0
	%	100	0	100	0	(*)	(*)	96.4	3.6	96.4	3.6	100	0
	10 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	11 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
6	1 #	16	12	18	10	(*)	(*)	24	4	13	15	15	13
	%	57.1	42.9	64.3	35.7	(*)	(*)	85.7	14.3	46.4	53.6	53.6	46.4
	2 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	3 #	4	24	17	11	(*)	(*)	12	16	2	26	3	25
	%	14.3	85.7	60.7	39.3	(*)	(*)	42.9	57.1	7.1	92.9	10.7	89.3
	4 #	28	0	28	0	(*)	(*)	27	1	28	0	28	0
	%	100	0	100	0	(*)	(*)	96.4	3.6	100	0	100	0
	5 #	2	26	23	5	(*)	(*)	17	11	2	26	3	25
	%	7.1	92.9	82.1	17.9	(*)	(*)	60.7	39.3	7.1	92.9	10.7	89.3
	6 #	27	1	27	1	(*)	(*)	27	1	27	1	28	0
	%	96.4	3.6	96.4	3.6	(*)	(*)	96.4	3.6	96.4	3.6	100	0
	7 #	21	7	8	20	(*)	(*)	6	22	5	23	7	21
	%	75.0	25.0	28.6	71.4	(*)	(*)	21.4	78.6	17.9	82.1	25.0	75.0

DOMAIN	Test Item	Item Ratings										Item Matching	
		FORMAT		NUMBERS		WORDING		BEHAVIOR		OVERALL			
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	MATCH	NO MATCH
6	8 #	16	12	21	7	(*)	(*)	20	8	15	13	11	17
	%	57.1	42.9	75.0	25.0	(*)	(*)	71.4	28.6	53.6	46.4	39.3	60.7
	9 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	10 #	27	1	26	2	(*)	(*)	26	2	24	4	26	2
	%	96.4	3.6	92.9	7.1	(*)	(*)	92.9	7.1	85.7	14.3	92.9	7.1
	11 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	12 #	20	8	5	23	(*)	(*)	9	19	6	22	9	19
	%	71.4	28.6	17.9	82.1	(*)	(*)	32.1	67.9	21.4	78.6	32.1	67.9
	13 #	27	1	28	0	(*)	(*)	27	1	27	1	28	0
	%	96.4	3.6	100	0	(*)	(*)	96.4	3.6	96.4	3.6	100	0
	14 #	28	0	28	0	(*)	(*)	27	1	27	1	26	2
	%	100	0	100	0	(*)	(*)	96.4	3.6	96.4	3.6	92.9	7.1
	15 #	16	12	8	20	(*)	(*)	11	17	8	20	6	22
	%	57.1	42.9	28.6	71.4	(*)	(*)	39.3	60.7	28.6	71.4	21.4	78.6
	16 #	28	0	28	0	(*)	(*)	28	0	28	0	27	1
	%	100	0	100	0	(*)	(*)	100	0	100	0	96.4	3.6
	17 #	21	7	14	14	(*)	(*)	19	9	13	15	13	15
	%	75.0	25.0	50.0	50.0	(*)	(*)	67.9	32.1	46.4	53.6	46.4	53.6
	18 #	28	0	28	0	(*)	(*)	28	0	28	0	27	1
	%	100	0	100	0	(*)	(*)	100	0	100	0	96.4	3.6
7	1 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	2 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	3 #	28	0	27	1	(*)	(*)	28	0	27	1	28	0
	%	100	0	96.4	3.6	(*)	(*)	100	0	96.4	3.6	100	0
	4 #	0	28	25	3	(*)	(*)	18	10	1	27	3	25
	%	0	100	89.3	10.7	(*)	(*)	64.3	35.7	3.6	96.4	10.7	89.3

DOMAIN	Test Item	Item Ratings										Item Matching	
		FORMAT		NUMBERS		WORDING		BEHAVIOR		OVERALL			
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	MATCH	NO MATCH
7	5 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	6 #	18	10	25	3	(*)	(*)	7	21	5	23	3	25
	%	64.3	35.7	89.3	10.7	(*)	(*)	25.0	75.0	17.9	82.1	10.7	89.3
	7 #	27	1	24	4	(*)	(*)	26	1	23	5	25	3
	%	96.4	3.6	85.7	14.3	(*)	(*)	92.9	7.1	82.1	17.9	89.3	10.7
	8 #	27	1	28	0	(*)	(*)	28	0	27	1	28	0
	%	96.4	3.6	100	0	(*)	(*)	100	0	96.4	3.6	100	0
	9 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	10 #	28	0	27	1	(*)	(*)	28	0	27	1	28	0
	%	100	0	96.4	3.6	(*)	(*)	100	0	96.4	3.6	100	0
	11 #	27	1	27	1	(*)	(*)	27	1	27	1	28	0
	%	96.4	3.6	96.4	3.6	(*)	(*)	96.4	3.6	96.4	3.6	100	0
	12 #	2	26	20	8	(*)	(*)	14	14	3	25	1	27
	%	7.1	92.9	71.4	28.6	(*)	(*)	50.0	50.0	10.7	89.3	3.6	96.4
	13 #	27	1	25	3	(*)	(*)	26	2	25	3	28	0
	%	96.4	3.6	89.3	10.7	(*)	(*)	92.9	7.1	89.3	10.7	100	0
	14 #	22	6	9	19	(*)	(*)	20	8	11	17	12	16
	%	78.6	21.4	32.1	67.9	(*)	(*)	71.4	28.6	39.3	60.7	42.9	57.1
	15 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	16 #	27	1	28	0	(*)	(*)	28	0	28	0	28	0
	%	96.4	3.6	100	0	(*)	(*)	100	0	100	0	100	0
	17 #	28	0	28	0	(*)	(*)	28	0	28	0	27	1
	%	100	0	100	0	(*)	(*)	100	0	100	0	96.4	3.6
	18 #	17	11	12	16	(*)	(*)	3	25	1	27	5	23
	%	60.7	39.3	42.9	57.1	(*)	(*)	10.7	89.3	3.6	96.4	17.9	82.1
	19 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0

DOMAIN	Test Item	Item Ratings										Item Matching	
		FORMAT		NUMBERS		WORDING		BEHAVIOR		OVERALL			
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	MATCH	NO MATCH
7	20 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	21 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	22 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	23 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	24 #	20	8	6	22	(*)	(*)	11	17	7	21	8	20
	%	71.4	28.6	21.4	78.6	(*)	(*)	60.7	39.3	25.0	75.0	28.6	71.4
	25 #	28	0	23	5	(*)	(*)	24	4	22	6	23	5
	%	100	0	82.1	17.9	(*)	(*)	85.7	14.3	78.6	21.4	82.1	17.9
	26 #	28	0	23	5	(*)	(*)	26	2	24	4	24	4
	%	100	0	82.1	17.9	(*)	(*)	92.9	7.1	85.7	14.3	85.7	14.3
8	1 #	28	0	(*)	(*)	27	1	24	4	24	4	25	3
	%	100	0	(*)	(*)	96.4	3.6	85.7	14.3	85.7	14.3	89.3	10.7
	2 #	23	5	(*)	(*)	21	7	25	3	19	9	15	13
	%	82.1	17.9	(*)	(*)	75	25	89.3	10.7	67.9	32.1	53.6	46.4
	3 #	28	0	(*)	(*)	28	0	28	0	28	0	27	1
	%	100	0	(*)	(*)	100	0	100	0	100	0	96.4	3.6
	4 #	28	0	(*)	(*)	28	0	28	0	28	0	28	0
	%	100	0	(*)	(*)	100	0	100	0	100	0	100	0
	5 #	27	1	(*)	(*)	26	2	25	3	21	7	21	7
	%	96.4	3.6	(*)	(*)	92.9	7.1	89.3	10.7	75.0	25.0	75.0	25.0
	6 #	17	11	(*)	(*)	10	18	9	19	8	20	7	21
	%	60.7	39.3	(*)	(*)	35.7	64.3	32.1	67.9	28.6	71.4	25.0	75.0
	7 #	28	0	(*)	(*)	28	0	26	2	26	2	24	4
	%	100	0	(*)	(*)	100	0	92.9	7.1	92.9	7.1	85.7	14.3
	8 #	28	0	28	0	(*)	(*)	27	1	27	1	28	0
	%	100	0	100	0	(*)	(*)	96.4	3.6	96.4	3.6	100	0

DOMAIN	Test Item	Item Ratings										Item Matching	
		FORMAT		NUMBERS		WORDING		BEHAVIOR		OVERALL			
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	MATCH	NO MATCH
8	9 #	0	28	24	4	(*)	(*)	17	11	2	26	8	20
	%	0	100	85.7	14.3	(*)	(*)	60.7	39.3	7.1	92.9	28.6	71.4
	10 #	25	3	26	2	(*)	(*)	25	3	23	5	24	4
	%	89.3	10.7	92.9	7.1	(*)	(*)	89.3	10.7	82.1	17.9	85.7	14.3
	11 #	15	13	25	3	(*)	(*)	12	16	9	19	15	13
	%	53.6	46.4	89.3	10.7	(*)	(*)	42.9	57.1	32.1	67.9	53.6	46.4
	12 #	18	10	10	18	(*)	(*)	12	16	8	20	2	26
	%	64.3	35.7	35.7	64.3	(*)	(*)	42.9	57.1	28.6	71.4	7.1	92.9
	13 #	28	0	27	1	(*)	(*)	26	2	25	3	27	1
	%	100	0	96.4	3.6	(*)	(*)	92.9	7.1	89.3	10.7	96.4	3.6
	14 #	28	0	28	0	(*)	(*)	27	1	27	1	28	0
	%	100	0	100	0	(*)	(*)	96.4	3.6	96.4	3.6	100	0
9	1 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	2 #	28	0	28	0	(*)	(*)	28	0	28	0	27	1
	%	100	0	100	0	(*)	(*)	100	0	100	0	96.4	3.6
	3 #	25	3	27	1	(*)	(*)	26	2	24	4	23	5
	%	89.3	10.7	96.4	3.6	(*)	(*)	92.9	7.1	85.7	14.3	82.1	17.9
	4 #	27	1	28	0	(*)	(*)	28	0	27	1	25	3
	%	96.4	3.6	100	0	(*)	(*)	100	0	96.4	3.6	89.3	10.7
	5 #	27	1	28	0	(*)	(*)	28	0	26	2	25	3
	%	96.4	3.6	100	0	(*)	(*)	100	0	92.9	7.1	89.3	10.7
	6 #	14	14	22	6	(*)	(*)	14	14	10	18	10	18
	%	50	50	78.6	21.4	(*)	(*)	50	50	35.7	64.3	35.7	64.3
	7 #	24	4	26	2	(*)	(*)	24	4	24	4	23	5
	%	85.7	14.3	92.9	7.1	(*)	(*)	85.7	14.3	85.7	14.3	82.1	17.9
	8 #	27	1	27	1	(*)	(*)	27	1	27	1	26	2
	%	96.4	3.6	96.4	3.6	(*)	(*)	96.4	3.6	96.4	3.6	92.9	7.1
	9 #	27	1	28	0	(*)	(*)	27	1	26	2	19	9
	%	96.4	3.6	100	0	(*)	(*)	96.4	3.6	92.9	7.1	67.9	32.1

DOMAIN	Test Item	Item Ratings										Item Matching	
		FORMAT		NUMBERS		WORDING		BEHAVIOR		OVERALL			NO
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	MATCH	MATCH
9	10 #	26	2	28	0	(*)	(*)	28	0	26	2	22	6
	%	92.9	7.1	100	0	(*)	(*)	100	0	92.9	7.1	78.6	21.4
	11 #	26	2	27	1	(*)	(*)	27	1	25	3	19	9
	%	92.9	7.1	96.4	3.6	(*)	(*)	96.4	3.6	89.3	10.7	67.9	32.1
10	1 #	23	5	17	11	(*)	(*)	21	7	14	14	11	17
	%	82.1	17.9	60.7	39.3	(*)	(*)	75	25	50.0	50.0	39.3	60.7
	2 #	27	1	27	1	(*)	(*)	27	1	26	2	22	6
	%	96.4	3.6	96.4	3.6	(*)	(*)	96.4	3.6	92.9	7.1	78.6	21.4
	3 #	28	0	18	10	(*)	(*)	24	4	18	10	16	12
	%	100	0	64.3	7.9	(*)	(*)	85.7	14.3	64.3	35.7	57.1	42.9
	4 #	28	0	28	0	(*)	(*)	28	0	28	0	27	1
	%	100	0	100	0	(*)	(*)	100	0	100	0	96.4	3.6
	5 #	28	0	28	0	(*)	(*)	27	1	27	1	26	2
	%	100	0	100	0	(*)	(*)	96.4	3.6	96.4	3.6	92.9	7.1
	6 #	9	19	17	11	(*)	(*)	16	12	6	22	4	24
	%	32.1	67.9	60.7	39.3	(*)	(*)	57.1	42.9	21.4	78.6	14.3	85.7
	7 #	28	0	27	1	(*)	(*)	27	1	27	1	25	3
	%	100	0	96.4	3.6	(*)	(*)	96.4	3.6	96.4	3.6	89.3	10.7
	8 #	27	1	28	0	(*)	(*)	28	0	27	1	26	2
	%	96.4	3.6	100	0	(*)	(*)	100	0	96.4	3.6	92.9	7.1
	9 #	27	1	27	1	(*)	(*)	27	1	26	2	26	2
	%	96.4	3.6	96.4	3.6	(*)	(*)	96.4	3.6	92.9	7.1	92.9	7.1
	10 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	11 #	8	20	19	9	(*)	(*)	16	12	7	21	3	25
	%	28.6	71.4	67.9	32.1	(*)	(*)	57.1	42.9	25.0	75.0	10.7	89.3
	12 #	21	7	25	3	(*)	(*)	22	6	18	10	18	10
	%	75	25	89.3	10.7	(*)	(*)	78.6	21.4	64.3	35.7	64.3	35.7
	13 #	28	0	27	1	(*)	(*)	27	1	27	1	27	1
	%	100	0	96.4	3.6	(*)	(*)	96.4	3.6	96.4	3.6	96.4	3.6

DOMAIN	Test Item	Item Ratings										Item Matching	
		FORMAT		NUMBERS		WORDING		BEHAVIOR		OVERALL			
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	MATCH	NO MATCH
10	14 #	26	2	27	1	(*)	(*)	26	2	25	3	25	3
	%	92.9	7.1	96.4	3.6	(*)	(*)	92.9	7.1	89.3	10.7	89.3	10.7
	15 #	23	5	14	14	(*)	(*)	9	19	5	23	1	27
	%	82.1	17.9	50.0	50.0	(*)	(*)	32.1	67.9	17.9	82.1	3.6	96.4
	16 #	27	1	15	13	(*)	(*)	17	11	13	15	10	18
	%	96.4	3.6	53.6	46.4	(*)	(*)	60.7	39.3	46.4	53.6	35.7	64.3
	17 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	18 #	10	18	21	7	(*)	(*)	18	10	8	20	8	20
	%	35.7	64.3	75	25	(*)	(*)	64.3	35.7	28.6	71.4	28.6	71.4
	19 #	27	1	28	0	(*)	(*)	28	0	27	1	27	1
	%	96.4	3.6	100	0	(*)	(*)	100	0	96.4	3.6	96.4	3.6
	20 #	24	4	24	4	(*)	(*)	18	10	18	10	11	17
	%	85.7	14.3	85.7	14.3	(*)	(*)	64.3	35.7	64.3	35.7	39.3	60.7
	21 #	25	3	13	15	(*)	(*)	17	11	11	17	10	18
	%	89.3	10.7	46.4	53.6	(*)	(*)	60.7	39.3	39.3	60.7	35.7	64.3
	22 #	28	0	28	0	(*)	(*)	28	0	28	0	27	1
	%	100	0	100	0	(*)	(*)	100	0	100	0	96.4	3.6
	23 #	27	1	28	0	(*)	(*)	28	0	27	1	25	3
	%	96.4	3.6	100	0	(*)	(*)	100	0	96.4	3.6	89.3	10.7
	24 #	11	17	20	8	(*)	(*)	15	13	9	19	9	19
	%	39.3	60.7	71.4	28.6	(*)	(*)	53.6	46.4	32.1	67.9	32.1	67.9
	25 #	28	0	27	1	(*)	(*)	27	1	27	1	27	1
	%	100	0	96.4	3.6	(*)	(*)	96.4	3.6	96.4	3.6	96.4	3.6
	26 #	5	23	20	8	(*)	(*)	12	16	5	23	2	26
	%	17.9	82.1	71.4	28.6	(*)	(*)	42.9	57.1	17.9	82.1	7.1	92.9
	27 #	19	9	21	7	(*)	(*)	16	12	13	15	13	15
	%	67.9	32.1	75.0	25.0	(*)	(*)	57.1	42.9	46.4	53.6	46.4	53.6
	28 #	25	3	27	1	(*)	(*)	28	0	24	4	26	2
	%	89.3	10.7	96.4	3.6	(*)	(*)	100	0	85.7	14.3	92.9	7.1

DOMAIN	Test Item	Item Ratings										Item Matching	
		FORMAT		NUMBERS		WORDING		BEHAVIOR		OVERALL			NO
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	MATCH	MATCH
11	1 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	2 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	3 #	23	5	15	13	(*)	(*)	19	9	11	17	4	24
	%	82.1	17.9	53.6	46.4	(*)	(*)	67.9	32.1	39.3	60.7	14.3	85.7
	4 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	5 #	15	13	3	25	(*)	(*)	3	25	0	28	0	28
	%	53.6	46.4	10.7	89.3	(*)	(*)	10.7	89.3	0	100	0	100
	6 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	7 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	8 #	3	25	21	7	(*)	(*)	17	11	1	27	0	28
	%	10.7	89.3	75.0	25.0	(*)	(*)	60.7	39.3	3.6	96.4	0	100
	9 #	27	1	27	1	(*)	(*)	27	1	27	1	28	0
	%	96.4	3.6	96.4	3.6	(*)	(*)	96.4	3.6	96.4	3.6	100	0
	10 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	11 #	3	25	19	9	(*)	(*)	8	20	2	26	1	27
	%	10.7	89.3	67.9	32.1	(*)	(*)	28.6	71.4	7.1	92.9	3.6	96.4
	12 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0
	13 #	22	6	7	21	(*)	(*)	10	18	7	21	4	24
	%	78.6	21.4	25.0	75.0	(*)	(*)	35.7	64.3	25.0	75.0	14.3	85.7
	14 #	5	23	19	9	(*)	(*)	14	14	2	26	1	27
	%	17.9	82.1	67.9	32.1	(*)	(*)	50.0	50.0	7.1	92.9	3.6	96.4
	15 #	28	0	28	0	(*)	(*)	28	0	28	0	28	0
	%	100	0	100	0	(*)	(*)	100	0	100	0	100	0

DOMAIN	Test Item	Item Ratings										Item Matching	
		FORMAT		NUMBERS		WORDING		BEHAVIOR		OVERALL			
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	MATCH	NO MATCH
11	16 #	1	27	20	8	(*)	(*)	10	18	3	25	1	27
	%	3.6	96.4	71.4	28.6	(*)	(*)	35.7	64.3	10.7	89.3	3.6	96.4
	17 #	27	1	27	1	(*)	(*)	27	1	27	1	28	0
	%	96.4	3.6	96.4	3.6	(*)	(*)	96.4	3.6	96.4	3.6	100	0
	18 #	7	21	12	16	(*)	(*)	9	19	2	26	1	27
	%	25.0	75.0	42.9	57.1	(*)	(*)	32.1	67.9	7.1	92.9	3.6	96.4
12	1 #	28	0	28	0	28	0	28	0	28	0	28	0
	%	100	0	100	0	100	0	100	0	100	0	100	0
	2 #	28	0	28	0	28	0	28	0	28	0	28	0
	%	100	0	100	0	100	0	100	0	100	0	100	0
	3 #	25	3	27	1	9	19	22	6	9	19	5	23
	%	89.3	10.7	96.4	3.6	32.1	67.9	78.6	21.4	32.1	67.9	17.9	82.1
	4 #	28	0	28	0	27	1	28	0	28	0	28	0
	%	100	0	100	0	96.4	3.6	100	0	100	0	100	0
	5 #	28	0	28	0	28	0	28	0	27	1	23	5
	%	100	0	100	0	100	0	100	0	96.4	3.6	82.1	17.9
	6 #	28	0	28	0	27	1	27	1	27	1	27	1
	%	100	0	100	0	96.4	3.6	96.4	3.6	96.4	3.6	96.4	3.6
	7 #	25	3	25	3	24	4	15	13	14	14	14	14
	%	89.3	10.7	89.3	10.7	85.7	14.3	53.6	46.4	50.0	50.0	50.0	50.0
	8 #	28	0	28	0	25	3	26	2	25	3	23	5
	%	100	0	100	0	89.3	10.7	92.9	7.1	89.3	10.7	82.1	17.9
	9 #	27	1	28	0	28	0	26	2	27	1	24	4
	%	96.4	3.6	100	0	100	0	92.9	7.1	96.4	3.6	85.7	14.3
	10 #	27	1	28	0	27	1	27	1	26	2	25	3
	%	96.4	3.6	100	0	96.4	3.6	96.4	3.6	92.9	7.1	89.3	10.7

NOTE
(*) indicates that the dimension of judgment did not apply to this item.

DOMAIN	Test Item	Item Ratings										Item Matching	
		FORMAT		NUMBERS		WORDING		BEHAVIOR		OVERALL			
		YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	MATCH	NO MATCH
12	11 #	28	0	28	0	28	0	28	0	28	0	28	0
	%	100	0	100	0	100	0	100	0	100	0	100	0
	12 #	23	5	13	15	25	3	19	9	11	17	10	18
	%	82.1	17.9	46.4	53.6	89.3	10.7	67.9	32.1	39.3	60.7	35.7	64.3
	13 #	27	1	28	0	26	2	26	2	25	3	23	5
	%	96.4	3.6	100	0	92.9	7.1	92.9	7.1	89.3	10.7	82.1	17.9

NOTE
(*) indicates that the dimension of judgment did not apply to this item.

APPENDIX C.2

Number and Percent of Teachers Expressing Judgments on Whether Collections of Self-Rated Content Valid Items Adequately Cover the Scopes of Domains, by Group.

	(*) Group 1		(*) Group 2		All Teachers	
Domain Number	Rating of Representativeness		Rating of Representativeness		Rating of Representativeness	
	YES	NO	YES	NO	YES	NO
1 #	16	12	20	8	36	20
%	57	43	71	29	64	36
2 #	17	11	16	12	33	23
%	61	39	57	43	59	41
3 #	23	5	22	6	45	11
%	82	18	79	21	80	20
4 #	25	3	27	1	52	4
%	89	11	96	4	93	7
5 #	23	5	24	4	47	9
%	82	18	86	14	84	16
6 #	24	4	22	6	46	10
%	86	14	79	21	82	18

NOTE (*)

Group 1 matched test items of domains 1-6 and rated test items of domains 7-12.

Group 2 rated test items of domains 1-6 and matched test items of domains 7-12.

		(*) Group 1		(*) Group 2		All Teachers	
Domain Number		Rating of Representativeness		Rating of Representativeness		Rating of Representativeness	
		YES	NO	YES	NO	YES	NO
7	#	25	3	22	6	47	9
	%	89	11	79	21	84	16
8	#	14	14	15	13	29	27
	%	50	50	54	46	52	48
9	#	21	7	16	12	37	19
	%	75	25	57	43	66	34
10	#	15	13	17	11	32	24
	%	54	46	61	39	57	43
11	#	19	9	20	8	39	17
	%	68	32	71	29	70	30
12	#	17	11	19	9	36	20
	%	61	39	68	32	64	36

NOTE (*)

Group 1 matched test items of domains 1-6 and rated test items of domains 7-12.

Group 2 rated test items of domains 1-6 and matched test items of domains 7-12.

APPENDIX C.3

Number and Proportion of Teachers Expressing Correct Judgments on Whether Test Items are Measures of the Domains

DOMAIN	Test Item	N1	P1	N2	P2	N3	P3
1	1	27	0.964	27	0.964	28	1.000
	2	26	0.929	23	0.821	28	1.000
	3	26	0.929	26	0.929	27	0.964
	4	18	0.643	2	0.071	22	0.786
	5	12	0.429	7	0.250	9	0.321
	6	24	0.857	25	0.893	25	0.893
	7	22	0.786	16	0.571	23	0.821
	8	28	1.000	28	1.000	28	1.000
	9	28	1.000	28	1.000	27	0.964
	10	28	1.000	28	1.000	28	1.000
	11	28	1.000	28	1.000	28	1.000
	12	28	1.000	28	1.000	28	1.000
	13	28	1.000	28	1.000	28	1.000

N1 and P1 are the respective number and proportion of teachers expressing correct judgments on whether, overall, an item is a measure of a domain.

N2 and P2 are the respective number and proportion of teachers indicating correct judgments on whether an item satisfies the domain specifications with regard to format, wording, number and behavior.

N3 and P3 are the respective number and proportion of teachers indicating correct judgments on whether an item corresponds to a domain definition.

DOMAIN	Test Item	N1	P1	N2	P2	N3	P3
1	14	25	0.893	25	0.893	26	0.929
	15	26	0.929	23	0.821	28	1.000
	16	24	0.857	15	0.536	24	0.857
	17	25	0.893	15	0.536	27	0.964
	18	27	0.964	27	0.964	27	0.964
	19	23	0.821	16	0.571	22	0.786
	20	26	0.929	26	0.929	27	0.964
	21	27	0.964	27	0.964	26	0.929
	22	21	0.750	2	0.071	23	0.821
	23	28	1.000	28	1.000	27	0.964
	24	27	0.964	24	0.857	28	1.000
	25	26	0.929	23	0.821	28	1.000
	26	25	0.893	16	0.571	26	0.929
	27	26	0.929	23	0.821	27	0.964
	28	27	0.964	28	1.000	27	0.964
	29	25	0.893	26	0.929	25	0.893
	30	28	1.000	15	0.536	28	1.000
	31	27	0.964	14	0.500	27	0.964
	32	26	0.929	16	0.571	26	0.929
	33	27	0.964	27	0.964	28	1.000
	34	25	0.893	25	0.893	27	0.964
	35	27	0.964	27	0.964	28	1.000
	36	26	0.929	11	0.393	26	0.929

DOMAIN	Test Item	N1	P1	N2	P2	N3	P3
2	1	28	1.000	28	1.000	28	1.000
	2	25	0.893	22	0.786	21	0.750
	3	18	0.643	7	0.250	17	0.607
	4	24	0.857	24	0.857	23	0.821
	5	26	0.929	26	0.929	23	0.821
	6	22	0.786	15	0.536	16	0.571
	7	28	1.000	28	1.000	28	1.000
	8	16	0.571	18	0.643	16	0.571
	9	28	1.000	28	1.000	28	1.000
	10	28	1.000	28	1.000	27	0.964
	11	14	0.500	9	0.321	14	0.500
	12	28	1.000	28	1.000	28	1.000
	13	27	0.964	14	0.500	28	1.000
	14	28	1.000	28	1.000	28	1.000
3	1	26	0.929	26	0.929	28	1.000
	2	26	0.929	26	0.929	25	0.893
	3	13	0.464	11	0.393	14	0.500
	4	25	0.893	25	0.893	25	0.893
	5	24	0.857	25	0.893	23	0.821
	6	23	0.821	23	0.821	20	0.714
	7	27	0.964	26	0.929	27	0.964
	8	9	0.321	5	0.179	12	0.429
	9	25	0.893	25	0.893	25	0.893

DOMAIN	Test Item	N1	P1	N2	P2	N3	P3
	10	27	0.964	27	0.964	28	1.000
	11	27	0.964	26	0.929	27	0.964
	12	28	1.000	28	1.000	28	1.000
	13	17	0.607	14	0.500	21	0.750
4	1	28	1.000	28	1.000	28	1.000
	2	28	1.000	28	1.000	28	1.000
	3	27	0.964	27	0.964	27	0.964
	4	28	1.000	28	1.000	27	0.964
	5	28	1.000	28	1.000	28	1.000
	6	28	1.000	28	1.000	28	1.000
	7	28	1.000	28	1.000	28	1.000
	8	27	0.964	11	0.393	27	0.964
	9	27	0.964	28	1.000	28	1.000
	10	28	1.000	28	1.000	27	0.964
	11	28	1.000	28	1.000	28	1.000
	12	28	1.000	28	1.000	28	1.000
	13	27	0.964	27	0.964	27	0.964
	14	24	0.857	17	0.607	22	0.786
	15	28	1.000	28	1.000	27	0.964
	16	28	1.000	28	1.000	28	1.000
	17	28	1.000	28	1.000	27	0.964
	18	28	1.000	28	1.000	28	1.000
	19	28	1.000	28	1.000	26	0.929

DOMAIN	Test Item	N1	P1	N2	P2	N3	P3
4	20	28	1.000	28	1.000	27	0.964
	21	28	1.000	28	1.000	26	0.929
	22	28	1.000	28	1.000	26	0.929
5	1	28	1.000	28	1.000	27	0.964
	2	28	1.000	28	1.000	28	1.000
	3	26	0.929	14	0.500	28	1.000
	4	28	1.000	28	1.000	28	1.000
	5	28	1.000	28	1.000	27	0.964
	6	28	1.000	28	1.000	28	1.000
	7	28	1.000	28	1.000	27	0.964
	8	28	1.000	28	1.000	27	0.964
	9	27	0.964	27	0.964	28	1.000
	10	28	1.000	28	1.000	28	1.000
	11	28	1.000	28	1.000	28	1.000
6	1	15	0.536	12	0.429	13	0.464
	2	28	1.000	28	1.000	28	1.000
	3	26	0.929	15	0.536	25	0.893
	4	28	1.000	27	0.964	28	1.000
	5	26	0.929	17	0.607	25	0.893
	6	27	0.964	27	0.964	28	1.000
	7	23	0.821	15	0.536	21	0.750
	8	13	0.464	5	0.179	17	0.607
	9	28	1.000	28	1.000	28	1.000

DOMAIN	Test Item	N1	P1	N2	P2	N3	P3
6	10	24	0.857	25	0.893	24	0.929
	11	28	1.000	28	1.000	28	1.000
	12	22	0.786	15	0.536	19	0.679
	13	27	0.964	27	0.964	28	1.000
	14	27	0.964	27	0.964	26	0.929
	15	20	0.714	10	0.357	22	0.786
	16	28	1.000	28	1.000	27	0.964
	17	15	0.536	9	0.321	15	0.536
	18	28	1.000	28	1.000	27	0.964
7	1	28	1.000	28	1.000	28	1.000
	2	28	1.000	28	1.000	28	1.000
	3	27	0.964	27	0.964	28	1.000
	4	27	0.964	23	0.821	25	0.893
	5	28	1.000	28	1.000	28	1.000
	6	23	0.821	18	0.643	25	0.893
	7	23	0.821	22	0.786	25	0.893
	8	27	0.964	27	0.964	28	1.000
	9	28	1.000	28	1.000	28	1.000
	10	27	0.964	27	0.964	28	1.000
	11	27	0.964	27	0.964	28	1.000
	12	25	0.893	18	0.643	27	0.964
	13	25	0.893	24	0.857	28	1.000
	14	17	0.607	14	0.500	16	0.571

DOMAIN	Test Item	N1	P1	N2	P2	N3	P3
	15	28	1.000	28	1.000	28	1.000
	16	28	1.000	27	0.964	28	1.000
	17	28	1.000	28	0.964	27	0.964
	18	27	0.964	11	0.393	23	0.821
	19	28	1.000	28	1.000	28	1.000
	20	28	1.000	28	1.000	28	1.000
	21	28	1.000	28	1.000	28	1.000
	22	28	1.000	28	1.000	28	1.000
	23	28	1.000	28	1.000	28	1.000
	24	21	0.750	17	0.607	20	0.714
	25	22	0.786	21	0.750	23	0.821
	26	24	0.857	23	0.821	24	0.857
8	1	24	0.857	23	0.821	25	0.893
	2	9	0.321	3	0.107	13	0.464
	3	28	1.000	28	1.000	27	0.964
	4	28	1.000	28	1.000	28	1.000
	5	21	0.750	23	0.821	21	0.750
	6	20	0.714	11	0.393	21	0.750
	7	26	0.929	26	0.929	24	0.857
	8	27	0.964	27	0.964	28	1.000
	9	26	0.929	23	0.821	20	0.714
	10	23	0.821	23	0.821	24	0.857
	11	19	0.679	12	0.429	13	0.464

DOMAIN	Test Item	N1	P1	N2	P2	N3	P3
8	12	20	0.714	11	0.393	26	0.929
	13	25	0.893	25	0.893	27	0.964
	14	27	0.964	27	0.964	28	1.000
9	1	28	1.000	28	1.000	28	1.000
	2	28	1.000	28	1.000	27	0.964
	3	24	0.857	24	0.857	23	0.821
	4	27	0.964	27	0.964	25	0.893
	5	26	0.929	27	0.964	25	0.893
	6	18	0.643	14	0.500	18	0.643
	7	24	0.857	24	0.857	23	0.821
	8	27	0.964	27	0.964	26	0.929
	9	26	0.929	26	0.929	19	0.679
	10	26	0.929	26	0.929	22	0.786
	11	25	0.893	25	0.893	19	0.679
10	1	14	0.500	8	0.286	17	0.607
	2	26	0.929	26	0.929	22	0.786
	3	10	0.357	9	0.321	12	0.429
	4	28	1.000	28	1.000	27	0.964
	5	27	0.964	27	0.964	26	0.929
	6	22	0.786	11	0.393	24	0.857
	7	27	0.964	27	0.964	25	0.893
	8	27	0.964	27	0.964	26	0.929
10	9	26	0.929	26	0.929	26	0.929

DOMAIN	Test Item	N1	P1	N2	P2	N3	P3
10	10	28	1.000	28	1.000	28	1.000
	11	21	0.750	11	0.393	25	0.893
	12	18	0.643	16	0.571	18	0.643
	13	27	0.964	27	0.964	27	0.964
	14	25	0.893	24	0.857	25	0.893
	15	23	0.821	14	0.500	27	0.964
	16	15	0.536	11	0.393	18	0.643
	17	28	1.000	28	1.000	28	1.000
	18	20	0.714	11	0.393	20	0.714
	19	27	0.964	27	0.964	27	0.964
	20	18	0.643	18	0.643	11	0.393
	21	17	0.607	9	0.321	18	0.643
	22	28	1.000	28	1.000	27	0.964
	23	27	0.967	27	0.964	25	0.893
	24	19	0.679	6	0.214	19	0.679
	25	27	0.964	27	0.964	27	0.964
	26	23	0.821	13	0.464	26	0.929
	27	13	0.464	14	0.500	13	0.464
	28	24	0.857	25	0.893	26	0.929
11	1	28	1.000	28	1.000	28	1.000
	2	28	1.000	28	1.000	28	1.000
	3	17	0.607	11	0.393	24	0.857
	4	28	1.000	28	1.000	28	1.000

DOMAIN	Test Item	N1	P1	N2	P2	N3	P3
11	5	28	1.000	25	0.893	28	1.000
	6	28	1.000	28	1.000	28	1.000
	7	28	1.000	28	1.000	28	1.000
	8	27	0.964	15	0.536	28	1.000
	9	27	0.964	27	0.964	28	1.000
	10	28	1.000	28	1.000	28	1.000
	11	26	0.929	16	0.571	27	0.964
	12	28	1.000	28	1.000	28	1.000
	13	21	0.750	16	0.571	24	0.857
	14	26	0.929	13	0.464	27	0.964
	15	28	1.000	28	1.000	28	1.000
	16	25	0.893	15	0.536	27	0.964
	17	27	0.964	27	0.964	28	1.000
	18	26	0.929	11	0.393	27	0.964
12	1	28	1.000	28	1.000	28	1.000
	2	28	1.000	28	1.000	28	1.000
	3	19	0.679	13	0.464	23	0.821
	4	28	1.000	27	0.964	28	1.000
	5	27	0.964	28	1.000	23	0.821
	6	27	0.964	27	0.964	27	0.964
	7	14	0.500	10	0.357	14	0.500
	8	25	0.893	25	0.893	23	0.821
	9	27	0.964	26	0.929	24	0.857

DOMAIN	Test Item	N1	P1	N2	P2	N3	P3
12	10	26	0.929	26	0.929	25	0.893
	11	28	1.000	28	1.000	28	1.000
	12	17	0.607	10	0.357	18	0.643
	13	25	0.893	25	0.893	23	0.821

N1 and P1 are the respective number and proportion of teachers expressing correct judgments on whether, overall, an item is a measure of a domain.

N2 and P2 are the respective number and proportion of teachers indicating correct judgments on whether an item satisfies the domain specifications with regard to format, wording, number and behavior.

N3 and P3 are the respective number and proportion of teachers indicating correct judgments on whether an item corresponds to a domain definition.

APPENDIX C.4

Number and Proportion of Teachers Expressing Correct Judgments on Whether Collections of Self-Rated Content Valid Items Adequately Cover the Scopes of Domains, by Group

Domain Number	(*) Group 1		(*) Group 2		All Teachers	
	Rating of Representativeness		Rating of Representativeness		Rating of Representativeness	
	N1	P1	N2	P2	N	P
1	16	0.57	20	0.71	36	0.64
2	11	0.39	12	0.43	23	0.41
3	23	0.82	22	0.79	45	0.80
4	25	0.89	27	0.96	52	0.93
5	23	0.82	24	0.86	47	0.84
6	24	0.86	22	0.79	46	0.82
7	25	0.89	22	0.79	47	0.84
8	14	0.50	13	0.46	27	0.48
9	21	0.75	16	0.57	37	0.66
10	13	0.46	11	0.39	24	0.43
11	19	0.68	20	0.71	39	0.70
12	17	0.61	19	0.68	36	0.64

NOTE (*)

Group 1 matched test items of domains 1-6 and rated test items of domains 7-12.

Group 2 rated test items of domains 1-6 and matched test items of domains 7-12.