DOCUMENT RESUME

ED 211 601 TM 820 051

AUTHOR Polin, Linda; Baker, Eva L.

TITLE Qualitative Analysis of Test Item Attributes for Domain Referenced Content Validity Judgments. Studies in Measurement and Methodology, Work Unit 1: Design and Use of Tests.

INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.

SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.

PUB DATE May 79

GRANT OB-NIE-G-78-0213

NOTE 86p.: Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April, 1979).

EDRS PRICE MF01/PC04 Plus Postage.

DESCRIPTORS Achievement Tests; *Criterion Referenced Tests; Elementary Secondary Education; Evaluation Methods; Item Analysis; *Test Construction; *Test Items; *Test Validity

IDENTIFIERS *Domain Specifications; Inter Rater Reliability; *Item Review Scale; Qualitative Analysis; Test Curriculum Overlap

ABSTRACT

This paper presents the interim results of a set of studies undertaken to develop a much needed methodology for establishing content validity in domain-referenced achievement tests. The study results are presented in the context of the larger issue of the improvement of test design. School teachers, administrators and graduate students were trained in using the Item Review Scale (IRS), and then served as item raters. The IRS provides structure for judgments of congruence between items and domain specifications. Raters were given a domain specification (elementary level geology) and eight related items to be rated for validity in terms of match. The items were systematically flawed along one or more of the dimensions of the IRS. Results demonstrated high inter-rater reliability on each dimension and the instrument as a whole. (Author/GK)

Deliverable: May 1979

Studies in Measurement and Methodology

QUALITATIVE ANALYSIS OF TEST ITEM ATTRIBUTES FOR DOMAIN REFERENCED CONTENT VALIDITY JUDGMENTS

Linda Polin and Eva L. Baker

Work Unit 1: Design and Use of Tests

Eva L. Baker, Project Director

Grant Number
OB-NIE-G-78-0213

Center for the Study of Evaluation
Graduate School of Education
University of California - Los Angeles

The authors wish to thank Jennie Pao Yeh and Shelly Niwa for their assistance in the data analyses and Vicki Porton and Osher Doctorow for their help in the early development of the Item Rating Scale.

The substance of the study reported herein was presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

The work upon which this publication is based was performed pursuant to a grant with the National Institute of Education, Department of Health, Education and Welfare. Points of view or opinions stated do not necessarily represent official NIE position or policy.

QUALITATIVE ANALYSIS OF TEST ITEM ATTRIBUTES FOR DOMAIN-REFERENCED CONTENT VALIDITY JUDGMENTS

Linda G. Polin
Center for the Study of Evaluation
UCLA
145 Moore Hall
Los Angeles, California 90024

Linda G. Polin and Eva L. Baker
Center for the Study of Evaluation
UCLA
Los Angeles, California 90024

This paper presents the interim results of a set of studies undertaken to develop a much needed methodology for establishing content validity in domain-referenced achievement tests. The study results are presented here in the context of the larger issue of the improvement of test design. One of the most neglected elements in designing tests is that of publicness; that is, the extent to which test specification are understandable and useable by all interested parties, including teachers, parents, taxpayers and even students themselves. Specifying test content and design accesses an entire set of content validity related issues, such as test bias and instructional sensitivity.

An underlying assumption of domain-referenced testing is that by limiting and defining a class of behaviors, skills, information i.e., a domain, we can create a set of rules for generating sets of test items that will accurately reflect the content of that domain. The power of such tests, or, the "goodness" of the test and test-biased decisions depends upon the match between the domain and the generated items. That is, is student test performance an accurate indicator of student capabilities vis a vis a particular domain of knowledge; is the test content valid?

At present there are no workable methodologies for determining content validity, expecially under the criterion of publicness as described above. Suion (1977) in psychological measurement and Popham (1975) in educational evaluation, describe the methodology for determining content validity using the term, "judgment," referring to the individual's thoughtful but subjective decision on the quality of fit between domain and items. Other methodologies attempting to systematize and objectify the judgment process have been explored and found lacking. In a separate project, one of the authors participated in item-objective "congruence methods of sorting, matching and even generating objectives from given items (Walker, Trithart and Polin, 1976).

The basic concept of making validity determiniations based upon "judged" degree of match between each item and the domain specification remains appealing largely because it is comprehensible to a wide range of persons interested in testing, test development and test selection. In approaching the content validity judgment issue anew, we have borrowed from the field of mathematics and set theory, the concept of "fuzzy sets." In fuzzy set theory, it is possible to specify strength of membership, or degree of inclusion for a set member along a range of decimal values from zero to one.

In our first study, this notion of degree of belongingness and the accompanying fuzzy set rating scale were applied to rating the match between test items and a domain specification. A group of raters experienced in test evaluation and construction made judgments of the match between items drawn from a variety of tests on the same subject and a domain specification from one of those tests. Although raters had no problem using the concept of fuzzy sets in making their judgments, there were marked inconsistencies in item and domain attributes upon which raters based their judgments. To remedy this problem and continue research into the application of fuzzy set judgments to content validity, we developed the Item Review Scale (IRS) to provide structure for judgments of congruence between items and domain specifications.

In drafting the IRS we have identified components or dimensions of items and provided an instrument encompassing these dimensions in a rating form conducive to making the judgments described above. This protocol for judging itmes distinguishes the following dimensions of items: Domain description, Content limits, Distractor domain or Response criteria, Format, Directions, Sample item (these six components identified as elements of our domain specification model); Linguistic complexity, and Thinking complexity. Linguistic complexity guides raters toward consideration of semantic, syntactic features of items affecting their difficulty and consequently, their match with the intentions of the domain specification. Thinking complexity guides raters toward consideration of the sophistication of cognitive processes involved in correctly responding to the item.

The study described in this paper was conducted with school teachers, administrators and graduate students as item raters. These raters were trained in using the IRS and were then given a domain specification (elementary level geology) and eight related items to be rated for validity in terms of match. The items used in the field study were systematically flawed along one or more of the eight dimensions of the IRS.

Results of the study demonstrated high inter-rater reliability on each dimension and the instrument as a whole. Interdimension correlations, by item, suggested that rating responses were sensitive to the item being rated. The three dimensions weighted in our instrument because of their hypothesized relative importance to the overall item quality (Content limits, Distractor domain or Response criteria and Thinking complexity), were indeed the only three dimensions correlated strongly with the overall item score for all items.

The results of this work are relevant for school personnel and curriculum developers as well as test developers. As districts and schools increasingly rely upon locally developed tests or objectives-based tests drawn from item banks, this method for making content validity judgments will assist in the development or selection of an instrument sensitive to instruction or curricula.

QUALITATIVE ANALYSIS OF TEST ITEM ATTRIBUTES FOR DOMAIN-REFERENCED CONTENT VALIDITY JUDGMENTS

Linda Polin and Eva L. Baker
Center for the Study of Evaluation
University of California, Los Angeles

A paper presented at the annual meeting
of the American Educational Research Association
April, 1979, San Francisco

Grant Number
OB-NIE-G-78-0213

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

QUALITATIVE ANALYSIS OF TEST ITEM ATTRIBUTES FOR DOMAIN-REFERENCED CONTENT VALIDITY JUDGMENTS

Linda G. Polin and Eva L. Baker
Center for the Study of Evaluation
UCLA
Los Angeles, California 90024

Linda G. Polin
Center for the Study of Eval
UCLA
145 Moore Hall
Los Angeles, California 90024

This paper presents the interim results of a set of studies undertaken to develop a much needed methodology for establishing content validity in domain-referenced achievement tests. The study results are presented here in the context of the larger issue of the improvement of test design. One of the most neglected elements in designing tests is that of publicness; that is, the extent to which test specification are understandable and useable by all interested parties, including teachers, parents, taxpayers and even students themselves. Specifying test content and design accesses an entire set of content validity related issues, such as test bias and instructional sensitivity.

An underlying assumption of domain-referenced testing is that by limiting and defining a class of behaviors, skills, information i.e., a domain, we can create a set of rules for generating sets of test items that will accurately reflect the content of that domain. The power of such tests, or, the "goodness" of the test and test-biased decisions depends upon the match between the domain and the generated items. That is, is student test performance an accurate indicator of student capabilities vis a vis a particular domain of knowledge; is the test content valid?

At present there are no workable methodologies for determining content validity, expecially under the criterion of publicness as described above. Guion (1977) in psychological measurement and Popham (1975) in educational evaluation, describe the methodology for determining content validity using the term, "judgment," referring to the individual's thoughtful but subjective decision on the quality of fit between domain and items. Other methodologies attempting to systematize and objectify the judgment process have been explored and found lacking. In a separate project, one of the authors participated in item-objective "congruence methods of sorting, matching and even generating objectives from given items (Walker, Trithart and Polin, 1976).

The basic concept of making validity determiniations based upon "judged" degree of match between each item and the domain specification remains appealing largely because it is comprehensible to a wide range of persons interested in testing, test development and test selection. In approaching the content validity judgment issue anew, we have borrowed from the field of mathematics and set theory, the concept of "fuzzy sets." In fuzzy set theory, it is possible to specify strength of membership, or degree of inclusion for a set member along a range of decimal values from zero to one.

In our first study, this notion of degree of belongingness and the accompanying fuzzy set rating scale were applied to rating the match between test items and a domain specification. A group of raters experienced in test evaluation and construction made judgments of the match between items drawn from a variety of tests on the same subject and a domain specification from one of those tests. Although raters had no problem using the concept of fuzzy sets in making their judgments, there were marked inconsistencies in item and domain attributes upon which raters based their judgments. To remedy this problem and continue research into the application of fuzzy set judgments to content validity, we developed the Item Review Scale (IRS) to provide structure for judgments of congruence between items and domain specifications.

In drafting the IRS we have identified components or dimensions of items and provided an instrument encompassing these dimensions in a rating form conducive to making the judgments described above. This protocol for judging itmes distinguishes the following dimensions of items: Domain description, Content limits, Distractor domain or Response criteria, Format, Directions, Sample item (these six components identified as elements of our domain specification model); Linguistic complexity, and Thinking complexity. Linguistic complexity guides raters toward consideration of semantic, syntactic features of items affecting their difficulty and consequently, their match with the intentions of the domain specification. Thinking complexity guides raters toward consideration of the sophistication of cognitive processes involved in correctly responding to the item.

The study described in this paper was conducted with school teachers, administrators and graduate students as item raters. These raters were trained in using the IRS and were then given a domain specification (elementary level geology) and eight related items to be rated for validity in terms of match. The items used in the field study were systematically flawed along one or more of the eight dimensions of the IRS.

Results of the study demonstrated high inter-rater reliability on each dimension and the instrument as a whole. Interdimension correlations, by item, suggested that rating responses were sensitive to the item being rated. The three dimensions weighted in our instrument because of their hypothesized relative importance to the overall item quality (Content limits, Distractor domain or Response criteria and Thinking complexity), were indeed the only three dimensions correlated strongly with the overall item score for all items.

The results of this work are relevant for school personnel and curriculum developers as well as test developers. As districts and schools increasingly rely upon locally developed tests or objectives-based tests drawn from item banks, this method for making content validity judgments will assist in the development or selection of an instrument sensitive to instruction or curricula.

The authors wish to thank Jennie Pao Yeh and Shelly Niwa for their assistance in the data analyses and Vicki Porton and Osher Doctorow for their help in the early development of the Item Rating Scale.

The substance of the study reported herein was presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

The work upon which this publication is based was performed pursuant to a grant with the National Institute of Education, Department of Health, Education and Welfare. Points of view or opinions stated do not necessarily represent official NIE position or policy.

TABLE OF CONTENTS

INTRODUCTION

Problem. In discussions of the quality of tests available to assess students' achievement, assumptions about item quality are clearly made. One of the presumably defining characteristics of emerging forms of testing technology is the desired homogeneity of the test items sampling a particular, well-described domain of competency. Domain-referenced achievement tests, for example, are to include both the careful description of the design principles which underpin a particular measure, and a set of test items which represent to some degree, the competency that is to be exhibited.

The development and description of procedures for preparing test specifications have been explored and exhorted by many (Hively, Patterson and Page, 1968; 1974; Baker, 1974; Millman, 1974; Popham, 1978). The utility of these specifications has been claimed in areas relating to: 1) the control of the test item writers' behaviors, i.e., that they can produce items which covary in appropriate ways; 2) the preparation of appropriate instructional plans clearly linked to the features incorporated in the test specifications; 3) the need for public access to the nature and design of tests used in making potentially important decisions affecting students; and, 4) the exploration of cognitive processing requirements of competent performance in various areas.

While no one would claim the components of this technology are in humming order, among the weaker segments is the manner in which item homogeneity is addressed. In its most common rendition, item homogeneity is attacked empirically by exploring the correlations and clusters of items thought to assess the same competency area. While sucn procedures are clearly useful in more general trait measurement, they lose cogency when one juxtaposes them with the idea that differential instruction should influence the patterns of performance that are exhibited on such items. When tests are designed, as at least part of their charge is to be responsive to instructional attempts, it is clear that making assumptions about the immutable quality of the item characteristics in the face of random instructional events is not possible. The tautology, although still present, is somewhat more tractable in such an experimental setting.

There is still, obviously, a prior question: before we have the confidence to move newly developed items into an empirical testing mode, what sense can we make of the relationship between what test writers have wrought and the original test specifications? How we can capture and describe human judgment of quality, judgment which would clearly precede empirical work in this area? Unfortunately, the responses of our field have been scattered and relatively impotent in this matter. We do have various recommendations for making judgments about the match between test items and specifications. In general, these procedures involve categorical decisions, often on-off choices about the item's fit with the specifications. Sometimes, as a refinement, a Likert-type scale is used, e.g., "certainly matches," "certainly doesn't match." However, careful attention has not been given to detailing the components of such judgments, that is, to determining which features are critical to check; nor has attention been given to the potential use of such information in subsequent empirical work.

Our interest, therefore, was to explore the development of a technique which would allow us to address, in rigorous terms, the nature of the judged differences in test item features as they relate to a set of standard test specifications. We believed that a procedure, if it could be sufficiently well developed, would be useful not only for the development and quality assurance of test items, but also for assessment of already available items as they might or might not conform to a user's specifications. Furthermore, we felt that our data might influence the basis upon which the test specifications themselves are created, and provide an empirical basis to refine the features and detail of such specifications.

Theoretical background. It was clear to us that available techniques did not adequately address the issues of concern to us, based as they were upon statistical methods of describing item characteristics. When we reconsidered the underlying notion of domain-referenced testing, it was obvious that the domain represented a "set" in the mathematical sense of the term, and that the test specifications were verbal definitions of the set's boundaries. It seemed natural for us then to explore what mathematicians studying in the area of sets and groups had developed that might help us consider the problem of describing an item's belongingness to a particular well-defined universe (domain). At the heart of our problem was the belief that items simply didn't either belong or not belong to a set of specifications. They coher in degrees.

Work in set theory described by Arbib (1968) and Doctorow (1977) provided some promising lines of inquiry. In the first place, in oversimplified form, rules for membership in sets and groups were described. Special subclasses of both sets and groups were described for instances where membership was relative rather than absolute. Because of the arithmetic involved, the subclass of interest for our work is called "fuzzy sets," a cuddly term which, in fact, denotes that instances can belong in matters of degree; that these degrees can be formulated in terms of probability of membership, along a 0 to 1 continuum.

In exploring the utility of such a theoretical notion, we first decided to see whether people could make such probability estimates. We suggested that, when confronted with an item and a set of specifications, people "bet" some proportion of 100 units on the membership of the item to the class (specification of the domain). Our first pilot study involved three specifications, ten items in the fields of mathematics and reading, and were judged by eight graduate students and staff members at the Center for the Study of Evaluation. In general, the responses we received were difficult to interpret. First, it became clear that features linking domain specifications and test items, e.g., distractor set, needed more complete definition. Secondly, people were inferring dimensions and clearly responding out of a wide range of experience which resulted in low inter-rater judgments on particular items. We believed that this matter might be addressed not only by including structure and difinition for judging features or dimensions, but also by providing brief training so that language use and attendant assumptions were explicitly treated.

Our instrument for judging set membership and inferring item quality was composed of dimensions representing a variety of features. Some of these are typically expected: the match between the cognitive complexity called for in the specifications and that exhibited in [illegible] item, the match between format called for and format provided, etc. Certain other areas were considered much more exploratory. We are, for example, interested in the impact of complexity of language used in the test items. Extant techniques for assessing language in test items sometimes maladapt "readability" formulas to the stacatto discourse of test items. Semantics are often addressed in terms of judged difficulty, and occasionally validity for various subcultural groups. Syntax is almost never directly studied. We planned to begin to inquire into the relationship of such language factors on the items, their relationship to the intellectual activity required, and their importance for refining and improving measures of achievement through the set belongingness idea.

METHOD

Overview. The Item Review Scale (IRS) was developed for making descriptive validity judgments about test items created from domain specifications. The study described here is part of a series of studies conducted to explore the utility and reliability of the IRS and to provide data for further refinement.

The field test was conducted with school teachers and administrators and with graduate students in education as item raters using the IRS. Raters were trained in using the IRS and were then given a domain specification and several items to "rate."

The items used in the field tests were systematically defaced in one or more of the IPS dimensions.

The Item Review Scale. The IRS, which presumes the availability of test specifications, is used in an item-by-item review of a test or item pool. The specifications may be those that accompany a particular test or those that have been locally developed to guide in creating or selecting test items. In either case, once a set of test specifications has been developed so as to express the testing intentions, the next task is to assemble test items that match these inteetions. This careful matching increases the likelihood that the test will provide a valid assessment of student performance in the content area and under the conditions described in the specifications.

The Item Review Scale, then, helps in judging the probability that any given item is a legitimate member of the hypothetical set, pool, or universe of items defined by the test specifications. More specifically, the IRS is used to judge, or rate, the probability of match between the test specifications and any given item along eight independent categories or dimensions.

The first six rating categories of the IRS parallel the basic structure of domain-referenced test specifications. These dimensions consist of: Domain Description, Content Limits, Distractor Domain or Response Criteria, Format, Directions, and Sample Item. In addition, two dimensions that reflect test item features of linguistic and thinking complexity are also included in the IRS since they affect the probability of a match between the item and the test intentions as embodied in the specifications.

The first category of the IRS concerns the general domain description. The second dimension, Content Limits, compares the description of eligible subject matter and item features with the test item's contents and features. The third dimension is the Distractor Domain or Response Criteria, depending on whether the item is a selected or constructed response type. For selected response items, the specification rules for creating wrong answer alternatives are compared with the actual wrong answer choices used in the test item. For constructed response items, the prescribed criteria for evaluating the response generated by the examinee are compared both to those criteria used and to the suitability of the item and conditions for eliciting a recordable response. Format and Directions dimensions are the fourth and fifth categores for rating the match between specifications and actual items. In those dimensions the concern is whether the layout of the item and the presentation of item and directions conform to the test specifications. The Sample Item provided in each specification is the final aspect of the test specification included in the Item Review Scale.

The two additional dimensions, Linguistic Complexity and Thinking Complexity, provide a structure for gaining access to some of the more subtle sources of complexity that may differentially affect students' performance in a way not described or intended by the specifications. These biasing elements are important, to the degree that the specifications and resulting items are intended to provide the same measure of performance for all students in the given area.

Each category described above appears in the IRS in boldface, 'ollowed by several statements describing test item features that are indicators of item match with the tset specification component under consideration. Raters are asked to use these statements as points for consideration in judging test items against their specifications.

Raters then assign a whole number value, from 0 to 10 inclusive, that best represents their judgment of the probability that the item and specification belong together on the particular dimension being rated, e.g., Domain description. In this 0 to 10 scale, 0 indicates a highly improbable match. After arriving at the probability rating for the item on the first dimension, i.e., Domain Description, raters proceed to the next dimension, rating the probability of item-specification match following the same process. To assist raters in their judgment, guidelines for rating values are presented as an attachment to the IRS.

When all eight dimensions have been individually scored, an overall probability rating may be calculated for the item. The final calculations are guided by the Overall Item Rating Scale which applies a weighting system to incorporate the scores in each dimension.

Interpretations of the ratings are offered in terms of the three features judged to be most critical--Content Limits, Distractor Domain or Response Criteria, and Thinking Complexity. Implications for item revision or, where necessary, specification revision, are also briefly stated. It is expected that this information will allow for more reliable, confident decisions by test makers to use, modify, or reject particular test items.

Development of Field Test Materials. The materials developed for the IRS field tests included a set of domain specifications and test items for rating, and a rating sheet.

The domain specifications were developed to parallel specifications assumed by the IRS. In particular, the components and organizational scheme are those developed in earlier work (Baker, 1974). Because the IRS field test would include a variety of local practitioners, including elementary and secondary teachers and administrators, with different subject matter familiarity, specifications covering a range of subject matter and grade levels were developed. Final specifications included English grammar; reading comprehension; mathematics: elementary level set theory, junior high level basic algebra; science: elementary level geology and secondary level biology.

Each specification was used to create eight to ten test items in their best possible form. These items were then deliberately altered so as to violate the domain specifications either explicitly or implicitly. Some items were loaded with complex linguistic or processing aspects to address those two dimensions of the IRS. Additionally some items were altered in several dimensions, some in only one. These violations were done with specific IRS categories in mind and defects were noted at that time in IRS terms. "Expert" ratings were recorded for two of the specifications-items sets. In the field testing reported here, the elementary science: geology specification-items set was used since it was felt that reading or math might be a more specialized or familiar subject area for some of the raters. Elementary science was considered the least likely of all the topics to create an imbalance in motivation and difficulty among raters due to differences in subject familiarity.

The rating sheet for recording the IRS scores by dimension was developed specifically to encourage note taking on item flaws. Also note, dimensions are listed down the left side and raters rate items from right to left. In this manner each item may be covered up after rating without covering up the list of IRS dimensions. This "covering up" was encouraged to maintain independence of ratings from item to item. It was pointed out to raters during field testing that it was often tempting for consistency's sake to "look back" at a rating, but that this was not desirable.

Raters. The IRS data reported here were gathered on two separate occasions. The first IRS field test was run as an in-service training session set up by a local school district. These raters (Group A), although primarily elementary level teachers, did include three principals and a school psychologist.

The second IRS field-test was run during the regular session of a graduate seminar in education, on instructional analysis. In this instance, raters (Group B) were all graduate students, though several of them reported themselves as teachers.

Rater training. Before using the Item Rating Scale, Group A raters participated in a workshop training module on constructing domain specifications. The culmination of this module was group construction of a domain specification. The IRS field test was presented to Group A as a second training module, complimented by and following the earlier module on developing and using domain specifications. Couched in terms of local district testing needs, especially regarding minimum competency test development or selection, the IRS was described as a systematic method for reviewing test item qualities for adherence to the domain specifications for generating those items. Preceding the IRS training for Group B, an abbreviated presentation of a domain specification training module covered the same content, omitting the group opportunity to write a specification. The IRS field test was similarly introduced as an instrument for verifying item fidelity to specifications.

Raters for both groups had at their disposal written materials that included: 1) the Item Rating Scale, presented dimension by dimension, with guidelines for item rating decisions subsumed under each dimension 2) rating sheets; 3) a rating interpretation guide; and 4) three domain specifications, each followed by eight related items.

Training of raters proceeded as follows: each of the eight item dimensions of the IRS were presented as the standard components of a good, complete domain specification. Examples of critical features described in such a specification drew upon earlier Module I materials and discussion.

The final two categories of linguistic and thinking complexity were presented as item qualities that are not necessarily explicitly included in domain specifications, but which may confound the content or skill (i.e. domain) tested by the item. In such an instance the item no longer offers a valid description of the student/test taker's ability in the original, intended domain.

The eleven point rating scale was characterized as representing a spectrum of probability of match between domain specification and item on the six explicit and two implicit dimensions of item qualities or features. The concept of our spectrum-like rating scale and the concept of "probability of fit" were each discussed at some length until the raters seemed confident and accurate in their understandings. The two concepts were linked by asking raters to consider a domain specification as a description of the hypothetical pool of items that would exhaustively sample competent behavior in the given content area. It was suggested that from this hypothetical universe of items, tests could be created by generating items that followed the domain specification. However, any test would necessarily only present a limited sample of that larger hypothetical universe of items described by the specification. For this reason, it is important to consider how adequately and fairly those test items sample the universe of competent/criterion behaviors in a particular content or skill area (i.e., a domain). This consideration could be described as a probability estimate of "belonggingness" of the test item to the larger pool or set of items detailed by the domain specifications. Thus item ratings for descriptive validity decisions could be presented as a judgment of likelihood of membership. In a set theory sense an item might be a partial or total member of a larger set of items, according to the membership rules (domain specifications) governing that set of items. The use of a 0 to 10 scale allows for probability-like statements of the membership relationship.

Raters were next directed to the overall item rating sheet in their materials. This sheet also indicated the additional weightings given to ratings for Content Limits, Distractor Domain/Response Criteria and Thinking Complexity, as well as the method of arriving at an overall scene for the rated item. From this point, raters were directed to the interpretation guide of recommended strategies for following up on ratings.

The above familiarization with materials, concepts and tasks required approximately forty-five minutes.

Rating Task. Raters were then lead through an item rating as a group. The specification used was included in their materials, as were the eight items to be rated. For both field-test settings an elementary geology--application level specification was used (see Appendix for specification and items). Raters were asked to read over to themselves the specification and the item before the actual group rating process began.

The first item was then rated, one dimension at a time. To do this individual raters were randomly called upon for their ratings and underlying rationale. Group consensus was reached before the next dimension was rated. Following the walk-through example, item number two was rated by individuals, after which ratings were compared and were discussed for those dimensions with major discrepancies among raters. This process took up approximately half an hour. Following this, the group of raters were asked to go ahead by themselves to rate the remaining items.

After the rating task was completed, raters were encouraged to discuss their experience with the items and the IRS instrument. This "feedback session" has been useful in our refinements and revisions of the IRS and the training session. Information and implications are described at greater length in the Discussion section.

RESULTS

The nature of the study reported here is exploratory rather than confirmatory. Thus, data are presented in the context of furthering development or refinement of the Item Review Scale, instead of in repsonse to a priori hypotheses. In this former context topics of interest are rater reliability, characteristics of the IRS dimensions and their sensitivity to a variety of item features. (Tables are contained in the Appendix.)

Rater reliability. Table 1 presents the inter-rater reliability coefficients for each of the eight dimensions of the IRS, as well as for the instrument as a whole. These reliabilities were calculated for items across all raters with recordable data. Note each dimension was found to be at or above r=.75. Specifically, Domain Description recorded r=.99; Content Limits, r=.76; Distractor Domain, r=.77; Format, r=.93; Directions, r=.80; Sample Item, r=.75; Linguistic Complexity, r=.80; Thinking Complexity, r=.94. The overall reliability for the IRS, across all raters, items and rating dimensions, reached r=.67.

Tables 2 through 9 present means and standard deviations for the rater group on each of the eight field test items. Item totals for the rater group are also listed.

It should be noted that the amount of variance associated with any one dimension differs across items. Recall that each of the eight items was intentionally "defaced" to present a variety of type and degree of defects possible that would mar a test specification and item match. Thus, for example, items 3 and 4 yield similar means and standard deviations for ratings of item congruence with the domain description component of the specification. However, these two items show very different ratings and group consensus, as indicated by means and standard deviations, for the Format dimension. Where item 3 is highly scored at an average of 9.9 (out of 10) with a standard deviation of 0.4, the match between domain specification and item 4 on the same dimension is rated by the group an average of 3.4, with very little consensus, standard deviation 4.4.

Perhaps the most interesting aspect of rater consensus is the finding that, despite IRS component variations within items, overall item scores show generally low standard deviations for the eleven point scale (0 to 10 inclusive). For example, five of the eight items have standard deviations of 2.0 about their average total ratings. (The remaining three items report standard deviations ≤ 3.0, i.e., 2.2, 2.3, 2.9).

Dimension characteristics. To determine the relationships among dimensions of the IRS, as well as the relationship between each dimension and the item total rating, interdimension correlations were calculated for each item. These are displayed on Tables 10 through 17.

Peculiar to items 1 and 2, interdimension correlations could not be determined for several cells, due to the lack of variance on one of the two dimensions. In item 1, dimension 4, Format, received a rating of 10 from all twenty-one raters. Accordingly, no dimension 4 correlations are reported for that item. Item 2 correlations for dimension 4, Format, are not available for the same reason. In addition, two cells of the dimension 5, Directions, correlations are similarly affected.

Again, due to the "built-in" variety in type and severity of item defects, correlations among dimensions across all items would not be meaningful and are not presented. To the extent that each IRS dimension is selectively sensitive to particular item features, it would be expected that different configurations of features within an item would affect a different pattern of relationships among rating dimensions. For example, an item with serious formatting problems would be expected to yield a different rating pattern than an item with serious linguistic complexity biases.

Dimensions correlating with total scores vary across items. However, Content Limits, Distractor Domain and Thinking Compexity are the only dimensions consistently and strongly correlated with the overall rating of each of the eight items.

Interrelationships among the dimensions also vary depending upon the item features of each of the filed test items. Although most dimensions do correlate highly on at least one item, ($P \leq .01$), several dimensions never share a relationship for this sample of items. Interestingly, dimension 5, Directions, is never correlated with two of the three critical features, Content Limits and Thinking Complexity. For those two critical features, dimension 5 is the only dimension with which they show no interrelationship. The third critical feature, Distractor Domain, also is never significantly correlated with only one other IRS dimension, Format, dimension 4.

Besides the critical features above, Format and Directions do not significantly correlate with other dimensions: Format with Domain Description (1) and Linguistic complexity (7); Directions with Sample Item (6). Finally, Domain Description (1) and Sample Item (6) do not appear to correlate of any of the eight rated items.

Sensitivity to item features. To determine the sensitivity and practicality of the Item Review Scale, we sought data reflecting on the success of the

IRS dimensions in guiding naive raters to detect both overt and subtle item flaws. This responsiveness of the IRS, as applied by raters to the given items, was determined in part by a discrepancy index representing the distance between the group's average ratings and an "expert rating." The expert ratings were compiled by the authors as part of the materials development described in the Methods section.

These discrepancy scores were calculated for each dimension of each item and for the overall item ratings of each item. The data are reported in Tables 2 through 9. As calculated (expert rating—group rating = discrepancy index), a positive difference indicates the group's average ratings were lower than the rating given by the expert; a negative difference indicates the reverse i.e., the group's average rating was higher than that of the expert.

With very few exceptions (see Tables 3,4 and 6), raters were more conservative than the "expert" in their ratings. In overall item ratings, only item 3 was scored higher by the raters than by the "expert" (discrepancy = -0.9; see Table 4.). A look at dimension rating discrepancies across items shows this consistent relationship.

For the Domain Description (dimension 1), only three of eight items were rated higher by the rating group than by the "expert." In two instances, items 3 and 5, the discrepancy was fairly large, -2.4 and -2.2 respectively (Tables 4 and 6). For the Content Limits, only one of eight items was rated higher by the rating group than by the "expert." That item, 3, showed a large difference, -2.9, between the two ratings (Table 4). For Distractor Domain, only two items are rated above the expert's ratings (items 2 and 3). The Format dimension, 4, is rated consistently at or lower thanthe score given by the expert. Directions (dimension 5) is only once (item 5) rated above the expert rating. Sample Item is three out of eight times rated higher by the field test group. Linguistic Complexity (dimension 7) is consistently rated lower and Thinking Complexity (dimension 8) only once rated higher than the expert's rating.

Besides the consideration of higher or lower judgments by the rating group, rating discrepancies provide a sense of the size of the differences between the perceptions of the rating group and those of the expert regarding each dimension of the items.

Item 1 shows little discrepancy with the exception of Linguistic Complexity, +3.3 (Table 2). Item 2 ratings also line up very closely with expert scores; only a -2.6 difference on the Sample Item seems indicative of a genuine difference in perceptions. Item 2 is further interesting because the low discrepancy indices hold up over a range of dimension scores, from 5.4 to 10.0 (see Table 3). Item 3 is a mixed message. Although half of the IRS dimensions are rated higher and half lower than the expert ratings, the overall rating is quite close (discrepancy = -0.9). The size of the discrepancies is either smallish (i.e., 0.0 to 1.1) or fairly large (i.e., 2.4 to 2.9). The same mixed pattern of higher/lower than expert ratings and big/little discrepancy sizes may be noted for item 5. Items 4 and 6 are the two items with the largest discrepancies on the most dimensions (see Tables 5 and 7). Item 8 shows very small discrepancy sizes except for the Distractor Domain's large +3.3 (see Table 9).

DISCUSSION

Empirical support for the IRS methodology. The study and data reported in this paper were undertaken in pursuit of a systematic, reliable methodology for making sensible judgments about the descriptive validity of domain-referenced test items.

Our initial concern for useful item validity decisions lead us away from the currently used YES/NO standard for item-specification match, toward providing a range of response choices indicating probability of a match between item and specification. In our first study using this concept of a range of judgment values, raters had no difficulty with the continuum concept, but were inconsistent both within and among themselves in the bases for their decisions and the importance of item features. Naturally, reliability for that early instrument was low.

The Item Review Scale used in the study reported here was developed from that early experience to structure the information sources used by raters making their judgments of match. Additionally, because of our concern with differences in the relative importance of item features affecting validity, our Item Review Scale was refined to include a weighting system. This instrument was first tried out in the study reported here.

The high reliability coefficients for each dimension of the IRS provide empirical support for our approach for structuring data sources used in the decision process about test item validity. Further, despite the potential for difficulty in training raters to use the Linguistic and Thinking Complexity dimensions tapping more subtle, implicit qualities of items, reliabilities for these categories are quite high ($r = .80$ and .94 respectively).

Concern with the weighting of dimensions was twofold. Primarily, we were hypothesizing those item features considered most important to overall item validity. The selection of those features was not at all random, however. The Content Limits, Distractor Domain and Thinking Complexity were considered to be the least obvious features and therefore those features most difficult to modify or otherwise take into account. Weighting ratings in these dimensions by a factor of three ensured they would tip the balance in any borderline item case.

The correlation between each of these features and the total item score supports our original choice of weightings. As pointed out in the Results section, those dimensions correlating with total item score varied greatly across the differently defaced items. However, our three "critical features" were consistently, strongly correlated with overall item ratings.

Further, the intercorrelations among dimesions varies with those features emphasized in the defective items suggesting raters' use of dimensions is sensitive to the variety of item flaws. That the discrepancy indices are generally low for both item totals and IRS dimension ratings, suggests that raters and categories are appropriately sensitive to item flaws.

Thus, our study data bear out the meaningfulness and reliability of the item dimensions on the IRS.

Further considerations of IRS dimensions and rating scale. In addition to the general approach of our IRS instrument, this study has provided formative information suggesting modifications and further directions for study. This inout falls primarily into two categories, IRS dimensions and the eleven point (0 to 10) rating scale.

Looking at the means and standard deviations among items for each dimension, it becomes apparent that certain item features occasion more rater variance than others. This may reflect from several different conditions. In the instance of several correlated dimensions for an item with a single gross defect, it might be the case that raters, as a group, were unable to nail down the problem to one category, although they were able to detect some sense of problem. Such a case may result in large variations among raters in their selection of which dimension to mark off; or, it might affect high intercorrelations among dimensions as raters generalize the problem across several dimensions. These intercorrelated dimensions would then be expected to correlate highly with the total.

Item 6 presents just such a case (see Table 7). According to expert ratings, representing built-in defects, Linguistic Complexity is the single msot important dimension affecting a mismatch between item and specification. Raters, on the other hand, show large varaitions around seven of the eight dimensions (standard deviations 2.7 to 3.6) along with high discrepancies between the group's average and the expert rater for five of these same seven dimensions (discrepancy indices ranging from 2.1 to 3.8). Further, these discrepancies indicate that the raters gave lower scores for these dimensions than did the expert. Each of these five dimensions turns up significantly correlated with the total item rating (P= .01). Interdimension correlations for item 6 suggest raters tended to mark down the item in the following sets of dimensions: (a) Domain Description--Distractor domain--Sample Item; (b) Format--Sample Item; (c) Linguistic Complexity--Thinking Complexity. These sets are natural and logical in their membership. Set (a) above, are each related to the more global intentions of the test item in describing kinds of behaviors expected. Set (b) keys in on the mode of presentation; set (c) suggests raters' concerns with adulterations in the item in categories not covered in the domain specifications.

For this and other items, dimension clusters of rater responses suggest the need for further investigations of the relationships among IRS features would be useful. It may be that raters are using surface features of language, e.g., length of words or prose passage, rather than deep structural features of syntax; or, that semantic complexity judgments are resulting in relationships between Linguistic and Thinking Complexity.

A final implication of the data on IRS dimensions concerns the advisability of returning to a binary scale (YES/NO, OFF/ON) for rating certain item features that are clearly either present or absent, a hit or a miss. Such features might include congruence with Domain Description, Format or Directions.

In addition to information on IRS dimensions, the post-task discussions with the rating groups uncovered issues related to the rating task itself. Raters suggested the desire for anchors for the scale and perhaps the need for a smaller range of values. Subsequent to the study reported here, a guideline for assigning ratings was developed and incorporated in additional field tests. The utility of collapsing the eleven point range of probability ratings remains to be evaluated in terms of the trade off in lost information.

Serendipity. As with most exploratory studies, our study resulted in unexpected treasures. Although these surprises are valuable information, they were not specifically planned for at this stage of IRS development.

The first pleasant discovery was that our IRS instrument and methodology were understandable by teachers, administrators and inexperienced graduate students. While it has always been the aim of our IRS development efforts to service the practioners as well as the research and development community, we have never been certain of the practioner response. Not only were the district and school level personnel able to understand and use the IRS, they were also happy and eager to do so. This reaction has been the rule in all of our field tests since.

The second happy discovery was that this audience found our instrument useful and "exportable." Many school districts are currently involved in creating their own competency testing or achievement monitoring systems. The IRS, especially as presented in conjunction with the module on domain specifications, was apparently immediately useful and valuable in these district development efforts as an item check instrument. We had the pleasurable experience of being asked for additional copies of our IRS at the field tests described here as well as those since.

The "exportability" or generalized utility of the IRS was demonstrated recently in another school district field test. On this occasion the district had already developed a system of "instructional models" from which district-wide test items in basic skill areas were being developed. The teacher group with whom we worked had no difficulty translating the six domain specification dimensions of the IRS to comparable components of the district's models.

A third serendipitous treasure was also uncovered at the district field test just mentioned. This unusually sophisticated district had, in addition to a model-based item development system, an "item checklist" designed to screen out item flaws beyond just the "specific determiners" level. When the teacher group of raters applied our IRS methodology and instrument to items already screened with the district device, additional serious flaws were detected in several items.

A final note on unexpected benefits of the IRS methodology is that the rating process consistently evokes a reconsideration of testing goals and often a revision of domain specifications (or "models") that embody those goals.

Future efforts. Meaningful item analyses, accessible to the layman and adaptable to a real-world variety of specification-based test development, appear to be among the benefits of our still-developing Item Review Scale. Additional studies will enable us to tighten the technical qualities and increase the ease of training. At some later date, the IRS will be used in conjuction with test items for which real student data exist. This will allow us to consider the kinds of distinctions our instrument is making. For instance, pending further refinement of the Linguistic Complexity dimension, field tests with such might help uncover those elusive biasing features in test items.

REFERENCES

Arbib, M.A., (ed.) Algebraic theory of machines, languages and semigroups. New York, N.Y.: Academic Press, 1968.

Baker, E.L., Beyond objectives: Domain-referenced tests for evaluation and instructional technology. Educational technology, 1974,14,10-16.

Doctorow, O., A preliminary comparison of semigroups and fuzzy sets in test domains. Unpublished paper, Center for the Study of Evaluation--UCLA, Los Angles, CA. 1977.

Guion, R.M., Content validity--the source of my discontent. Applied psychological measurement, 1977,1,1-10.

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

Millman, J., Sampling plans for domain-referenced tests. Educational technology,1974,14,17-21.

Popham, W.J., Criterion-referenced measurement. Englewood Cliffs, New Jersey: Prentice-Hall, Incorporated, 1978.

APPENDIX

THE ITEM REVIEW SCALE

(IRS)

DIRECTIONS

The Item Rating Scale (IRS) is intended for use in making systematic content validity judgments for domain-referenced tests by comparing test specifications with items. The Scale is devised in such a way as to provide feedback, as well, for revising items or specifications as necessary. In using the IRS, one test item at a time is rated against a set of test specifications.

1. Get a copy of the test specifications and the items you wish to rate.

2. Go through the categories of the IRS using the statements in each section to direct you in judging the compatability of your item with the six test specification features and the two additional categories concerned with complexity issues.

3. In each section, rate the probability that your item is a member of the hypothetical set of items described by the test specifications in that category. Use a scale of 0 to 10 to rate your item, letting 0 indicate a highly improbable match and 10 a highly probable one.

   The following guidelines are suggested for assigning number ratings in each section:

   0,1,2 *This rating range should be used for items that are completely unrelated to the specification in the dimension you are rating.*
   3,4,5 *This rating range should be used for items that are vaguely related and/or inadequate.*
   6,7 *This rating range should be used for items you feel would definitely require a second look and some revision, but which you feel reluctant to totally abandon.*
   8,9 *This rating range should be used for items that you feel are good representative match-ups with the specifications although slightly off.*
   10 *This rating should be used for items that are beyond a doubt perfect examples of the specification.*

   Enter your rating in the box provided.

   Space for taking notes has been provided with each section or category. It is strongly suggested that you take advantage of this to make comments about the item as you rate it. Such notes will be useful later in revising the item or the specifications.

4. Complete the Overall Item Rating sheet by carrying over the rating scores from each section to the appropriate line of the rating sheet. Make the calculations indicated in the directions there, applying the rating weights where indicated.

5. Refer to the Interpretation Guide for rating explanations.

6. REMEMBER YOU ARE RATING THE MATCH BETWEEN THE ITEM AND THE SPECIFICATION, NOT THE ITEM AND YOUR EXPECTATIONS OR STANDARDS! ALSO, EACH IRS CATEGORY SHOULD BE RATED INDEPENDENTLY OF THE OTHERS, FOR EXAMPLE, DOMAIN DESCRIPTION RATINGS DO NOT INCLUDE CONTENT LIMIT CONSIDERATIONS. USE THE STATEMENTS PROVIDED TO GUIDE YOUR JUDGMENTS.

I. DOMAIN DESCRIPTION

1. The test item is a good and fair representative of the subject area outlined in the domain description of the test specifications. It does not assess an obscure or unusual aspect of the domain.

2. Test item conditions are not at odds with test intentions. This is especially important in constructed items.

3. The test item content is closely related to the instructional objective(s) stated or implied in the domain description.

II. CONTENT LIMITS--SELECTED RESPONSE ITEMS ONLY

1. The item and additional accompanying material (e.g., graphs, maps, reading selections) follow the content limits on length and general difficulty level.

2. The item and additional accompanying material follow the content limits on eligible content, descriptive detail and completeness of information provided.

3. The solution processes required by the student to answer the item match those described or implied in the content limits.

II. CONTENT LIMITS--CONSTRUCTED RESPONSE ITEMS ONLY

1. The item matches the content limits on eligible content, descriptive detail, or completeness of the prompting information provided.

2. The item provides a context for responding that is similar to that described in the content limits (e.g., time restrictions, length of written/oral response, equipment or aid restrictions, warmup or false start provisions).

3. The mental processes required by the student to respond to the item seem to match those described or implied in the content limits.

III. DISTRACTOR LIMITS--SELECTED RESPONSE ITEMS ONLY

1. The alternative answers, or distractors, provided in the item require the test taker to discriminate important features or factors described in the distractor domain as differentiating correct from incorrect answers. Distinctions between correct and incorrect answers are not based on trivial or irrelevant features.

2. The distractors provided in the item correspond to the content limits on number, length, and general level of difficulty.

III. RESPONSE CRITERIA--CONSTRUCTED RESPONSE ITEMS ONLY

1. The rules used to judge the student's response are those described by the response criteria.

2. The item prompt provides a context for responding that is appropriate to the response criteria for juding the content and style/form of the response (i.e., likely to elicit a judgeable response).

3. Problems arising from incomplete or inadequate answers are dealt with in a way that upholds the testing intentions of the specifications.

IV. FORMAT

1. The organization and display (layout) of the item conforms to the format description in the test specifications.

2. FOR SELECTED RESPONSE ITEMS ONLY: The organization and display of any additional information (e.g., maps, graphs, pictures, reading selections) conforms to the format description.

   FOR CONSTRUCTED RESPONSE ITEMS ONLY: The context or conditions for responding to the item (e.g., time limits, space limits, available equipment) conform to the format description.

V. DIRECTIONS

1. The directions for completing the test item correspond to the description of test directions in the test specifications.

2. The reading level and complexity of the directions follow the description of test directions in the test specifications; or seem to be within suitable range for the intended test takers.

VI. SAMPLE ITEM

1. The sample item and the test item being rated could come from the same set of items described by the test specifications.

2. The sample item and the test item are very similar in content and either distractors or response criteria.

3. The sample item and the test item are very similar in format and directions.

LINGUISTIC COMPLEXITY

1. Vocabulary used in the item is consistent with the test specifications for item difficulty. Words are not used that have different or unfamiliar meanings for different students or student groups.

2. Item language structure (including, e.g., the use of compound, complex sentences, antecedents) is consistent with the test specifications for item difficulty.

VIII. THINKING COMPLEXITY

1. Those mental processes required for the solution or performance of the test item, but that are not described in the domain description or content limits (i.e., are assumed) are readily available to all test takers at some necessary level of competence (.e.g, drawing ability, handwriting legibility, short-term memory capacity, imagination, ability to separate relevant from irrelevant, detail from generalization).

2. Directions for completing the test item provide the same amount of information and structure for all test takers. Everyone has the same understanding of what is expected and of what the limits or rules for answering are.

3. FOR ITEMS WITH NONVERBAL COMPONENTS, it is reasonable to assume that these components conform with the content limits or distractor domain in their intended meaning, and that this interpretation is stable across all groups of test takers.

OVERALL ITEM RATING

1. Recopy item ratings from each section, making the indicated weighting adjustments for the starred features: Content Limits, Distractor Limits or Response Criteria, and Thinking Complexity.

DOMAIN DESCRIPTION ______

*CONTENT LIMITS (____ x 3) = ______

*DISTRACTOR LIMITS OR RESPONSE CRITERIA (____ x 3) = ______

FORMAT ______

DIRECTIONS ______

SAMPLE ITEM ______

LINGUISTIC COMPLEXITY ______

*THINKING COMPLEXITY (____ x 3) = ______

TOTAL ______

2. Total the scores. Divide the total by 14. This number is the overall item rating.

OVERALL ITEM RATING ______ ÷ 14 = ______

3. Refer to the Interpretation Guide for assistance in making decisions about the item and for suggestions for modifying the item according to its rating.

INTERPRETATION GUIDE

ITEMS RATED 7 OR BETTER

IF ALL THREE STARRED CRITICAL FEATURES ARE RATED 8 OR BETTER*, your item is good, basically in conformity with the test specifications. Review and rewrite efforts should be directed toward other features that scored low, e.g., Format. Use the statements in the IRS rating categories to guide your work.

IF ONE CRITICAL FEATURE RECEIVED A RATING OF 7 OR LOWER*, go back to the specifications on that feature. Try to better align your item with the testing intentions described in the specifications. Use the statements in the IRS to help direct your thinking. You also have problems with other features. Rewrite the item but review it again to be certain all critical features are up to par.

IF MORE THAN ONE CRITICAL FEATURE RECEIVED A RATING OF 7 OR LOWER*, the item has serious validity problems. If this is tne kind of test item you want, then you should reconsider the specifications you are using. They may need to be better conceptualized, reconceptualized, or more complete in their description of item qualities. If the specifications are closer to what you want to be testing, throw out the item. Find or write a new item.

ITEMS RATED BELOW 7

IF ALL THREE STARRED CRITICAL FEATURES ARE RATED 8 OR BETTER*, your item is potentially a good item but has serious problems in presentation. Go back to the specifications for those features receiving the low ratings. Clean up your item. Use the statements in the IRS rating categories to guide your efforts.

IF ONE OR MORE OF THE CRITICAL FEATURES SCORED 7 OR LOWER*, your item isn't worth the fix-up effort. Before you start over, reconsider the specifications with which you are working; they may need to be better conceptualized or more complete in their description of item features.

* Before rating weights are applied.

SPECIFICATION BEING RATED________________________________

RATER TITLE________________________________

COMMENTS: (additional comments can be made on the reverse side)

RATING SCALE

Domain Description			
*Content Limits			
*Distractor Domain or Response Criteria			
Format			
Directions			
Sample Item			
Linguistic Complexity			
*Thinking Complexity			
TOTAL			
÷ 14 =			

SPECIFICATION BEING RATED ____________________________________

RATER TITLE ____

COMMENTS: (additional comments can be made on the reverse side)

ITEM NUMBER

Domain Description	____	____	____
*Content Limits	____	____	____
*Distractor Domain or Response Criteria	____	____	____
Format	____	____	____
Directions	____	____	____
Sample Item	____	____	____
Linguistic Complexity	____	____	____
*Thinking Complexity	____	____	____
TOTAL	____	____	____
: 14 =	____	____	____

SPECIFICATION BEING RATED________________________________

RATER TITLE________________________________

COMMENTS: (additional comments can be made on the reverse side)

ITEM NUMBER			
Domain Description	____	____	____
*Content Limits	____	____	____
*Distractor Domain or Response Criteria	____	____	____
Format	____	____	____
Directions	____	____	____
Sample Item	____	____	____
Linguistic Complexity	____	____	____
*Thinking Complexity	____	____	____
TOTAL	____	____	____
: 14 =	____	____	____

SPECIFICATION BEING RATED________________________________

RATER TITLE________________________________

COMMENTS: (additional comments can be made on the reverse side)

ITEM NUMBER

Domain Description			
*Content Limits			
*Distractor Domain or Response Criteria			
Format			
Directions			
Sample Item			
Linguistic Complexity			
*Thinking Complexity			
TOTAL			
: 14 =			

SPECIFICATION BEING RATED ______________________________

RATER TITLE ______________________________

COMMENTS: (additional comments can be made on the reverse side)

ITEM NUMBER

Domain Description	____	____	____
*Content Limits	____	____	____
*Distractor Domain or Response Criteria	____	____	____
Format	____	____	____
Directions	____	____	____
Sample Item	____	____	____
Linguistic Complexity	____	____	____
*Thinking Complexity	____	____	____
TOTAL	____	____	____
: 14 =	____	____	____

SAMPLE DOMAIN SPECIFICATION AND
ITEMS USED IN THE FIELD TESTING

Grade Level: Grade 7 and/or 8

Subject: Elementary science-geology

Domain Description: Applying knowledge of destructional forces and constructional forces that alter the earth's surface, to make predictions of effects, given causes, and/or to hypothesize the causes given effects.

Content Limits:

1. Constructional forces include the following:

   volcano — pressure forces magma (lava) to break through the earth's crust

   folding — forces press the earth's sediments sideways, causing rock layers to become folded upward

   earthquake/faults — settling and shaking down the earth's crust

   Destructional forces include the following:

   erosion — flowing water bumping and wearing away the rock and land, pulling away pebbles and boulders that hammer away at the land as the water flows over it

   wind erosion, sand storm blasting and wearing away the surface of the land

   glacier action — scrape and drag ice and rock across the surface of the land deepening valleys and smoothing out rocky mountains and hills

   lichens — break up rocks by acid secretions

   sunlight/freezing — cracks--expansion and contraction of rocks causes break-up

2. Items should not use key terms, e.g., erosion, to pose the problem, but should use description or definitions to convey the process, function.

3. Pictorial representations of causes or effects may be used if labelled and accompanied by a verbal prompt of the given part of the item.

4. All prose material below grade level readability.

Distractor Domain:

1. Wrong choice alternatives must be the result of mismatches (mixups) between causes and effects, or misidentification of causes or effects (misnaming).

2. Wrong choices must not be from outside the conten limits.

3. Wrong choices must not be the result of inability to decipher the meaning of pictorial representations

Format: Multiple choice, four alternatives.

Directions: Students will be asked to select the correct answer. Each item will pose its own specific question.

Sample Item:

Flowing water, such as a river, can change the surface of the land by_______ _____________.

a. creating a strong pressure against sediments forcing them upward into mountains.
b. giving off acids which slowly eat away at the rocks making them crumble apart.
c. causing sudden shifts upward or downward of great rock masses and layers.
*d. carrying pebbles and rocks that scratch and hammer away at the land and rock surface.

Circle the letter of the correçt answer for questions 1-3.

1. Glaciers that once existed in North America are responsible for changing of the surface of the earth by ______________.

[✓ a) scraping and dragging ice and rock across the land.]
b) flooding the land with melted ice and snow.
c) erupting and spreading lava (magma) over the land.
d) settling and shaking the layers of the earth's crust.

2 The Dust Bowl was caused by ______________.

a) wind
[✓ b) erosion]
c) floods
d) glacier

3. Destruction of the earth's surface can be caused by ______________.

a) faults
[✓ b) sun]
c) volcanos
d) earthquakes

4. Match each cause with its effects.

CAUSES

_____ faults

_____ flowing water

_____ sunlight & freezing temperatures

_____ glacier action

EFFECTS

a) acid causes rocks to crumble and break up
b) deepens valleys and smooths out rocky mountains and hills
c) expansion and contraction causes rocks to crack
d) washes away pebbles and soil, causing erosion
e) magma breaks through to the surface as lava

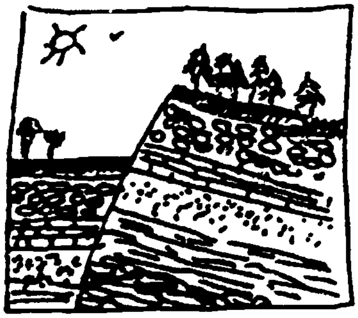
Circle the letter of the correct answer for question 5-8.

5. Lichens are __________.

   a) tall, block-shaped mountains formed by faults
   b) tiny organisms that form sediment deposits in river beds.
   [✓ c) plants that grow on rocks and secrete an acid from the roots.]

6. Solar energy can be conceived of as a destructive force in modifying the face of the earth in that the thermal effects upon rocks is to expand them; while in the absence of solar heat, cold temperatures affect contraction. This expansion and contraction cycle ______.

   a) causes fault lines to deepen and widen creating cracks in the earth's surface.
   b) exerts pressure upon the magma below the earth's crust, sometimes leading to volcanic eruption.
   [✓ c) affects cracks and, eventually, promotes break-up of large boulders and rock surfaces.]
   d) erodes the topsoil, allowing the winds to transform valuable farmlands into veritable wastelands.

7. 

In this picture, layers of the earth's crust have tilted and shifted, creating ___________.

   a) a volcano.
   [✓ b) a fault.]
   c) rocks.
   d) a glacier.
   e) a landslide

8. Wind and rushing water can cause _____ ____.

   a) high tides.
   b) cracks in rocks.
   c) pressure on sediments.
   [✓ d) erosion.]

TABLES

Table 1

Inter-rater reliability across items, by IRS dimension

		r	$\bar{x}$
Dimension 1:	Domain description	.99	8.5
Dimension 2:	Content limits (Wgtd.)	.76	6.2
Dimension 3:	Distractor domain or Response criteria (Wgtd.)	.77	5.9
Dimension 4:	Format	.93	7.2
Dimension 5:	Directions	.80	9.3
Dimension 6:	Sample item	.75	6.7
Dimension 7:	Linguistic complexity	.80	7.5
Dimension 8:	Thinking complexity (Wgtd.)	.94	6.9
TOTAL		.67	6.9

Table 2

Mean ratings across all raters: Item 1

(N)	Dimension	$\bar{x}$	sd	discrepancy (E - R)
21	1. Domain description	9.6	0.9	+0.4
21	*2. Content limits (Wgtd.)	6.7	1.8	+1.3
21	*3. Distractor domain or Response criteria (Wgtd.)	6.6	2.4	+0.4
21	4. Format	10.0	0.0	0.0
21	5. Directions	9.8	0.9	+0.2
21	6. Sample item	8.9	2.0	+1.1
21	7. Linguistic complexity	6.7	2.0	+3.3
21	*8. Thinking complexity (Wgtd.)	9.0	2.5	0.0
20	TOTAL	8.2	0.9	+0.5

* Although these categories are weighted by a factor of three before item totals are calculated, they are presented here in preweighted ratings for easier comparison with other dimensions' ratings.

Table 3

Mean ratings across all raters: Item 2

(N)	Dimension	$\bar{x}$	sd	Discrepancy (E - R)
20	1. Domain description	8.1	3.0	-0.1
19	*2. Content limits (Wgtd.)	5.4	4.2	+0.6
19	*3. Distractor domain or Response criteria (Wgtd.)	5.6	2.9	-0.6
19	4. Format	10.0	0.0	0.0
19	5. Directions	9.9	0.5	+0.1
19	6. Sample item	7.6	2.3	-2.6
19	7. Linguistic complexity	9.8	1.3	+1.1
18	*8. Thinking complexity (Wgtd.)	6.2	3.7	+1.8
18	TOTAL	6.8	2.2	+0.3

* Although these categories are weighted by a factor of three before item totals are calculated, they are presented here in preweighted ratings for easier comparison with other dimensions' ratings.

Table 4

Mean ratings across all raters: Item 3

(N)	Dimension	$\bar{x}$	sd	Discrepancy (E - R)
21	1. Domain description	9.4	1.2	-2.4
21	*2. Content limits (Wgtd.)	9.9	3.6	-2.9
21	*3. Distractor domain or Response criteria (Wgtd.)	8.1	2.9	-1.1
21	4. Format	9.9	0.4	+0.1
21	5. Directions	9.9	0.5	+0.1
21	6. Sample item	7.4	2.3	-0.4
21	7. Linguistic complexity	10.0	4.7	0.0
20	*8. Thinking complexity (Wgtd.)	7.9	2.4	-2.9
20	TOTAL	8.1	1.5	-0.9

* Although these categories are weighted by a factor of three before item totals are calculated, they are presented here in preweighted ratings for easier comparison with other dimensions' ratings.

Table 5

Mean ratings across
all raters: Item 4

(N)	Dimension	$\bar{x}$	sd	Discrepancy (E - R)
21	1. Domain description	9.3	1.2	+0.7
21	*2. Content limits (Wgtd.)	7.5	3.0	+2.5
21	*3. Distractor domain or Response criteria (Wgtd.)	6.3	3.7	+3.7
21	4. Format	3.4	4.4	+3.6
21	5. Directions	6.7	4.2	+0.3
21	6. Sample item	2.7	3.6	+3.3
21	7. Linguistic complexity	7.2	2.5	+2.8
20	*8. Thinking complexity (Wgtd.)	6.1	3.1	+3.9
20	TOTAL	6.4	2.3	+2.9

* Although these categories are weighted by a factor of three before item totals are calculated, they are presented here in preweighted ratings for easier comparison with other dimensions' ratings.

Table 6

Mean ratings across
all raters: Item 5

(N)	Dimension	$\bar{x}$	sd	Discrepancy (E - R)
21	1. Domain description	5.2	4.3	-2.2
21	*2. Content limits (Wgtd.)	5.2	3.9	+0.8
21	*3. Distractor domain or Response criteria (Wgtd.)	4.5	3.9	+1.5
21	4. Format	3.4	4.3	+4.6
21	5. Directions	8.7	3.1	-0.7
21	6. Sample item	5.3	3.8	-0.3
21	7. Linguistic complexity	6.7	3.4	+2.3
20	*8. Thinking complexity (Wgtd.)	6.9	3.1	+1.1
20	TOTAL	5.7	2.9	+0.2

* Although these categories are weighted by a factor of three before item totals are calculated, they are presented here in preweighted ratings for easier comparison with other dimensions' ratings.

Table 7

Mean ratings across
all raters: Item 6

(N)	Dimension	$\bar{x}$	sd	Discrepancy (E - R)
21	1. Domain description	7.9	3.2	+2.1
21	*2. Content limits (Wgtd.)	6.3	2.9	+3.7
21	*3. Distractor domain or Response criteria (Wgtd.)	6.2	2.7	+3.8
21	4. Format	8.3	3.6	+1.7
21	5. Directions	9.3	1.7	+0.7
21	6. Sample item	7.8	3.0	+2.2
21	7. Linguistic complexity	3.7	3.5	+1.3
20	*8. Thinking complexity (Wgtd.)	4.0	3.2	+3.0
20	TOTAL	6.3	1.7	+2.7

* Although these categories are weighted by a factor of three before item totals are calculated, they are presented here in preweighted ratings for easier comparison with other dimensions' ratings.

Table 8

Mean ratings across all raters: Item 7

(N)	Dimension	$\bar{x}$	sd	Discrepancy (E - R)
21	1. Domain description	8.7	2.5	+1.3
21	*2. Content limits (Wgtd.)	6.0	3.5	+1.0
19	*3. Distractor domain or Response criteria (Wgtd.)	5.3	2.9	+1.4
20	4. Format	4.8	4.7	+3.2
20	5. Directions	9.7	0.9	+0.3
20	6. Sample item	7.3	3.2	+1.7
20	7. Linguistic complexity	8.6	2.2	+1.4
19	*8. Thinking complexity (Wgtd.)	8.3	2.3	+0.7
19	TOTAL	6.9	1.9	+1.4

* Although these categories are weighted by a factor of three before item totals are calculated, they are presented here in preweighted ratings for easier compari-on with other d' ensions' ratings.

Table 9

Mean ratings across
all raters: Item 8

(N)	Dimension	$\bar{x}$	sd	Discrepancy (E - R)
20	1. Domain description	9.5	1.1	+0.5
20	*2. Content limits (Wgtd.)	6.7	3.6	+0.3
20	*3. Distractor domain or Response criteria (Wgtd.)	5.7	1.1	+4.3
20	4. Format	9.2	2.3	+0.8
20	5. Directions	9.7	0.9	+0.3
20	6. Sample item	8.3	2.3	+1.7
20	7. Linguistic complexity	9.0	1.1	+1.0
19	*8. Thinking complexity (Wgtd.)	8.6	1.7	+1.4
19	TOTAL	7.7	8.6	+1.7

* Although these categories are weighted by a factor of three before item totals are calculated, they are presented here in preweighted ratings for easier comparison with other dimensions' ratings.

TABLE 10

Interdimension correlation by item: Item 1

	2	3	4	5	6	7	8	TOTAL
1 Domain description	.12	.34	**	.69*	.33	-.40	.00	.50*
2 Content limits		.52*	**	.34	.66*	.01	.56*	.59*
3 Distractor domain or Response criteria			**	.35	.57*	-.24	.55*	.85*
4 Format				**	**	**	**	**
5 Directions					.34	-.27	.00	.57
6 Sample item						-.43	.86*	.70*
7 Linguistic complexity							-.24	-.20
8 Thinking complexity								.60*

*P ≤ .01

**Correlation coefficient cannot be computed as dimension 4 showed no variance; all raters scored 10 for item 1 on Format.

TABLE 11

Interdimension correlation by item: Item 2

	2	3	4	5	6	7	8	TOTAL
1 Domain description	.68*	.70*	**	-.16	.17	-.26	.77*	.84*
2 Content limits		.76*	**	-.10	.52*	-.26	.46	.88*
3 Distractor domain or Response criteria			**	-.37	.40	-.23	.64*	.91*
4 Format				**	**	**	**	**
5 Directions					-.25	.53*	**	**
6 Sample item						-.31	.20	.45
7 Linguistic complexity							-.04	-.13
8 Thinking complexity								.81*

*P ≤ .01

**Correlation coefficient cannot be computed as dimension 4 showed no variance and dimension 5 no variance for two cells.

TABLE 12

Interdimension correlation by item: Item 3

	2	3	4	5	6	7	8	TOTAL
1 Domain description	.10	.29	-.10	.19	.12	09	.59*	.41
2 Content limits		.27	-.22	-.18	.57*	.29	.26	.83*
3 Distractor domain or Response criteria			.01	.61*	.41	.12	.34	.81*
4 Format				-.07	-.29	.00	-.21	.02
5 Directions					.06	.02	.08	-.08
6 Sample item						-.02	.18	.58*
7 Linguistic complexity							-.43	.29
8 Thinking complexity								.47

*P ≤ .01

TABLE 13

Interdimension correlation by item: Item 4

	2	3	4	5	6	7	8	TOTAL
1 Domain description	.39	.29	.31	.39	.28	.12	.18	.46
2 Content limits		.61*	.27	-.11	.40	.59*	.75*	.91*
3 Distractor domain or Response criteria			.20	.22	.30	.36	.42	.80*
4 Format				.39	.52*	.15	.42	.50*
5 Directions					.42	-.22	.03	.23
6 Sample item						.45	.57*	.65*
7 Linguistic complexity							.68*	.61*
8 Thinking complexity								.82*

*P ≤ .01

TABLE 14

Interdimension correlation by item: Item 5

	2	3	4	5	6	7	8	TOTAL
1 Domain description	.50*	.24	.45	.19	.27	.48*	.43	.57*
2 Content limits		.77*	.71*	.44	.48*	.83*	.75*	.96*
3 Distractor domain or Response criteria			.44	.36	.08	.79*	.74*	.85*
4 Format				.19	.54*	.44	.52*	.71*
5 Directions					.40	.31	.24	.47
6 Sample item						.19	.17	.43
7 Linguistic complexity							.91*	.90*
8 Thinking complexity								.87*

*P < .01

TABLE 15

Interdimension correlation by item: Item 6

	2	3	4	5	6	7	8	TOTAL
1 Domain description	.38	.75*	.17	.32	.42	.02	-.01	.62*
2 Content limits		.22	.06	.22	-.03	.13	.14	.57*
3 Distractor domain or Response criteria			.08	.27	.42	-.07	-.05	.57*
4 Format				-.00	.67*	.12	.20	.44
5 Directions					.23	.18	.12	.34
6 Sample item						.11	.17	.52*
7 Linguistic complexity							.35*	.61*
8 Thinking complexity								.55*

*P ≤ .01

TABLE 16

Interdimension correlation by item: Item 7

	2	3	4	5	6	7	8	TOTAL
1 Domain description	.25	-.04	-.07	.10	-.05	.14	.37	.26
2 Content limits		.46	.47	-.02	.55*	.46	.46	.84*
3 Distractor domain or Response criteria			.21	.02	.49	.36	.24	.69*
4 Format				-.10	.56*	.38	.23	.61*
5 Directions					-.02	-.26	-.14	.03
6 Sample item						.52*	.49	.74*
7 Linguistic complexity							.85*	.71*
8 Thinking complexity								.68*

*P ≤ .01

TABLE 17

Interdimension correlation by item: Item 8

		2	3	4	5	6	7	8	TOTAL
1	Domain description	-.03	-.05	-.16	-.18	-.22	.07	.16	.02
2	Content limits		.63*	.38	-.08	.62*	.57*	.52*	.93*
3	Distractor domain or Response criteria			.38	.35	.78*	.50*	.46	.94*
4	Format				.02	.39	-.04	-.24	.39
5	Directions					.18	-.17	-.06	.13
6	Sample item						.46	.32	.78*
7	Linguistic complexity							.58	.67*
8	Thinking complexity								.63*

*P ≤ .01

QUALITATIVE ANALYSIS OF TEST ITEM ATTRIBUTES FOR DOMAIN-REFERENCED CONTENT VALIDITY JUDGMENTS

Linda Polin and Eva L. Baker
Center for the Study of Evaluation
University of California, Los Angeles

A paper presented at the annual meeting
of the American Educational Research Association
April, 1979, San Francisco

Grant Number
OB-NIE-G-78-0213

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.