DOCUMENT RESUME

ED 121 845 TM 005 277

AUTHOR Rovinelli, Richard J.; Hambleton, Ronald K.
TITLE On the Use of Content Specialists in the Assessment of Criterion-Referenced Test Item Validity.
PUB DATE [Apr 76]
NOTE 37p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage
DESCRIPTORS *Content Analysis; *Criterion Referenced Tests; Data Collection; Evaluation Methods; *Item Analysis; Statistical Analysis; *Test Construction; *Test Validity

ABSTRACT

Essential for an effective criterion-referenced testing program is a set of test items that are "valid" indicators of the objectives they have been designed to measure. Unfortunately, the complex matter of assessing item validity has received only limited attention from educational measurement specialists. One promising approach to the item validity question is through the collection and analysis of the judgments of content specialists. The purpose of this paper are twofold: First, several possible rating forms and statistical methods for the analysis of content specialists' data are discussed. Second, the results of item validation work with science teachers and three of the more promising rating forms are presented. The overall results of the study clearly support the recommendation for expanded use of content specialists' ratings is the item validation process. (Author/RC)

On the Use of Content Specialists in the Assessment of Criterion-Referenced Test Item Validity

Richard J. Rovinelli
National Board of Medical Examiners

and

Ronald K. Hambleton
University of Massachusetts, Amherst

Abstract

Essential for an effective criterion-referenced testing program is a set of test items that are "valid" indicators of the objectives they have been designed to measure. Unfortunately, the complex matter of assessing item validity has received only limited attention from educational measurement specialists. One promising approach to the item validity question is through the collection and analysis of the judgements of content specialists. The purposes of this paper are twofold: First, we will discuss several possible rating forms and statistical methods for the analysis of content specialists' data. Second, we will present the results of our item validation work with science teachers and three of the more promising rating forms. The overall results of the study clearly support the recommendation for expanded use of content specialists' ratings in the item validation process.

On the Use of Content Specialists in the Assessment of Criterion-Referenced Test Item Validity[1]

Richard J. Rovinelli
National Board of Medical Examiners

and

Ronald K. Hambleton
University of Massachusetts, Amherst

The amount of interest and energy that has been expended in the area of criterion-referenced testing and measurement in the last few years has been impressive. A wide variety of theoretical and practical problems have received considerable attention from educational measurement specialists (see for example, Fremer, 1972; Hambleton & Novick, 1973; Livingston, 1972; Millman, 1974; and Popham & Nusek, 1969). Considering its importance, educational measurement specialists have given relatively little attention to the problem of item validation, i.e., the problem concerning the extent to which items are measures of the objectives they have been designed to measure.

The problem of item validation is of particular importance with criterion-referenced tests because of the way the test score information is used. The success of an individualized program depends to a considerable

---

1 A paper presented at the annual meeting of AERA, San Francisco, 1976. The paper has been published as Laboratory of Psychometric and Evaluative Research Report No.24. Amherst, Mass: The University of Massachusetts, 1976.

extent upon how effectively teachers make decisions concerning student mastery of specific instructional objectives. Unless one can say with a high degree of confidence that the items in a criterion-referenced test measure the instructional objectives, any use of the test information for instructional decision-making is questionable.

To date, the two most popular approaches to the problem of assessing item validity have been through the use of item generation rules (Hively, et al., 1973) and the empirical analysis of examinee test data. Relative to the first of these approaches, while the use of item generation rules is intuitively appealing and represents an excellent solution when the rules can be applied, at the present time it would seem that the approach is not practical in content areas besides mathematics. Relative to the second approach, while the use of a variety of empirical methods on examinee test data have been popular among criterion-referenced test developers, at best this approach provides only partial data for the determination of item validity (Millman, 1974; Rovinelli, 1976). A third approach to the problem, which has received very little attention from test developers, is the use of the judgements of content specialists. However, before this approach can become a practical solution to the problem of assessing item validity, there is a need for the generation, organization and comparative analysis of possible data collection techniques and methods of analyzing content specialists' ratings.

Purposes of the Study

In spite of the importance of the item validity problem to the criterion-referenced testing area, to date there does not exist a method-

ology for conducting item validation studies. What does exist is a disorganized set of techniques that address different aspects of the item validity problem. As recently as 1974, Popham posed two important questions that still remain for criterion-referenced test developers:

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

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

Further, Skager (1974) adds the following important questions:

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

2. How does one identify poorly written items by means of item analysis procedures when the frequency of correct response may be extremely high or low, accurately reflecting the achievement status of a particular group of learners?

Given the importance of the item validity question and the shortage of research on the use of content specialists' ratings, this study was designed to achieve two purposes:

1. To generate and to organize appropriate judgemental data techniques and methods of data analysis and reporting,

2. To examine three different techniques for the collection of judgemental information with regard to the type, reliability, and validity of the information provided.

An Organization of Item Validation Approaches

We feel that it is useful to organize existing item validation methods around three rather different approaches: Item generation rules,

empirical methods, and the use of content specialists' ratings.

Through the use of item generation rules, one attempts to ensure item validity by developing a direct relationship between an item on objective during the item construction phase (Anderson, 1972; Bormuth, 1970; Hively et al., 1968, 1973; Millman, 1974). As such, it is an a priori approach as compared to the other a posteriori procedures which are designed to assess whether or not a direct relationship between an item and an objective exists through analyses of data conducted after the item is written. However, the use of item generation rules as currently formulated contain inherent problems which make their implementation in many objectives-based programs impractical.

The second approach, the use of empirical procedures (for example, see Popham, 1971; Brennan and Stolurow, 1971) has been very popular but there remain many problems. For example,

1. The procedures are dependent upon the characteristics of the group of examinees and the effects of instruction.

2. They often require sophisticated statistical techniques and/or computer programs which are not available to the practitioner.

3. When item statistics derived from empirical analyses of test data are used to "select" the items for a criterion-referenced test, the test developer runs the risk of obtaining a non-representative set of items from the domain of items measuring the objectives included in the test.

4. Empirical methods in many instances require pre-test and post-test data on the same test items and this is rarely done in classroom settings.

In situations where a large sample of examinees is available and 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 inform-

ation with regard to the assessment of item validity.

The third approach, the use of the judgements of content specialists, appears to offer considerable promise as a means for assessing item validity. The approach is not dependent on group composition or instructional effects; may not require sophisticated statistical techniques; is not restricted to highly structured content domains; and can be implemented easily in practical settings.

A Methodology for the Use of Content Specialists' Ratings

The first step in the development of a methodology for the use of the judgements of content specialists to assess item validity is to clearly delineate the important issues. Five of the most important issues are:

1. Can the content specialists make meaningful (valid) judgements about the relevance of items to instructional content?

2. Is there agreement amongst the ratings of content specialists?

3. What information is one seeking to obtain from the judgemental data?

4. What variables effect the judgemental techniques?

5. What techniques can be used for collecting content specialists' ratings of test items?

Only the second question above has received serious attention. With respect to the other four issues, we have little information and few clear guidelines.

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

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

Very poor Poor Fair Good Very good

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

0 .20 .40 .60 .80 1.00

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

Same Slightly better Somewhat better
Much better Very much better

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

Not relevant Somewhat relevant Quite relevant Very relevant

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

While Ryan's (1968) study is a step in the right direction, his conclusions on the issue or relevance is weakly supported in that one does not know whether the teachers perceived the judgement of quality the same as a judgement of relevance. On the other hand, the judgement of difficulty correlated highly with actual difficulty which gives a more conventional substantiation of judgemental validity.

The second question concerning the consistency of agreement amongst the content specialists, i.e., the reliability of the ratings, has been examined by a number of researchers (Lu, 1971; Cohen, 1960; Light, 1971 Fleiss, 1971; and Brennan and Light, 1973). It is not our intention to review this extensive literature here. However, a description of one

prominent method for assessing agreement amongst content specialists will follow.

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

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

where X_{ij} represents the category assignment on the ith item by the jth judge. We will assume that the rating scale consists of t ordered categories.

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

		Categories			
		C_1	$C_2 \ldots C_k \ldots$		C_t
Judges	J_1	n_{11}	n_{12}	n_{1k}	n_{1t}
	J_2	n_{21}	n_{22}	n_{2k}	n_{2t}
	J_j	n_{j1}	n_{j2}	n_{jk}	n_{jt}
	J_m	n_{m1}	n_{m2}	n_{mk}	n_{mt}

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

$$y_k = \sum_{r=1}^{k-1} p_r + \tfrac{1}{2} p_k , \qquad k = 1,2, \ldots t$$

where $p_r = nr./mn$

An analysis of variance is then performed on the transformed data. The coefficient of agreement A is calculated as

$$A = \frac{\sigma_H^2 - S_i^2}{\sigma_H^2}$$

where σ_H^2 is the expected within subject variance under the condition that all ratings are equally likely, and

S_i^2 is the observed within subject variance.

A test of the significance of A is conducted indirectly by testing "the hypothesis that the assignments of subjects to the categories by the judges are equally likely for all categories." That is

$$E(S_i^2) = \sigma_H^2$$

The statistic

$$\theta = \frac{S_i^2}{\sigma_H^2} \qquad \text{is } \chi^2/df \text{ distributed}$$

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

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

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

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

The choice of the types of information which is to be collected under the second category will vary from study to study as the choice is dependent on secondary goals or methodological considerations. Examples of secondary goals would be the determination of whether or not the content specialists can judge the difficulty of the items or whether the items were well written. An example of a methodological consideration would be the collection of data which would help validate the rating instrument.

The fourth question concerning the variables which effect the judgements of content specialists is particularly important. In comparing methods for judging the similarity of personality inventory items, Girard and Cliff (1973) found that "the criteria by which subjects were instructed to judge similarities between items in a pair made a large difference in the judgements." Four of these variables which are felt to be important are:

1. Judgemental Procedures: Whenever possible, one should use the simplest of techniques available to collect data. For example, usually, categorical judgements obtained from sorting, rating and ranking procedures are less complex than comparative judgements obtained from similarity, dissimilarity or choice procedures.

2. Format of Presentation: The response task should not be tedius and time consuming. For example, while there are methods which can be used to reduce the number of required responses (Torgeson, 1958), generally the method of paired comparisons should be avoided if the number of stimuli (items) is large, because of the great number of responses involved.

3. Definition of Task: When describing the response task, one should ensure that all the judges are operating under the same assumptions. If one merely asks the judges to rank or choose items according to personal preference, the judges could obtain significant results based not on real differences in the items but on the dimension of preference. For example, the judges could have been ranking the items on any one of the following levels of the preference dimension:

    A. Simplicity/Complexity of item,
    B. Closeness of match to hypothetical objective,
    C. Response mode required,
    D. Style in which the item was written.

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

4. Settings for data collection: In choosing an instrument for collecting the judgements of content specialists, the setting in which the data is to be collected must be taken into consideration. That is, the practicality of its use in both research and non-research settings is a key factor in the choice of instrument.

The fifth question outlined at the beginning of this section is concerned with the choice of instrument which will be used to collect the judgemental data. It is suggested that the test developer choose a technique which conforms as closely as possible to the guidelines set forth under the discussions of questions 1, 2 and 4 above, while at the same time, providing the information described in question 3.

Judgemental Techniques

Three techniques for the collection and analysis of the judgements of content specialists will be described in this section. These techniques were chosen primarily to provide information on the efficacy of

the use of content specialists as a means for assessing item validity and not to provide a definitive answer to the question of which techniques are most appropriate.

(a) An Index of Item Homogeneity

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

According to Hemphill and Westie (1950)

> This index was adopted to give a single numerical evaluation of each item with respect to its homogeneity. Agreement among judges that the item applied to a dimension and agreement that it did not apply to other dimensions in the description were given approximately equal weight in the value of this index.

> The index of "homogeneity of placement" differs in two ways from certain other techniques for examining item content. First it is based on "expert" judgement of probable response to the items, not on actual item response data. Second, unlike indices such as "internal consistency," "homogeneity," or "unidimensionality" all of which refer to relationship among items, the index of "homogeneity of placement" involves both relationships among items (as reflected by judge agreement that certain items apply to the same dimension) and independence of relationship of the item to other dimensions making up the same general heuristic system.

The index appears to be a valid procedure for collecting and analyzing judgemental data on item validity.

The mechanics for collecting data through the use of the Hemphill-Westie consists of having the content specialists rate each item on each of the objectives by assigning a value of +1, 0 or -1. The three possible ratings have the following meaning:

+1 = definite feeling that an item is a measure of an objective

0 = undecided about whether the item is a measure of an objective

-1 = definite feeling that an item is not a measure of an objective.

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

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

where

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

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

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

X_{ijk} is the rating (1,-1 or 0) of item k as a measure of objective i by content specialist j.

While the Hemphill-Westie procedure is conceptually appropriate for the task of collecting judgemental data from content specialists for the purpose of assessing item validity, the computational formula given above has some serious deficiencies. First, the maximum and minimum values are .67 and -.40, respectively. (The maximum value of this index occurs when each content specialist assigns a +1 to the item for the appropriate objective and a -1 for all the other objectives. The minimum value occurs when content specialists assign a -1 to the item for the appropriate objective and a +1 for all the other objectives.) For ease of interpretation it is convenient if the range of the index is from -1 to +1. Second, and an even more serious problem with the index is that its value varies as a function of the number of content specialists and objectives, clearly an undesirable situation since it complicates the problem of

interpreting the index.

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

1. That perfect item objective congruence should be represented by a value of +1 and will occur when all the specialists assign a +1 to the item for the appropriate objective and a -1 to the item for all the other objectives.

2. That the worst judgement an item can receive should be represented by a value of -1 and will occur when all the specialists assign a -1 to the item for the appropriate objective and a +1 to the item for all the other objectives.

3. That the assignment of a 0 to an item is poorer than a +1 but better than a -1. This is in effect saying that it is better for a specialist to not be able to definitely decide whether an item is a measure of an appropriate objective than it is for the judge to feel that the item is definitely not a measure of the objective.

4. That this index should be invariant to the number of content specialists and the number of objectives.

The new computational formula is

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

(All variables on the right-hand side of the expression have the same meaning as in the Index of Homogeneity.)

The choice of a cutoff score for this index to separate "good" from "bad" items can be based on some absolute standard relating to specific proportions of perfect ratings for the items. For example, if one-half of the content specialists judged an item to be a perfect match to an objective, while the others were not able to make a decision, the computed

value of the index would be .50. Thus, test constructors obtaining I' values of .50 would know that at a minimum, at least 50 percent of the content specialists gave a perfect rating to the item.

As with the Hemphill-Westie Index there is no means for determining the statistical significance of the values for the Index of Item-Objective Congruence. However, the use of Lu's coefficient of agreement amongst the judges will give an indication of how reliable (or consistent) the judgements are. This indication of consistency of judgements along with the known values that the index would take with specific proportions of perfect rating will give the test constructor a very good idea as to how meaningful a particular I' value is for an item.

(b) Semantic Differential Technique

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

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

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

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

Directions

Given the objective and item above, your task is to make judgements on the relationship between it and the item on the adjective scales indicated below.

Scale 1: very relevant — relevant — no feeling — irrelevant — very irrelevant

Scale 2:

very important		unimportant		no feeling
	important		very important	

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

(c) A Matching Procedure

A third procedure used to obtain the judgements of content specialists involves the use of a matching task. The content specialists are presented with two lists. The first list contains a set of items. The second list is a set of objectives. The content specialists match items to objectives that they feel they measure. A contingency table can then be constructed to represent the number of times each item is assigned to each objective across the content specialists. A visual examination of a contingency table will provide information concerning the deviant items. Statistical tests can also be done.

An Empirical Study of Several Judgemental Methods

In this section, two studies used to collect the judgements of content specialists on items from two different tests designed to measure performance on an individualized science learning package will be described. In Study One, twenty-one science teachers were administered

an item validation questionnaire which was designed to determine the extent to which they thought a set of items were measures of the intended objectives. The teachers (or content specialists as we will refer to them) were asked to make judgements on forty items and eleven objectives using the Hemphill-Westie categorizing technique. Table 1 contains the expected match between the items and objectives.

In Study Two, a more complex research design and item validation questionnaire were used to obtain the judgements of content specialists on a set of forty-eight science items and twelve science objectives. The twelve instructional objectives and their matched items (see Table 2) were divided into three subgroups. Each of these subgroups (denoted subgroup one, two, and three) consisted first of four objectives and their four corresponding items for a total of 16 test items. Next, two additional objectives from the initial pool of twelve objectives, without their corresponding items, were assigned to each subgroup resulting in a final subgroup composition of six objectives and sixteen items. Finally, three different forms of an item validation questionnaire were constructed by assigning each of the three subgroups of items and objectives to one of three judgemental procedures, the Hemphill-Westie categorizing technique, the semantic differential rating tecbnique and the matching technique. All three judgemental procedures were described in previous sections. The form of each questionnaire is as follows:

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

TABLE 1

EXPECTED MATCH BETWEEN THE TEST ITEMS AND THE OBJECTIVES THEY WERE DESIGNED TO MEASURE

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

TABLE 2

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

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

Ten science teachers (not the same teachers as in Study One) were randomly assigned to complete each form of the questionnaire. Thus for any one subgroup of objectives and items, there was information available from three different groups of content specialists using three different judgemental procedures.

The data collected from both studies were examined, where appropiate, with regard to the following questions:

1. Does the judgemental data provide information which can be used to assess the extent to which an item is a measure of an instructional objective?

2. Is the information obtained reliable in the sense that there is consistency of agreement amongst the content specialists?

3. Is the data valid?

The Hemphill-Westie Categorizing Procedure

For both Studies One and Two, a decision was made to set the cutoff score for the index of item-objective congruence, the numerical representation of the Hemphill-Westie data, to be .70. That is, items having item-objective congruence indices less than .70 were identified as not being valid measures of their intended objectives. The results of the calculation of these indices are presented in Tables 3 and 4. In Study One, items 3, 4, 7, 8, 9, 10, 15, 18, 19, 20, 26, 31 and 34 were identified as not being valid measures of the intended objectives. In Study Two, items 8, 10, 13, 14, 16, 22, 23, 24, 35, 40 and 41 were identified as not being valid measures of the intended objectives.

The Hemphill-Westie procedure requires that the content specialists judge each item against all of the objectives. If an item is judged

TABLE 3

VALUES FOR THE INDEX OF ITEM OBJECTIVE

CONGRUENCE ON TEST ITEMS IN DATA SET ONE

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

TABLE 4

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

(Index/SD Statistic)

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

TABLE 4 (continued)

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

to be a measure of more than one objective, its item-objective congruence index will be lowered. For both studies, the item-objective congruence indices were always considerably higher when the items were matched to the intended objectives than when they were matched to the other objectives. It appeared that the content specialists could make' meaningful judgements in the assessment of item validi'y.

The next analyses were concerned with determining whether levels of item-objective congruence indices were based on reliable data. That is, were the content specialists consistent in their judgements of the test items? The assessment of the consistency of agreement amongst judges was made by calculating Lu's (1971) coefficient of agreement. A coefficient of agreement was obtained for each objective subgroup for both Data Sets One and Two. The results are presented in Tables 5 and 6. For all twenty-three objectives, the coefficients of agreement were signigicant. These findings support the hypothesis that the Hemphill-Westie judgemental data was reliable in the sense that there was substantial consistency of agreement amongst the judges.

For the purposes of this study, validity of the judgemental data was defined as the degree of agreement between different groups of content specialists assessing item validity through the use of different judgemental procedures. For Study One, there was insufficient data to check for validity. For Study Two, the degree of agreement was obtained by correlating two rank orderings of the items based on the sized of judgemental statistics calculated from the categorizing and rating procedures. The first rank ordering of the items was established by using values of the index of item objective congruence. The second rank order-

TABLE 5

LU'S COEFFICIENT OF AGREEMENT FOR THE

OBJECTIVE SUBGROUPS OF DATA SET ONE

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

*$p < .01$

TABLE 6

LU'S COEFFICIENT OF AGREEMENT FOR THE

OBJECTIVE SUBGROUPS OF DATA SET TWO

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

*$p < .01$

ing of the items was established by using values of the index of item objective congruence. The second rank ordering was established by using values of a statistic (SD) calculated from the semantic differential ratings on the items. This statistic was computed using the following algorithm:

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

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

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

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

The minimum value for SD is given as

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

where n is the number of content specialists,

N is the number of objectives, and

k is the highest value of the rating scale.

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

For each of the three subgroups of objectives, consisting of 16 items each, Spearman's coefficient of rank difference was calculated between the item-objective congruence indices and the item SD statistics. The three Spearman coefficients reported in Table 7 were statistically significant and above .65 , suggesting the substantial agreement as to the quality of test items across the two methods for judging items.

TABLE 7

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

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

*p<.01

The Semantic Differential Rating Procedure

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

An assessment of the reliability of these ratings was made through an examination of the standard deviations of the ratings on an item and the objective to which it was matched. With the exception of a few items these standard deviations were quite small, indicating that the content specialists were making the same ratings with respect to the item. These results are presented in Table 8.

The Matching Procedure

For the matching technique the content specialists were asked to match each item to the objective they felt it measured. The data col-

TABLE 8

SEMANTIC DIFFERENTIAL RATINGS ON THE

TEST ITEMS FROM DATA SET TWO

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

TABLE 9

CONTINGENCY TABLES FOR DATA COLLECTED FROM THE CONTENT SPECIALISTS IN THE TEST ITEMS TO THE OBJECTIVES IN DATA SET TWO

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

lected from the use of this technique is different from the data collected from the use of the other two techniques in that the content specialists were not required to judge each item on all the objectives.

An (m x N) contingency table of items (m) and objectives (N) was constructed. The mN cell frequencies consisted of the number of times a content specialists matched an item to a particular objective. Discrepancies between the expected matches and the actual matches were used to identify invalid items. A minimum criterion that seventy percent of the content specialists must have correctly matched an item to an objective before the item could be declared valid was established. Using this criterion, the results presented in Table 9 show that for Data Set Two, items 8, 25, 28, 35, 41 and 47 were identified as not having item validity. The relatively high number of correct matches is an indication that this information can be used to assess item validity.

One means for assessing the reliability of the data collected through the use of a matching technique is to calculate the amount of agreement between the expected matches and the actual standard. Light (1971) has developed a statistic (G) which provides a numerical representation of this amount of agreement which can be tested statistically for significance. However, because of the relatively small number of judgements required of the content specialists, it was not calculated for this data.

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

Summary and Conclusions

In this study, three techniques for collecting and analyzing the

judgements of content specialists as a means for assessing item validity were discussed. All three techniques were shown to provide information which could be used to ascertain if an item was a measure of an objective. However, there were differences in the types of data which were collected through the use of these techniques. For example, there were many more low SD statistics than low item-objective congruence indices for the same items. This is an indication that the content specialists when using the semantic differential rating procedure judged the items to be relevant meausurs of objectives other than the intended ones more often than when using the categorizing procedure. It appears that these two procedures are tapping different dimensions.

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

On the other hand, there are drawbacks to the use of the Hemphill-Westie procedure which could be rectified through the use of other judgemental techniques. These drawbacks are given as follows:

1. The procedure cannot be used to collect information on such topics as quality of the item, and type of distractors.

2. The dimensionality of the data must be known in advance of its use.

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

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

Basic to an effective criterion-referenced testing program is a set of test items that are "valid" indicators of the objectives they were designed to measure. Unfortunately, the matter of assessing item validity has recieved only limited discussion in the voluminous criterion-referenced testing literature. It is clear from this study that one promising approach to the item validity question is through the collection and analysis of the judgements of content specialists.

Our expectation is that the results reported in the study will provide some direction for the continued development of methodologies for the collection and analysis of content specialists' judgements.

REFERENCES

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

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

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

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

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

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

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

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

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

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

Hively, W., Maxwell, G., Rabehl, G., Sension, D., and Lund n, S. Domain-referenced curriculum evaluation: a technical handbook and a case study from the Minnemast Project. Monograph Series in Evaluation, No.1. Los Angeles: Center for the Study of Evaluation, University of California, 1972.

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

Light, R.J. Issues in the analysis of qualitative data. In R. Travers (Ed.), Second handbook of research on teaching. Chicago: Rand McNally, 1971, 318-381.

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

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

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

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

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

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

Rovinelli, R.J. Methods for validating criterion-referenced test items. Unpublished doctoral dissertation, University of Massachusetts, Amherst, 1976.

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

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