CENTER FOR THE
STUDY OF
EVALUATION

UCLA
Graduate School
of Education
Los Angeles, California

# DETERMINING HOW WELL A TEST MEASURES YOUR OBJECTIVES

Stephen P. Klein and Jacqueline P. Kosecoff

CSE Report No. 94
April 1975

# DETERMINING HOW WELL A TEST MEASURES YOUR OBJECTIVES

Stephen P. Klein
Jacqueline P. Kosecoff

CSE Report No. 94
April 1975

Program for Research on Objectives-Based Evaluation
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California

# ABSTRACT

This paper describes a procedure for in depth analysis of a limited number of tests being considered for selection by a school, district, project, or state personnel. This procedure involves listing the objectives that it would be desirable to measure, determining the relative importance of each of these objectives, having "judges" match test items to these objectives, and then correlating the relative importance of each objective with the extent to which it is covered by a test. Variations in the procedures are presented as well as adjustments for differences in average item difficulties across clusters of items and for differences in test length. There also is a discussion of how well a cluster of items assigned to an objective actually covers that objective. Finally conditions under which the procedures described are and are not applicable are considered and appendices illustrating specific directions and procedures are provided.

# DETERMINING HOW WELL A TEST MEASURES YOUR OBJECTIVES

A frequent problem for educators is finding the "best" test to use for evaluating an instructional program. What may be best for one program in a given area may not necessarily be best for another program in that area, even when both programs use the same instructional materials. There are a wide range of factors that have to be considered in any test selection procedure. Some of these factors include the presence of national norms to satisfy special score reporting requirements, the ability level(s) of the students to be tested, and the conditions under which the testing takes place. The importance of these factors varies with the reason for testing.

In many instances resource materials are available that can help narrow the range of possible measures that might be selected. The CSE test evaluation books (Hoepfner, et al., 1970, 1971, 1972, & 1974), for example, list all the published standardized measures and evaluate each test's general validity, reliability, examinee appropriateness, and administrative usability. Burros (1972) edits an extensive collection of test reviews; such reviews are authored by psychometricians and/or subject matter specialists. There also are compendiums describing the general characteristics of certain criterion referenced test systems (e.g., Klein & Kosecoff, 1973). An examination of these resource materials and the manuals supplied by test publishers usually provides sufficient information for educators to determine which two or three tests meet their general needs and requirements.

Once the number of eligible tests has been delimited, the next step is to make the final decision as to which one of them will be used. This decision process usually involves a small committee (such as teachers within a school) examining each test in terms of how well it matches and/or emphasizes the important objectives of their instructional program. In other words, the committee must determine which test is most consistent with what they are trying to teach. Sometimes this test review only involves reading what the publisher has to say about what the test contains while in other instances, the test questions themselves are actually examined. In either case, the appraisal rarely involves a systematic and objective procedure whereby the overall "goodness of fit" between each test and the program is actually measured and compared. Further, test review committees may be unduly influenced by the opinions of one or two members and/or by the persuasiveness of a publisher's representative.

## PURPOSE

The remainder of this paper describes one objective, practical, and efficient procedure whereby educators can determine how closely a test matches a program's objectives. This technique is appropriate for examining whether a test is consistent with a national educational program (such as Title VII of ESEA), a state or district adopted text, a commercial curriculum package, or even teacher developed instructional materials. It also is equally appropriate for norm- and criterion-referenced tests.

## BASIC PROCEDURAL STEPS

Step 1: Develop a list of program objectives. Most current programs and curriculum materials contain a description of their instructional objectives. These objectives should be written at a level of generality so that there are about 25-75 objectives that might have to be considered in any major area for which a given test has to be selected (such as mathematics at grade 6).

Step 2: Rate the relative importance of each objective. This usually involves committee members (such as all the grade 6 mathematics teachers in the program) independently judging each objective on a five-point scale (1 = very unimportant to 5 = very important). Instructions should require that at least a few objectives be assigned to each of the five levels of importance to ensure that relative importance of the objectives is actually judged. Appendix A contains a sample of a set of directions for making these judgments.

Step 3: Construct a table (matrix) with one row for each judge and one column for each objective in order to record the ratings of importance. The average score in a column provides an index of the relative importance of the objective in that column. This table also would permit an examination of the inter-rater agreement (either by a simple inspection or by analysis of variance; see Winer, 1962). If there is wide disagreement among the raters, then there should be some discussion of the ratings by the committee in order to achieve greater consensus as to which objectives are most important.

Step 4: Place each test question on a card. An identification number should be placed on the back of the card to indicate from which test it came and which question number it was in that test. Appendix B contains detailed set of instructions for preparing this deck of cards.

Step 5: Assign each item to the objective(s) it measures. The objectives established in Step 1 are used for this purpose; the judges may be the same or different than those involved in Step 2. If there are a large number of items to be reviewed (such as several long tests) and the test committee is limited in the time it can devote to the item review process, it would be possible to divide the work up so that just two or three members evaluate each item rather than the whole committee having to do this.

It should be noted that a given item may be assigned to more than one objective if satisfactory performance on the item requires proficiency on a small set of objectives. This kind of overlapping should be distinguished from a situation in which a series of "enroute" objectives are required for mastering a "terminal" objective. For example, the addition items in a mathematics test should be assigned only to objectives dealing with addition. They should not be assigned to the objectives involving multiplication even though one could argue that proficiency in addition is a necessary requirement for solving multiplication problems. The general rule of thumb, therefore, would be to assign an item to the "most advanced" objective it directly measures and only make multiple assignments in those instances where the objectives assessed are at an equivalent step or phase in the normal learning sequence.

Appendix C contains a sample set of directions for the assignment of items to objectives.

Step 6: Record the item-objective assignment assignments. This is done separately for each test. The purpose of this step is to determine how many items (if any) measure each of the program objectives. Two procedures for recording the data are described below.

6a: Arbitrary rule technique. This procedure involves setting an arbitrary rule as to when a test item measures an objective. For example, "When 50% or more of the judges agree that an item measures a given objective, then it is assigned to that objective." The data from this approach may be recorded by marking a "1" in the appropriate cells in Table 1 (e.g., if 7 out of 10 committee members said item #1 from test A measures objective #1, then a mark would be placed in column one, row one). The row marked "Total" contains the total number of items assigned to each objective; i.e., the number of tally marks in each column are recorded on this row.

Table 1
Item Assignment Record Form*

Test Name ______________________________

Date ___________________________

| | | Objective # | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Item # | Item ID code | 1 | 2 | 3 | 4 | ................ | K | no fit |
| 1 | | | | | | | | |
| 2 | | | | | | | | |
| 3 | | | | | | | | |
| .<br>.<br>.<br>. | | | | | | | | |
| N | | | | | | | | |
| Totals | | | | | | | | |

*A separate table would be constructed for each test being reviewed.

6b: Proportional weighting technique. This approach involves inserting in Table 1 the actual proportion of judges who assigned a given item to each objective. For example, if 7 out of 10 committee members said that item #3 measured objective #2, then a "70" would be placed in row three, column two. The row marked "total" contains the sum of the proportions for each objective. For example, if in column #2, items 3, 5, and 7 had proportions of .50, .70, and .30 respectively; then their sum would be 1.50. In essence, this means that the test contains "1.5 items measuring objective #2."

The arbitrary rule technique is relatively easy to use and explain but it may lead to including or excluding an item on the basis of a single committee member's opinion. The proportional weighting procedure, on the other hand, provides more sensitivity as to the total committee's judgment regarding the total number of items that should be assigned to each objective. The major disadvantage of this method is that it does not provide a one-to-one correspondence between an item and an objective, since fractional parts of items can be matched to several objectives. In comparing the two approaches, it would seem that the proportional weighting technique would be generally preferred for the following reasons: (1) the tallying for method 6a requires the actual proportions in order to apply the arbitrary rule, i.e., 6a requires an extra computational step since it needs essentially all the data reported in 6b (except the totals themselves) before the tally marks can be assigned; and (2), should it be necessary to indicate a single objective to which each item has been assigned, then it would be easy to impose the arbitrary rule technique, i.e., one can go from method 6b to 6a, but not vice versa. The same method should, of course, be used consistently for all the tests being reviewed.

Past experience (Carlson, 1974; Dahl, 1971; Niedermeyer, 1974) has indicated that judges generally agree on which items measure which objectives. Thus, the distinctions between the arbitrary rule and proportional weighting techniques may be of only academic interest. In the event that it is necessary to examine the degree of agreement, it could be done by merely inspecting the average proportion across all the objectives; i.e., the higher the average proportion, the greater the agreement among judges. Average proportions of .85 or more would indicate a high degree of agreement. If more precise information is needed about agreement levels, then it would be possible to compute the inter-rater reliability using an analysis of variance approach (Winer, 1961) for each objective separately. Only the items selected by at least one judge for a given objective would be used in the analysis for that objective.

Step 7: Compute the correlation between the relative importance of each objective and the number of items assigned to it. This should be done separately for each test. The higher the correlation, the greater the correspondence between the test and the instructional program.

Step 8. Count the number of objectives which are not measured by any item from a given test. The fewer the number of unmeasured items the greater the coverage of the objectives by the test.

## CONSIDERATIONS AND ADJUSTMENTS

### Minding your P's and Q's

The proportion of students passing an item is referred to as the item's "p value." The proportion of students failing an item is called the "q value."* The product of p and q yields the item's variance. For example, if 80% of the students pass an item, its variance is .16 (pq = .8 x .2 = 16). The greater an item's variance, the more influence it has on determining a student's relative score on a test. The maximum value of the variance is .25 which occurs when p=q=.5.* Thus, items with p values of about .5 have a much greater impact than those with larger or smaller p values. The reason for this phenomena is explained in standard texts on test theory (Lord & Novick, 1968).

*Note: p + q = 1, the prop of S's failing, plus the prop of S's passage equals the entire sample. Similarly q = 1 - p.

In the context of this paper, the important consideration is that items with middle difficulties (p values of .4 to .6) carry the most weight in determining a student's <u>relative position in the distribution</u> of scores on a test. Thus, if two objectives have the same number of items, the one whose items have an average p value closer to .5 will in fact have more influence on the total score on the test. Further, an inspection of standardized test manuals will often indicate wide differences in average p values across objectives within the same test (Klein, 1970). Thus, this is a real problem when the total score on a test will be used (as opposed to separate scores for each objective).

In order to adjust for differences in average p values across objectives, one must first know the p values. Such information can often be obtained by examining technical test reports supplied by the publisher. The next step would be to compute the average p and q values for the cluster of items assigned to each objective. Finally, one could multiply the total number of items assigned to an objective by its average pq before computing the correlation between the importance rating and the number item assigned to each objective. For example, if an objective had five items and their average p and q values were .7 and .3, respectively; then would multiply 5 times .21 (since pq = .7 x .3) to get an adjusted total number of items for each objective. This adjusted total would then be used in the correlations with importance ratings to provide a more precise estimate of the degree of overlap in emphasis between the test and the instructional program it is supposed to measure.

## Test Length

The correlation between the importance of an objective and the number of items assigned to it (either adjusted or unadjusted for the average item pq value) may be influenced somewhat by the length of the test; i.e., the total number of items in the test. If two or more tests are being compared and if the tests differ more than 10-20% in length, then one might be tempted to make a correction for this difference. The simplest and most immediate correction might be to transform the values in the row marked "totals" in Table 1 to their respective proportions of the total item assignment. However, this correction is unnecessary since it is an integral part of the correlation procedure.

Appendix D contains a sample worksheet and a numerical example illustrating the adjustment for average item p-q value.

## Item Coverage

It is evident from the preceding set of seven steps and the subsequent discussion that a "good test" is one which emphasizes the same objectives as those the test committee wants to achieve most and which covers all the objectives addressed by the instructional program, i.e., considers most important. There is, however, one further factor that should be considered in the review process. This consideration deals with the degree to which the cluster of items assigned to a given objective represents an adequate and appropriate sampling of the kinds of items that should be used to measure that objective. The need for addressing this issue was illustrated by a recent experience of the authors when they reviewed a set of mathematics objectives and items constructed by a school district. In the process of their review, they noted that for an objective dealing with solving word problems, essentially all the items involved solving just "time-rate-distance" problems; and, even these were limited to just solving for time (as opposed to some items involving solving for rate or distance). Thus, the items used to measure the objectives were not an adequate sample of all the items that might be reasonably constructed to measure that objective.

There is no apparent way of statistically detecting or adjusting for this consideration. It is, however, something that should be checked before the decision is made to select a test, especially when there are several items assigned to each objective. Finally, it is quite possible to get very misleading results from Step 7 (the correlation between importance of and number of items assigned to each objective) when the kinds of tests being compared are very different in their basic construction for objective coverage. For example, it would not be appropriate to compare a test having 2 items for each of 20 objectives with a test having 40 items spread over just 4 or 5 of these same objectives. It is unlikely that such situations will arise since tests containing only a very few items for each of several "objectives" generally involve highly specific objectives and such objectives can usually be grouped into more global terms to match the ones in tests having several items per "objective." The process of initial screening of potential tests on several factors also would presumably delimit the instruments chosen for in depth analysis to just the ones that are likely to be comparable in nature.

## APPENDIX A

### Sample Directions for Judging Objectives

There are two frequently applied methods for rating or judging objectives. The first method uses a rating form with which each objective can be rated against a single rating scale (e.g., 1 = unimportant, 5 = very important). The second method uses cards (each containing one objective) to be sorted according to the levels of a rating scale. Card assignments are then transcribed onto a summary sheet. The same rating scale can be used with either method; however, the first method is recommended for shorter lists of objectives or in situations with time constraints. Directions for both methods of judging objectives follow.

Directions for judging objectives using a rating form

On the rating form below the objectives of the (name) (grade level) instructional program are listed. Using the five point scale provided rate each objective according to its educational importance:

1 = unimportant

2 = not very important

3 = average importance

4 = above average importance

5 = very important

Do not base your decision on the ease or difficulty in measuring an objective or on the availability of current tests in this area. Base your judgments only on how important it is for students to achieve a given objective as part of the (name and grade) program.

Procedures:

1. Read the entire list of objectives. Select an objective that matches each of the five categories on the rating scale.

2. Rate the remaining objectives marking an "X" in the column that is labeled with the appropriate rating category. At least _____ objectives should be assigned to each category.

3. An objective can only be assigned one value. Choose the category that is the closest match to the objective's importance. There are no correct or incorrect answers. Rate objectives according to your opinion of their importance.

4. All objectives must be rated. Do not spend too much time on any objective. Leave difficult decisions to the end.

A sample rating form that might be used with this technique is depicted in Table 2.

Table 2
Sample Rating Form

______ Program Objectives

rater: _______________ date: _________________

| OBJECTIVE | unimportant<br>1 | not very important<br>2 | average importance<br>3 | above average importance<br>4 | very important<br>5 |
|---|---|---|---|---|---|
| 1. | | | | | |
| 2. | | | | | |
| 3. | | | | | |
| 4. | | | | | |
| | | | | | |

## Directions for judging objectives using a card sort*

The envelope labeled "objectives" contains a set of (#) cards. Each card contains one objective of the (name and grade) instructional program. Use the rating scale provided to rate each objective according to its educational importance. Do not base your decisions on the ease or difficulty in measuring an objective or the availability of current tests in this area. Base your judgments solely on how important it is for students to achieve a given objective as part of the (name and grade) program.

Procedures:

1. From the manilla envelope take the five envelopes marked:

1 = unimportant

2 = not very important

3 = average importance

4 = above average importance

5 = very important

Place the envelopes on a table from left to right in order of increasing importance.

2. Take the objective cards from the envelope. Read all the objectives. Select one objective that matches each of the 5 categories on the importance scale.

3. Sort the remaining objectives into the five piles. At least ______ objectives should be assigned to each pile.

4. An objective can only be assigned to one pile; choose the category that is the closest match to the objective's importance. Remember that there are no correct or incorrect answers. Sort objectives according to your opinion of their importance.

*This procedure and its directions are based on a technique presented in the CSE Needs Assessment Kit (Hoepfner, Bradley, Klein, and Alkin, 1972).

5。 All objectives must be sorted。 Do not spend too much time on any objective。 Leave difficult decisions to the end。

6. When all objectives are sorted put the objective cards into the respective envelopes。 (NOTE: the values in each envelope can be transcribed onto a summary form by a clerk after <u>all</u> the envelopes have been returned from each rater).

# APPENDIX B

## Sample Instructions for Preparing Deck of Test Question Cards

1. Obtain a set of blank cards (e.g., 3" x 5" cards). There should be one card for each item on each test being reviewed. Cards should be of the same size and color.

2. Prepare test and item identification codes and unique ID's for each item.

   test codes: Assign a single-digit numerical code to each test being reviewed (e.g., the first test in alphabetical order can be code "1," the second test "2," and so on).

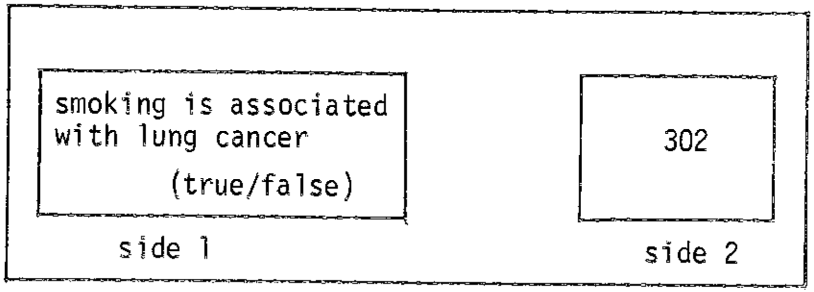
   item codes: Assign a two-digit numerical code to each item on a test*. To avoid confusion, whenever possible this code value should reflect the actual numbering of test items (e.g., the first item on a test is coded "01" and the eleventh item "11"). Note that the eleventh item of each test will have the same two-digit item code.

   unique item ID: A unique identification code can be assigned to each item by combining test and item codes. For example, the code "112" represents the twelth item on the test coded "1".

3. For each item on each test print the statement of the item on one side of a card and the identification code on the reverse side. The statement of the item should include just the test item. All identifying information that can be intuited from the item format should be omitted; however, when directions are necessary, they should be noted in lower right hand corner (e.g., "check all that apply" or "true/false"). Center the item statement and the sides of each card.

An example of a completed card for a true/false item (the second item on the test coded "3") is presented in Figure 1.

Figure 1
Sample Test Question Card



smoking is associated with lung cancer

(true/false)

side 1

302

side 2

*If a test has more than 99 items it will be necessary to have a 3-digit item code.

# APPENDIX C

## Sample Directions for Assignment of Items to Objectives

A matching form that might be associated with the sample directions is displayed in Table 3.

Table 3
Sample Matching Form

Item-Objective Matching Form

rater: ______________________ date: ______________________

| | | | | | | |
|---|---|---|---|---|---|---|
| | Program Objectives | | | | | |
| Item ID Code | 1 | 2 | 3 | ................................ | K | no fit |
| 101 | | | | | | |
| 102 | | | | | | |
| .<br>.<br>.<br>. | | | | | | |
| N | | | | | | |
| | | | | | | |

The purpose of this matching task is to identify which test items best measure the objectives of the __(name and grade)__ program. The test whose items provide the best fit to these objectives will be used to assess students' achievement in the program.

On the matching form the columns list the program objectives and the rows list the items from each test being reviewed. (Only a short summary of each objective and the ID codes for test items are printed on the form.) Use this form to match each item with the objective(s) it measures.

Procedures:

1. Remove the list of objectives from the manilla envelope. This list contains a complete statement of each of the (name) program objectives.

2. Remove the envelope containing test question cards. Each card has the statement of the test item on one side and an identification code on the reverse side. Read the test item statements.

3. Match each test item (specified on a test question card) to the objective it measures by marking an "X" in the box assigned to the appropriate item and objective combination. For example, if Item 302 measures the fourth objective, and "X" would be placed in the box formed by the fourth objective column and the row labeled "302."

4. An item can be associated with more than one objective if satisfactory performance on the item requires proficiency on several objectives. This kind of overlapping should not be confused with "en route" objectives. For example, an item dealing with addition should not be matched to arithmetic and multiplication objectives, even though addition skills are required for multiplication. This item should only be matched with the addition objective. Similarly, a multiplication item should only be matched with multiplication objectives and NOT with addition objectives. The rule is to assign an item to the "most advanced" objective that it measures directly. Assignments of an item to several objectives should only occur when the objectives are all measured by the item and are in an equivalent step or phase in the normal learning sequence.

5. If an item does not fit well with any objective, choose the closest objective. If there still is no objective which can be fit to the item, then an "X" should be placed in the last column (marked "no fit") of the appropriate row.

6. All items must be matched to at least one objective or the "no fit" category. Do not spend too much time on any item. Leave difficult decisions to the end.

# APPENDIX D

## Sample Worksheet and Example for Adjusting Item Assignments for p-q Values

A sample worksheet for adjusting item assignments for p-q values is illustrated in Table 4.

Table 4
Sample Worksheet for Adjusting
Item Assignments for p-q Values

Item Assignment Record Form
with Adjustments for p-q

test: ______________________ date: ____________________

| | Objective # | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item # | 1 | | | 2 | | | | K | | | no fit | | |
| | wt | p | q | wt | p | q | ......... | wt | p | q | wt | p | q |
| 1 | | | | | | | | | | | | | |
| 2 | | | | | | | | | | | | | |
| 3 | | | | | | | | | | | | | |
| .<br>.<br>.<br>n | | | | | | | | | | | | | |
| total | | | | | | | | | | | | | |
| mean | ▨ | | | ▨ | | | | ▨ | | | ▨ | | |
| adjusted total | | ▨ | ▨ | | ▨ | ▨ | | | ▨ | ▨ | | ▨ | ▨ |

To complete this worksheet three kinds of information are required for each item-objective assignment, the item's weight, p value and q value.

As in Table 1, the weight of an item for a given objective is the value placed in the box formed by the item and the objective. This value will vary according to which item assignment technique is used, the arbitrary rule or the proportional weighting technique. Recall from sections 6a and 6b, for the first technique a weight of "1" is assigned to an item each time it is matched with an objective. For the second technique an item's weight on an objective is the proportion of judges who assigned the item to the objective. For this first technique the only possible non zero weight is "1" while for the second technique weights can assume fractional values as well. The p and q values are the average proportion of students passing and failing the item.

The following procedure outlines how to compute adjusted total scores using the sample worksheet in Table 4. Data used to illustrate this procedure can be found in Table 5 on the following page.

Table 5

Sample Data for Computation of Adjusted Scores
(arbitrary rule technique is being used)

Item Assignment Record Form with Adjustments for p-q

test: ______________________ date: ______________________

| Item # | Objective # | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | | | 2 | | | 3 | | | 4 | | |
| | wt | p | q | wt | p | q | wt | p | q | wt | p | q |
| 1 | 1 | .2 | .8 | | | | | | | | | |
| 2 | 1 | .3 | .7 | 1 | .1 | .9 | | | | | | |
| 3 | 1 | .5 | .5 | | | | | | | | | |
| 4 | | | | 1 | .2 | .8 | | | | | | |
| 5 | 1 | .6 | .4 | | | | | | | | | |
| 6 | | | | 1 | .5 | .5 | | | | | | |
| 7 | | | | 1 | .3 | .7 | | | | | | |
| 8 | | | | 1 | .4 | .6 | | | | | | |
| .<br>.<br>.<br>N | | | | | | | | | | | | |
| total | 4 | 1.6 | 2.4 | 5 | 1.5 | 3.5 | | | | | | |
| mean | ▨ | .4 | .6 | ▨ | .3 | .7 | ▨ | | | ▨ | | |
| adjusted total | 9.6 | ▨ | ▨ | 1.05 | ▨ | ▨ | | ▨ | ▨ | | ▨ | ▨ |

Procedures:

1. Assign all items to objectives as described in sections 6a and 6b. This requires assigning an item weight to each objective. (Note that the arbitrary rule technique is reported to have been used in Table 5 and consequently all non-zero item weights have the value "1.")

2. Record the p and q values of the items matched to each objective. These p and q values should be obtained from the publisher (usually in a technical manual). [The average p values for objective #1 are .2, .3, .5 and .6. The average q values for objective #2 are .9, .8, .5, .7, .6.]

3. Sum the item weights for the items assigned to each objective. (The sum of item weights for objective #1 is 4 and for objective #2 is 5.)

Sum the p values for the items assigned to each objective. (The sum of the p values for objective #1 is 1.6 (1.6 = .2 + .3 + .5 + .6) and for objective #2 is 1.5.)

Sum the q value for the items assigned to each objective. Note that the sum of the q values for a given objective is the sum of the item weights minus the sum of the p values. (The sum of the q values for objective #1 is 2.4 (2.4 = .8 + .7 + .5 + .4 = 4 - 1.6) and for objective #2 is 3.5.)

For each objective record the sum of the item weights, p values and q values in the row labeled "totals."

4. For each objective divide the total p value by the total item weight to get the mean p value. The mean p value for objective #1 is .4 (4 = 1.6 ÷ 4) and for objective #2 is .3).

For each objective divide the total q value by the total item weight to get the mean q value. (Note that the mean q value is equal to 1 minus the mean p value. The mean q value for objective #1 is .6 (.6 = 2.4 ÷ 4 = 1 - .4) and for objective #2 is .7)

Record the mean p and q values for each objective in the row labeled "mean."

5. For each objective obtain the adjusted total by multiplying the total item weight by the product of the mean p value and the mean q value. That is: adjusted total = (total item weight) x (mean p value) x (mean q value). (The adjusted total for objective #1 is 9.6 (9.6 = 4 x .4 x .6) and for objective #2 is 1.05.) Record this value in the row labeled "adjusted total."

# REFERENCES

Buros, O. K. (Ed.) *Mental measurement yearbook* (7th ed.). Highland Park, New Jersey: Gryphon Press, 1972

Carlson, D. Personal communication, 1974.

Dahl, T. Toward an evaluative methodology for criterion-referenced measures: Objective-item congruence. CSE Working Paper No. 15. Los Angeles: Center for the Study of Evaluation, University of California, May 1971.

Hoepfner, R., et al. *CSE elementary school test evaluations.* Los Angeles: Center for the Study of Evaluation, University of California, 1970.

Hoepfner, R., et al. *CSE/ECRC preschool kindergarten test evaluations.* Los Angeles: Center for the Study of Evaluation, University of California, 1971.

Hoepfner, R., et al. *CSE-RBS test evaluations: Tests of higher-order cognitive, affective, and interpersonal skills.* Los Angeles: Center for the Study of Evaluation, University of California, 1972.

Hoepfner, R., et al. *CSE secondary school test evaluations.* (3 Vol.) Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Hoepfner, R., Bradley, P. A., Klein, S. P., & Alkin, M. C. *CSE elementary school evaluation KIT: Needs assessment.* Boston: Allyn and Bacon, 1972.

Klein, S. P. Evaluating tests in terms of the information they provide. *Evaluation Comment*, 1970, *2*(20), 1-6. ED 045 699

Klein, S. P., & Kosecoff, J. P. Issues and procedures in the development of criterion referenced tests. TM Report 26. Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurement, and Evaluation; Educational Testing Service, September 1973.

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

Neidermeyer, F. Personal communication, 1974.

Winer, B. J. *Statistical principles in experimental design.* San Francisco: McGraw-Hill, 1962.