DOCUMENT RESUME

ED 194 584 TM 800 713

AUTHOR Roth, Rodney
TITLE Evaluating Validity Data from a State Assessment Test.
PUB DATE Sep 80
NOTE 7p.: Paper presented at the Annual Meeting of the Evaluation Network (6th, Memphis, TN, September 29-October 1, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Education: *Minimum Competency Testing: *Models: *Probability: Sampling: State Programs: *Student Evaluation: *Teacher Participation: Testing Programs: *Test Validity
IDENTIFIERS Arkansas: Chi Square Test: *Teacher Expectations

ABSTRACT

In 1979-80, the Arkansas minimum competency tests were administered to a sample of 5,000 students in grades 3, 6, and 8. To determine how well test objectives matched the curriculum, their teachers estimated how many of the four items per objective a randomly selected student would answer correctly. Because chi square test comparisons of teacher estimates with actual responses were statistically questionable, a probability model was applied instead. The probability of obtaining 0, 1-2, or 3-4 correct answers by chance alone for the four items per objective was calculated, to obtain a validity coefficient. Results were presented by a 3 by 3 contingency table. Two applications of the model yielded validity coefficients of .78 and .98. An objective was considered invalid only when a large number of teachers predicted 0-1 correct and many students achieved at that level. (CP)

EVALUATING VALIDITY DATA FROM A STATE ASSESSMENT TEST*

Rodney Roth, Director
Bureau of Field Studies and School Services
University of Arkansas

The Arkansas Department of Education developed minimum performance tests in reading and math for Grades Three, Six, and Eight, during the 1979-80 school year. Since this was the first year of the minimum performance testing program in Arkansas, a major goal for the program was to develop reliable and valid tests before the tests are administered to all the students in Grades Three, Six, and Eight during the 1981-82 school year.

The minimum performance test items were based on the objectives in a booklet entitled "Basic Educational Skills", published by the Arkansas Department of Education. Arkansas teachers wrote the test items to match the objectives. Four items were written for each objective on a test form. A student answering three or four of the four items correctly was considered to have mastered the objective.

The Third Grade test has 16 math and 23 reading objectives for a total of 156 items. The Sixth Grade test has 16 math and 26 reading objectives for a total of 168 items. The Eighth Grade test has 23 math and 22 reading objectives for a total of 180 items.

The tests were administered to a random sample of about 15,000 students during April, 1980. Approximately 5,000 students were selected at each of the three grade levels. The sampling plan was in two stages and was designed to insure proportionality among five geographic regions in the state. The

* Paper presented at the annual meeting of the Evaluation Network, Memphis, Tennessee, September 29 - October 1, 1980.

first step in the sampling was a random selection of schools. After the schools had been selected, the schools were requested to send student lists for the selected grade level. From these lists, the actual students to be tested were randomly selected by the Arkansas Department of Education.

In addition to the sample of students responding to the tests, each teacher who had students selected for testing was asked to respond to a Teacher Survey. The Teacher Survey was a process that asked a teacher to estimate the level of mastery on each tested objective for a randomly selected student. The principals of the teachers who had students selected for test administration randomly selected one student for each teacher. The teachers were instructed to read each test objective and the four items measuring the objective. They were then instructed to indicate how many of the four items the selected student would answer correctly. The options were zero or one of the four items correctly, two items, and three or four items.

The estimations by the teachers were compared to the actual number of items the selected students answered correctly. This procedure produced a 3x3 contingency table per objective as presented in Figure 1.

FIGURE 1.

		TEACHER RESPONSES 0 - 1	2	3 - 4
STUDENT RESULTS	0-1	1	2	3
	2	4	5	6
	3-4	7	8	9

N

This table enables a person to compare the teachers' performance estimations with the actual student results per objective.

For example, the number in Cell Five is the number of students the teachers

estimated would answer two of the four items correctly and who did in fact correctly answer two for an objective.

A primary purpose of the Teacher Survey was to obtain validity data about the objectives assessed on the test. These data would help indicate if specific objectives were "out of sequence" with the actual curriculum sequence in the local schools. A basic assumption for the Teacher Survey was that teachers would respond with three or four correct if the objective had been taught and mastered by the student or if the objective had been at least exposed to the students, the teacher would probably respond with two correct, or if the objective had not been part of the curriculum or the specific student did not know the objective, the teacher would respond with a zero or one correct.

A primary purpose of this paper is to present a method, with a rationale, for evaluating the Teacher Survey results.

A "traditional" way to evaluate the Teacher Survey results might be with a Chi-square test. This procedure would indicate if there was a significant association between the teachers' estimations and the students' results. One could also compute a Cramer's V in order to find the apparent strength of the association.

The following actual results for an objective from the Teacher Survey presents some problems with traditional statistics.

A sixth-grade objective is Using a Telephone Directory. The results for this objective are presented in Table 1.

TABLE 1.
TEACHER RESPONSES

STUDENT RESULTS	0 - 1	2	3 - 4	
0 - 1	0	1	4	5
2	0	5	7	12
3 - 4	3	20	122	145
	3	26	133	162

The Chi-square value for this table was 6.67 with a p more than .10. The Cramer's V was .014.

The results presented in Table 1. indicate a tremendous problem with using Chi-square as a way to analyze these data to help determine validity of test objectives.

One problem is a "rule of thumb" guide for using Chi-square. This rule states that the expected frequencies of each cell should be at least five. This is not the case in Table 1. and in many other Teacher Survey tables.

An additional problem with the Chi-square test for the results in Table 1 is that the Chi-square value indicates non-significance or a lack of statistical association. An examination of the data, however, would lead one to feel quite confident about the validity of this objective for sixth graders. This is based on the fact that 90 percent (145/162) of the students mastered the objective. Furthermore, the teachers estimated that 82 percent (133/162) of the students would master the objective. The teacher also accurately predicted mastery for 75 percent (122/162) of the students.

This writer will now present an alternative model for evaluating the validity data. The model is based to a large extent on probability theory. Each objective on the test was measured by four items and each item had four responses. The probabilities for obtaining correct answers by chance alone for the four items per objective are:

None or one correct = .74

Two correct = .21

Three or Four correct = .05

The proposed model for evaluating the validity data is the ratio of the sum of Cells 2 - 9 (See Figure 1.) to total N. This ratio should be quite large

for a valid objective. (One could also use the ratio of Cell 1 to total N. In this case, the ratio should be quite small.)

The rationale for the model is based on the teacher judgement plus the actual results. The sum of Cells 7, 8, and 9 represents the number of students who "mastered" the objective. This sum would be quite accurate for student mastery since the probability for 3 or 4 correct by chance is extremely small.

The sum of Cells 3 and 6 represents the number of students predicted to master the objective, but didn't reach mastery. These two cells should, however, be included since teachers believe that mastery will occur.

Cells 4 and 5 represent partial mastery by the students. These two cells should also be summed for validity purposes since the probability of answering two of the four items by chance is also small. In other words, the students have probably been at least exposed to the objective. Furthermore, Cell 2 should be included in the summation. This cell indicates that teachers feel that some exposure to the objective has taken place.

An example for using the evaluation model is presented in Table 2.

TABLE 2.
TEACHER RESPONSES

STUDENT RESULTS	0 - 1	2	3 - 4	
0 - 1	22	10	4	36
2	21	18	5	44
3 - 4	26	40	21	87
	69	68	30	167

The results presented in Table 2 are from a sixth grade objective concerning centigrade temperatures.

At first glance it might appear that the objective was "out of sequence" with the sixth grade curriculum since the teachers predicted that 41 percent

of the students would answer correctly zero or one item. The actual number answering 0-1 was only 22 percent. Furthermore, the teacher predicted mastery for only 18 percent while 52 percent actually mastered the objective.

An application of the proposed validity evaluation model would yield a ratio coefficient of .87. This value is considered quite "acceptable".

The data presented in Table 2 indicated that the teachers had underestimated the students' performance. The data presented in Table 3 is an example of teacher overestimation. The model for validity, however, does apply for this situation.

TABLE 3.

		TEACHER RESPONSES			
		0 - 1	2	3 - 4	
STUDENT RESULTS	0-1	4	9	17	30
	2	4	17	40	61
	3-4	3	15	58	76
		11	41	115	167

The results in Table 3 are for a sixth grade objective concerning length measurement.

The teachers obviously feel that the majority of students will master this objective since the predicted rate was 69 percent. The percent of students mastering the objective was only 46 percent. The application of the validity model would indicate an extremely valid objective since the ratio coefficient is .98.

In conclusion, this paper has presented a model for evaluating validity data from a state assessment program. The model combines teacher estimations of student performance and actual student performance levels in order to obtain a validity coefficient. An objective should be considered not valid only in the situation where a larger number of teachers predict "no" achievement (0-1 correct) on the objective and a large number of students do in fact achieve at the 0-1 performance level.