DOCUMENT RESUME

ED 074 150 TM 002 515

AUTHOR Millman, Jason
TITLE Psychometric Characteristics of Performance Tests of Teaching Effectiveness.
PUB DATE Feb 73
NOTE 12p.; Paper presented at annual meeting of the American Educational Research Association (New Orleans, Louisiana, February 25-March 1, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Effective Teaching; *Performance Tests; *Psychometrics; Scoring; Speeches; Teacher Evaluation; Technical Reports; Testing; Test Reliability; Test Validity

ABSTRACT

Teaching performance tests are measures which assess a teacher's ability to accomplish prespecified instructional objectives. Although possessing much face validity, little psychometric information is available about such assessment devices. Three separate studies were conducted to provide information about the validity, reliability, administration, and scoring of performance tests of teaching effectiveness. (Author)

1922

Psychometric Characteristics of Performance Tests of Teaching Effectiveness[1,2]

Jason Millman
Cornell University

The importance of a valid way of measuring teaching effectiveness is obvious. The performance test approach described in the previous paper by W. James Popham has much appeal because of its face validity -- effectiveness is measured by changes in student behavior. Optimum use of performance tests of teaching effectiveness as criteria and dependent variables, however, requires much more information than presently exists about the psychometric characteristics of these measures. The purpose of this paper is to describe the results of three investigations designed to get some information on such characteristics.

[1] Paper presented at the annual meeting of the American Educational Research Association, New Orleans, February 1973.

[2] The cooperation of the Pasadena Public Schools is gratefully acknowledged.

STUDY 1

282 Students and 34 student teachers in an urban school participated. Students were in grades 4, 5 and 6. 45% Of the students were black, 17% represented Spanish and other nonblack minority groups.

Instructional Materials

Three lessons were employed. Their titles and the statement of the instructional task as given to the teacher follow:

Light Rays. Given a diagram containing parallel light rays, air, and glass, the student will indicate whether the light rays will tend to converge or tend to spread apart by circling the correct term.

Adjectives. Given sentences containing an underlined adjective, the pupil will designate which sentences contain specific adjectives by circling a letter S placed before the sentence.

Folkways (short lesson). When presented with a list of folkways, the pupil will be able to designate which folkways are mores and which are not.

Folkways (long lesson). When presented with a list of folkways, the pupil will be able to designate which folkways are mores and which are not, and which are primary and which are secondary.

Instruments

A cognitive and an affective criterion test was constructed for each lesson. The cognitive criterion test consisted of items consistent with the instructional objective and similar to the sample questions provided the student teacher. The affective tests for the two Folkways lesson were identical and differed from the affective test for the other two lessons only in the reference to the name of the lesson. The affective tests solicited each student's opinion (a) of the lesson and (b) of the teacher. These two types of questions were combined into a total score when two separate factors failed to appear.

A number of "control" measures were also available. These included a test of liking for school (e.g., Do you like staying at home and watching T.V. better than going to school?), a general vocabulary test, and separate measures for each lesson designed to measure skills related to, but distinct from, the cognitive skills to be taught. Also available for each child was his sex, grade level, race, and a school rating of general ability expressed on a 3-point scale.

Design

The variables were: Time for instruction (15 minutes); Class size (4-6 pupils, 9-14 pupils); Trial -- i.e., attempt at teaching a minilesson (1, 2, 3); and minilesson (light rays, adjectives, folkways short, folkways long). No child was taught the same lesson twice nor (in Study 1) did the student teacher teach the same lesson twice. The teacher taught the same children for the first two trials and a different group of children for a third trial. Assignment of student to group was supposedly on a stratified random basis (grade and academic level rating the stratifying variables).

Calculation of Teaching Effectiveness Measure

The adjusted mean scores of students being taught a lesson (by a student teacher) on the affective and cognitive criterion tests were the teaching effectiveness scores for that teacher. The mean scores were adjusted (using analysis of covariance) for initial differences on control measures and for the unreliability of the control measures. Based on preliminary multiple-regression analyses in which the control measures were used to predict the criterion measures, only one to three control measures were used to adjust any one of the criterion measures. These effectiveness measures were standardized to a mean of 50 and standard deviation of 10.

Results

A. Criterion Test Reliabilities and Multiple R with Predictors

	Cognitive		Affective	
Lesson	KR20	R	KR20	R
Light Rays	.47	.16	.58	.20
Adjectives	.41	.45	.54	.15
Folkways (Short)	.72	.35	.71	.28
Folkways (Long)	.56	.56		

Since a teacher's effectiveness score was the mean score on the test of students being taught by him, the functioning reliability of the dependent variable is higher than the values shown. This is because group means are, in general, more reliable than an individual's score.

The low values of R suggest that the control measures were unsuccessful in equating pupil differences existing prior to the insturction.

B. Test-Retest Teache ffectiveness Reliability Data

Since teachers taught more than one lesson and since students were taught more than once, it was possible to correlate effectiveness scores over trials. Specifically, the (a) row in the table below indicates that a teacher's effectiveness score on the cognitive criterion when she taught one lesson was essentially uncorrelated (.02) with her corresponding score based on another minilesson taught by her to the same group. Although 24 paired observations were involved in the calculations, correlations were computed separately for each pair of minilessons and, consequently, the degrees of freedom associated with the mean r of .02 is only 18.

	Cognitive Criterion Tests			Affective Criterion Tests		
	N	df	r	N	df	r
(a) Both lessons taught to the SAME pupils (same teachers)	24	18	.02	24	18	.34
(b) Both lessons taught to DIFFERENT pupils (same teachers)	38	28	-.03	42	32	.31
(c) Both lessons taught to the SAME pupils (different teachers)	33	23	.34	35	25	-.20

The correlations in row (a) were predicted to be the highest; those in row (c) the lowest. The value of .34 in row (a) and more especially in row (c) differences in pupil ability prior to instruction were not controlled. The value of .02, however, is perplexing.

C. Analysis by Teaching Trial

The mean effectiveness scores analyzed by trial are shown in the table below. Data are for the 21 teachers who taught all three minilessons.

	Cognitive			Affective		
Trial	N	Mean	S.D.	N	Mean	S.D.
First	21	50.4	11.5	21	48.9	8.5
Second	21	52.6	7.9	21	50.6	10.2
Third	19*	50.3	10.9	21	49.3	11.6

*Students in two groups were not administered control tests for the cognitive criterion test so adjusted teaching effective scores could not be computed.

Teaching trial had a negligible effect in this study.

D. Analysis by Length of Teaching Time

The design permitted a comparison between effectiveness scores when teachers were given a short time to teach (15 minutes) and scores when they were given a longer teaching time (30 minutes). Since a different criterion test was employed for the Folkways lesson under the two time conditions, the short-time/long-time comparison is possible only for the other two lessons. Results are below.

	Cognitive Criterion						Affective Criterion					
	Short			Longer			Short			Longer		
Lesson	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Light Rays	10	47.6	6.9	17	51.4	11.3	10	47.6	11.4	17	51.5	8.8
Adjectives	21	50.1	10.2	8	50.0	9.4	21	49.7	10.5	8	50.9	8.3
Total	31	49.3	9.3	25	51.0	10.7	31	49.0	10.9	25	51.3	8.6

The extra teaching time did not result in increased effectiveness scores of teachers for the adjectives lesson. There was about .4 of a standard deviation difference on the light rays lesson. This difference was not statistically significant.

E. Analysis by "Class" Size

The design permitted a comparison between effectiveness scores when teachers taught larger groups (usually 8-12) and scores when teachers taught smaller groups (usually 4-6). Results are shown below.

	Cognitive Criterion						Affective Criterion					
	Small Class			Larger Class			Small Class			Larger Class		
Lesson	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Light Rays	13	45.8	5.9	14	53.9	11.4	13	49.9	11.4	14	50.1	8.5
Adjectives	14	46.9	11.9	15	53.0	6.6	14	51.0	8.1	15	49.0	11.3
Folkways	10	49.2	12.1	11	50.8	7.6	11	53.9	9.1	12	46.2	9.3
Total	37	47.1	10.4	40	52.7	8.9	38	51.5	9.8	41	48.6	10.0

The larger class size was associated with higher performance on the cognitive tests (about one-half of a standard deviation), but with somewhat lower ratings on the affective scales. For all lessons combined, the difference of 5.6 points on the cognitive criterion was significant ($p<.05$); the corresponding difference of -2.9 points on the affective criterion measure was not significant.

F. Analysis by Familiarity with Content

Minilesson instructors were given the following question: "Pretend you were given, as a test, items like the sample test items but were not allowed to read the explanatory material. How well do you think you would do on such a test? (a) Poorly. I'd have to guess at the items. (b) O.K. I would know some and miss some. (c) Well. I was familiar enough with the ideas ahead of time that I would probably have answered virtually all the items correctly."

The data in the table below do NOT support the hypothesis that instructors who rate themselves as knowledgeable about the content they teach will have higher effectiveness scores than instructors who do not rate themselves as high.

	Cognitive Criterion			Affective Criterion		
	N	Mean	S.D.	N	Mean	S.D.
(a) would do POORLY	27	48.4	10.5	27	50.4	9.5
(b) would do O.K.	32*	51.4	10.8	33	50.1	9.8
(c) would do WELL	18*	50.1	7.0	19	49.1	10.9

*Students in two groups were not administered control tests for the cognitive criterion test so adjusted teaching effectiveness scores could not be calculated.

G. Analysis of Effect of Instruction on CONTROL Test Performance

All control tests were administered to students after they had instruction. This was done for administrative convenience, as students only had to be tested in one sitting (for each minilesson). However, this can be a "dangerous" practice, because the instruction might change the test performance and the very effects one wants to measure might be "controlled away." To see if control test performance was influenced by instruction, 4 or 5 students from each of 9 classrooms were given two control tests prior to any instruction. The two tests chosen were the control test on attitudes toward school and the control test for the folkways lesson -- the control test asking students which of a number of practices were laws and which were not laws. Results are shown below.

	Attitude Control Test			Folkway Control Test		
Time of Testing	N	Mean	S.D.*	N	Mean	S.D.*
Before Instruction	38	4.84	1.26	19*	7.95	.97
After Instruction	38	4.39		19*	5.95	

*Of the difference scores.
**There are fewer cases here because only half of the students taking this control test prior to instruction were taught the folkway minilesson.

The differences are of statistical and practical significance. Although these two control tests were the ones judged most susceptible to the effect being studied, and although the change of scores does not mean that the several teachers would be affected differentially, the results do suggest that caution be exercised if this tactic of administrative convenience is to be used.

STUDY 2

155 Students and 18 student teachers in an urban school. 144 students were in grade 3; 11 in grade 2. 34% Of the students were Black, 18% represented Spanish and other nonblack minority groups.

Instructional Materials

Two "minilessons" were employed. The titles and the statement of the instructional task as given to the teachers follow.

Decoding. Given a list of "words" containing symbols that have a sound and symbols that do not have a sound, the children will be able to circle the "words" which have a SHORT sound. (Using nonsense symbols, the task was similar to that for determining long and short vowel sounds in printed words that follow consonant-vowel-consonant and consonant-vowel patterns.)

Rhythms. Given an orally presented sentence or verse having either an iambic or dactylic rhythm, the children will be able to distinguish between the meters by circling a picture of either a balloon or a bumblebee.

Instruments

A cognitive and an affective criterion test was constructed for each lesson. The cognitive criterion test consisted of items consistent with the instructional objective and similar to the sample questions provided the student teacher. The affective tests for the two lessons were identical except when referencing the name of the lesson. The affective tests solicited each student's opinion (a) of the lesson and (b) of the teacher. These two types of questions were combined into a total score when two separate factors failed to appear.

"Control" measures were also available. These included a test of liking for school (e.g., Do you like staying at home and watching T.V. better than going to school?), and separate measures for each lesson designed to measure skills related to, but distinct from, the cognitive skill being taught. Also available for each child was his sex, race, and a 3-point rating of general ability made by school personnel.

Design

The plan was for each teacher to teach three times -- one lesson twice and the other lesson once. Further, each teacher was to teach the same group twice and another group once. Only 12 of the 18 teachers completed all three lessons.

Assignment of students to groups was on a stratified random basis with academic level rating the stratifying variable. Group size was limited to 6, with the students in any one group coming from different homerooms. Kept constant was instructional time (15 minutes).

Calculation of Teaching Effectiveness Measures

The same procedure as employed in Study I was used. This involved computing the mean score of students in a group on the criterion measures and adjusting these means for initial group differences on the "control" measures.

Results

A. Criterion Test Reliabilities and Multiple R with Predictors

	Cognitive		Affective	
Lesson	KR20	R	KR20	R
Decoding	.38	.76	.50	.16
Rhythms	.18	.46	.43	.18

The test reliabilities are barely satisfactory and the multiple correlations predicting the affective criteria are unsatisfactory. The multiple correlations predicting the cognitive criteria are surprisingly high given the KR20 values of the dependent variables.

B. Test-Retest Teacher Effectiveness Reliability

Because of the design used, it was possible to correlate effectiveness scores under three conditions -- when the teacher taught the same lesson both times (to different groups of students, of course); when different lessons were taught the two times, but the same group of students were used; and finally when different lessons and different students were involved in the two teaching trials. The correlations shown below are actually means of several correlations -- each computed on a small number of teachers having similar teaching experiences. For example, a separate correlation was computed for the three teachers who taught the decoding lesson both times as their first and second teaching attempts.

	Cognitive Criterion Tests			Affective Criterion Tests		
	N	df	r	N	df	r
Same lesson, different students	12	6	-.23	11*	5	.26
Different lessons, same students	12	6	.71	12	6	.18
Different lessons, different students	12	8	.09	11*	7	-.22

*In one group of students, the control affective measure was not administered, and thus a teaching effectiveness score could not be computed.

C. Analysis by Teaching Trial

The data are based on the 12 teachers who taught the two minilessons a total of three times. One teacher whose group did not take a "control" test in the affective area was excluded.

	Cognitive Criterion Tests			Affective Criterion Tests		
	N	Mean	S.D.	N	Mean	S.D.
A teacher's first trial	12	47.3	6.7	11	50.8	9.1
A teacher's second trial	12	50.1	11.1	11	54.5	7.9
A teacher's third trial	12	53.8	12.2	11	47.1	10.3
The 1st time a teacher taught a minilesson	12	47.3	6.7	11	50.82	9.1
The 2nd time the teacher taught the SAME minilesson	12	53.3	12.2	11	51.73	10.6
1st time a group of students taught	12	47.8	9.6	11	55.3	7.6
2nd time the same group of students taught	12	51.5	10.9	11	49.7	8.6

The experience of previously teaching the same minilesson on the same group is associated with higher effectiveness scores on the cognitive measures. Familiarity with the students was associated with lower ratings on the affective criterion, an observation noted previously by John McNeil.[3] Although suggestive, none of the differences is statistically significant at $p<.05$ two-tailed.

D. Analysis by Familiarity of Content

As was done in Study 1, instructors were asked how well they would do on the posttest type of questions if they had not been given background information about the subject. On the decoding lesson, the posttest consists of "words" which use "nonsense" type symbols, and all teachers would have to guess on such questions if the arbitrary key was not given.

RESULTS FOR RHYTHM TEST ONLY	Cognitive Criterion Test			Affective Criterion Test		
	N	Mean	S.D.	N	Mean	S.D.
(a) Would do poorly	1	58.0	-	1	67.0	-
(b) Would do O.K.	6	47.5	9.4	6	47.2	10.55
(c) Would do well	6	50.7	10.5	6	48.8	6.84

Note -- scores reported above do not have means and S.D.s of 50 and 10 because the instructor's second trial results are not shown.

Again, these ratings do not seem particularly related to the teacher's effectiveness score.

[3] person communication.

STUDY 3

Subjects

A convenience sample of 58 teachers from Long Beach, California and Hawaii. Students were elementary-aged children.

Instructional Materials and Instruments

The light rays, folkways (long), rhythms, and decoding lessons and criterion tests (see studies 1 and 2 above) were used. In many cases, control tests were not administered.

Design

No planned design was followed. In a few cases, teachers taught more than one minilesson with relatively little confounding between order and minilesson.[4] Data are reported only for teachers instructing with more than one minilesson. Class sizes ranged from four to six with the vast majority of them being six in number. The method by which students were assigned to teachers is not known.

Calculation of Teaching Effectiveness Measures

When control test data were available and correlated at least .15 with criterion tests, mean class scores were adjusted in accordance with the procedures described for Study 1.

Results

A. Test-Retest Teacher Effectiveness Reliability

Cognitive Criterion Tests			Affective Criterion Tests*		
N	d.f.	$\bar{r}$	N	d.f.	$\bar{r}$
13	9	.05	12	8	-.22

*Effectiveness scores not adjusted.

[4] The standard scores reported do not have a mean of 50 (and S.D.=10) because the subsample on which multiple minilesson data were available is not exactly representative of the total sample of which the standardization of scores were based.

C. Analysis by Teaching Trial

	Cognitive Criterion Tests			Affective Criterion Tests		
Trial	N	Mean	S.D.	N	Mean	S.D.
First	17	44.3	10.9	18	47.3	8.2
Second	17	53.6	8.8	18	48.3	10.5

cognitive The tendency found in Study 2 of improved effectiveness scores on the/criterion measures with practice was evident here also ($t_{16}=2.62$, $.01<p<.02$).

*　*　*　*　*　*

CONCLUDING THOUGHTS

To most teaching performance test advocates, the most disturbing results reported in these studies are the erratic and low test-retest reliabilities. Although one reviewer of previous data on this question concluded that results from such studies were not inconsistent with a hypothesis of zero reliability,[5] more encouraging findings were reported in one of the larger, better designed studies.[6]

Why the low reliabilities? Several possible answers come to mind.

Reason: The performance test results may be reflecting the true state of affairs; teaching is not a generalizable act. Reply: It is true that like happiness, anxiety and other traits that vary over occasions, it is no doubt the case that a teacher's "real" effectiveness varies with the situation. Nevertheless, one could have hoped to obtain a more interpretable pattern of correlations in which reliabilities would be highest when the situations (e.g., minilesson being taught, student group involved) were most similar.

Reason: The abilities and attitudes of the learners are what really make the difference--the teacher's input accounts for a negligible amount of the variance. Reply: We obviously were not able to control adequately the pretreatment individual differences variables, but even when the same students were involved in the test-retest correlations, results were erratic.

Reason: The conduct of the investigation during the data gathering phases was faulty. Reply: There is a risk involved when a study is being carried out 2,600 miles away from the principal investigator. Except for study 3 which was never intended to be a controlled investigation, to the best of my knowledge the data were collected in accordance with the planned design.

[5] Glass, Gene V. Statistical and measurement problems in implementing the Stull Act. Stanford, California: Stanford University Invitational Conference on the Stull Act, 1972.

[6] Belgard, M., et al. Pages 182-209 in Westbury and Bellack (Eds), Research into classroom processes: recent developments and next steps. New York: Teachers College Press, 1971.

Reason: The analysis lacked power in the statistical sense of the word. Reply: Clearly the erratic results are due in large measure to the few teachers involved in most comparisons. For example, the 95% confidence interval for a correlation of 0 computed on 8 teachers exceeds ± .60.

Reason: The control and criterion instruments were not that reliable. Reply: It may be that with more reliable measures utilizing more items collected on larger students groups after longer instructional sessions that the teaching performance tests will be a more reliable indicator of teaching effectiveness.

Clearly more definitive work is needed on teaching performance tests. It is to this end that my future empirical work is directed.