DOCUMENT RESUME

ED 076 608 TM 002 588

AUTHOR Baker, Eva L.
TITLE Teaching Performance Tests as Dependent Measures in Instructional Research.
PUB DATE 1 Mar 73
NOTE 16p.; Paper presented at Annual Meeting of American Educational Research Association (New Orleans, Louisiana, February 25 - March 1, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Classroom Observation Techniques; Effective Teaching; *Evaluation Methods; Instructional Improvement; Instructional Programs; *Measurement Techniques; *Performance Tests; Student Evaluation; *Teacher Evaluation; Technical Reports

ABSTRACT

The need for common measures in research on teaching is legend, and the merits of teaching performance tests to meet this requirement are explored here. A regression study where teacher performance tests were used as dependent measures is described. Sixth-four subjects were given objective-based lessons to teach. During their lesson, they were rated on the use of six instructional principles. Following instruction, learners were administered a short test of achievement and interest. Step-wise regression analyses were conducted, and variables related to the performance criteria described. Suggested modifications of performance tests to enhance their suitability as dependent measures are discussed. (Author)

Paper to be Presented at the Annual Meeting of the
American Educational Research Association
New Orleans, Louisiana
February 25-March 1, 1973

Teaching Performance Tests as Dependent Measures in Instructional Research

Eva L. Baker
University of California, Los Angeles

After the clattering controversy, the claims and counterclaims which have enveloped the question of teaching performance tests, this paper was intended to provide a crystalline moment of unity. Whereas teaching performance tests have been characterized and challenged as useful in teaching assessment, teaching analysis and teaching improvement, the purpose of this paper was to discuss such tests in an important and blessedly undisputed role: as dependent measures for instructional experiments involving teacher behavior. The assumed agreement among us is based on three factors: 1) there is a continuing need for reasonable dependent measures for instructional experiments, especially ones which can be used under a wide number of conditions and might conceivably function as standard referents for evaluating the sensitivity of experiments; 2) performance tests are logically related to significant aspects of the instructional task; 3) the tests are "effectiveness-based" so scores on them are suitable for the desired outcome nature of a dependent variable.

This paper intended to sanctify the use of such tests with a neat and unassailable example of performance tests functioning as anticipated. Unfortunately, thi felicitous rite must be deferred until the future and tenor of the paper shifted to one of faltering exploration rather than

resounding demonstration.

In lieu of conducting a convocational exercise, I shall rather provide the details of a research study where performance tests were used as the dependent measure, present the obtained data analyses and discuss some of the difficulties, insurmountable and otherwise, inhibiting unwavering confidence in the use of performance tests as possible dependent measures. The study is presented as a reference for discussion. The particular finding should not be considered to be the central topic of this presentation.

Overview:

Data were obtained for this study from sixty-four teacher education candidates, each asked to teach short performance tests to their peers as part of the instructional requirements for a course in curriculum and instruction. During these lessons the students were rated by eight trained observers according to their use of six instructional techniques. Correlation and step-wise regression analyses were planned to be conducted on the data using achievement scores and interest ratings of students as dependent measures.

Performance tests:

Eleven different performance tests were used in the study. These were assigned at random within each group of subjects. The reason for the great number was to insure that the task would be new to each subject. These tests each consisted of a statement of an operationally defined instructional objective, a sample test item, and approximately two pages of relevant content on the topic covered by the objective. Students taught by teachers were asked to complete a short posttest following the fifteen

minute instructional period and were also asked to rate the lesson in interest on a five point scale. Topics for the tests were concepts relevant to the course, but likely not to have been encountered by these subjects, e.g., erosion measures in educational evaluation. Behaviors called for by the objectives were either discrimination of examples of concepts, e.g., "Is X an example of A?" or classification, "Which of the following four procedures is Z an instance of?"

Subjects:

Sixty-four senior and graduate students preparing to be secondary level teachers were involved in the study. To be admissible for teacher education work these students normally need to have a 3.0 grade point average at UCLA or a comparable institution. These subjects were randomly assigned to "minilesson" groups.

Raters:

Eight different raters were involved in the observation phase of the study. These students were, for the most part, teacher education candidates who were exceptionally successful in completing the curriculum and instructional course. Operating on a Keller (1968) model, these students received course credit for acting as an assistant/leader to one or two groups of from five to eight students. The raters, by virtue of their selection, had already demonstrated capability in the analysis of instruction according to the techniques for observation. A two-hour training session, followed by weekly reviews for three weeks was administered to each rater.

Instructional Techniques:

Six instructional techniques were to be observed and considered as

the independent variables in this investigation. These techniques can be exhibited in teacher behavior and have been demonstrated to be reliably judged (Baker, 1969). The techniques observed were the following:

Direct Practice:	Did the teacher provide opportunity for the class to practice the criterion behavior described in the objective?
Knowledge of Results:	Did the teacher inform the students whether their responses were adequate?
Prompting:	Did the teacher provide cues to allow responses to be more easily made at the outset of instruction and then reduce the students' dependency upon cues?
Individualization:	Did the teacher respond to the individual attributes and experiences of the students by varying instruction for certain individuals?
Task Description:	Did the teacher communicate in unambiguous language what task the learners should focus upon?
Motivation:	Did the teachers attempt to provide incentives or explanations designed to enhance the appeal of attending to instruction?

Justification for the selection of these variables can be found in almost any consideration of the literature in instruction. Full descriptions of techniques 1, 2, 3, and 5 are presented in a recently prepared set of materials (Baker and Quellmalz, 1972).

Procedures:

All enrolled students in the Curriculum and Instruction class were to complete at least one "minilesson" during the course of the quarter. These lessons were assigned at random within groups one week in advance to each student, prior to the scheduled conduct of the lesson and the administration of measures. During each lesson, raters surreptitiously completed brief rating forms where the "teachers'" use of six principles

was assessed. These forms were concealed in a notebook, and the rater sat off to the side of the instructor. No mention or discussion of the dimensions recorded on the form followed any lesson. Following the conclusion of each fifteen minute lesson, students were permitted to complete the posttest and the interest rating form. There was no time limit imposed in the testing. Data were collected over a three week period.

Data Analysis:

Average scores for the students in each teacher's lesson were computed in percentage terms for each teacher. Average rating from 1 to 5 on interest was also computed.

Analyses were first conducted on the raw data, that is, the average percentage of achievement and average interest rating for each teacher. Means and standard deviations by test for cognitive and interest rating are presented below in Table 1.

Table 1. Means and Standard Deviations by Test Forms for Achievement and Interest

Test	Achievement		Interest		
	$\bar{X}$	s	$\bar{X}$	s	n
1	84.69	19.38	1.90	.74	16
2	79.00	14.18	2.02	.32	14
3	81.67	10.36	1.93	.58	12
4	74.00	0.00	2.60	.00	1
5	71.00	0.00	1.90	.00	1
6	100.00	0.00	2.30	.00	1
7	80.67	11.85	2.10	.40	3
8	56.67	11.37	2.20	.70	3
9	61.33	10.07	2.40	.17	3
10	86.00	2.16	1.70	.49	4
11	85.00	0.00	2.00	.00	1
xx (uncoded)	76.00	10.30	2.12	.13	5
TOTAL	79.55	15.14	2.00	.61	64

From the distribution of scores one can see that the idea that the tests would be approximately equally distributed among the student didn't work out. Explanations for the disproportionate assignment of topics is related to absenteeism in the minilesson sessions. In all, subsequent analyses, data were considered for subjects in Tests 1, 2, and 3, where the distribution of subjects was most equitable (N = 16, 14, 12 respectively).

Observation data were averaged for all subjects. Means and standard deviations of ratings are presented in Table 2.

Table 2. Means and Standard Deviations of Ratings of Six Instructional Techniques

Variable	$\bar{X}$	s	n
Practice	3.06	1.33	64
Knowledge of Results	3.06	1.44	
Prompting	2.17	1.27	
Individualization	2.30	1.28	
Task Description	3.31	.88	
Motivation	2.68	1.27	

Table 3. Means and Standard Deviations of Ratings of Six Instructional Techniques. (Tests 1, 2, 3 only)

Variable	$\bar{X}$	s	n
Practice	3.00	1.41	42
Knowledge of Results	2.93	1.52	
Prompting	2.19	1.37	
Individualization	2.38	1.43	
Task Description	3.26	0.91	
Motivation	2.81	1.29	

Correlation coefficients were computed for the variables and the dependent measures and are reported in Table 4.

Table 4. Correlation Matrix of Rated Techniques (N = 42)

	Practice	Knowledge of Results	Prompting	Individualization	Task Descrip.
Practice	.00				
Knowledge of Results	.94				
Prompting	.49	.52			
Individualization	.40	.45	.60		
Task Description	.45	.38	.27	.09	
Motivation	.11	.17	.37	.34	.29

The interrelationship among the independent variables is obvious.

Practice was found to correlate significantly (beyond .01 level) with each of the independent variables, with the exception of motivation. The pattern of relationship for the variable, Knowledge of Results (as expected for the .94 correlation), was identical to that of Practice. Prompting correlated significantly with each of the other independent measures. Individualization was not found to correlate only Task Description. Motivation significantly correlated with Prompting, Task Description, and Individualization.

Data from both dependent variables, achievement and interest, were transformed into standard scores ($\bar{X} = 50$, $s = 10$) within each test and them combined.

Correlations of the independent variables with both raw and transformed scores for the dependent measures are presented below.

Table 5. Correlation of Independent and Dependent Variables

	Achievement		Interest	
	Raw	Transformed	Raw	Transformed
Practice	.28*	.28*	.36**	.34*
Knowledge of Results	.18	.19	.37**	.35*
Prompting	.26*	.30*	.15	.23
Individualization	.29*	.29*	.09	.18
Task Description	.07	.02	.04	.00
Motivation	-.05	-.02	.30*	.34*

* $p < .05$

** $p < .01$

N = 42, Tests 1, 2, 3 only

Equating the tests by transforming scores resulted in little change in correlational values. To summarize, the variables found to be most related to achievement were the observed use of Practice, Prompting and Individualization. Variables significantly related to interest ratings by students were Practice, Knowledge of Results and Motivation. A previous study (Baker, 1969) using a sample of 80 teachers and 20 different performance tests obtained significant correlations, of about the same magnitude for the techniques of Practice, Individualization and Knowledge of Results.

Step-wise regression analyses were conducted as intended and are described in the Appendix. However, because of the intercorrelations among the predictor's conclusions from these analyses are tenuous at best. For a brief resume, please see the Appendix section.

Discussion of the Study

Certain variables were found to relate significantly to achievement and ratings of interest. Because of the high correlation of practice and knowledge of results, perhaps only one such variable should be observed in future investigation, probably practice, because it was found to be related positively to both achievement and interest dimensions.

Exploring Problems with the Dependent Measure

The multiple performance test approach

If one were in a position to use a single performance test as a dependent measure, problems would be both dissipated and created. The use of a single performance test would have precluded the necessity to transform scores, but would reduce the generalizability (if any) of the obtained results. Transformation is required, not to weight the contribution of each

test but to neutralize differences in the difficulty of each test.

Multiple tests also raise the issue of possible performance test/independent variable interactions, in that a given test might be more conducive to the use of certain techniques than others.

The question of reliability

Jason Millman will report his work on the psychometric properties of performance tests. Clearly stability coefficients (test-re-test of teachers) would be important to obtain. The present study was designed so that each subject be taught only once and thus the design precludes such analysis. Reliability analyses were computed for Tests 1, 2, 3 where 16, 14, 12 teachers were involved. In Table 6 below, these findings are summarized.

Table 6. Achievement Reliability on Test Items for Performance Tests 1, 2, and 3.

Test	No. of items	$\bar{X}$	s	alpha	N
1	10	8.97	1.67	.78	78
2	5	3.76	1.27	.60	71
3	10	8.07	1.41	.45	60

Because performance test items are designed to be homogeneous measures of the objective, difficulty persists in interpreting such data, particularly in the light of the pursuit of substitute reliability-determination procedures for objective-referenced tests. Especially, it is not clear that one can generalize the results of analyses performed on instructor groups with a wide distribution of teaching talent (as measured by performance) to groups of instructors trained to behave more homogeneously and to produce more similar results.

Differences in Requirements for Performance Tests for Evaluation and Research Purposes

The informal banter in educational circles alleges that instructional research should be carefully done with enormous attention to detail, and in contrast, evaluation studies may be permitted to proceed with a much more casual view regarding design, controls, and the other catch-words of the educational research community. The performance test notion provides an example where, I would suggest, the practiced precision requirements must be reversed. When performance tests are to be used for decision purposes, such as the evaluation or selection of teachers, one would need, on ethical grounds, a clear concern with the consistency and validity of the experience for the individuals tested. If instructional improvement programs, or in some cases, career chances are modified at all by the use of such measures, one would wish to have heightened confidence in the basis for the decision. On the other hand, and at the risk of sounding wild and contemptuous of standards of research rigor, constraints on the use of performance tests to investigate the relationship of independent variables might be somewhat relaxed. For instance, if the tests have imperfect reliability coefficients, in light of imperfect methodology, the research ethos is to report the data, qualify one's conclusions and encourage replication. The trained consumer of educational research takes his or her own risks. Each should be able to evaluate the validity of an investigation, particularly if the designer of the study is careful to disclaim proving anything and clearly indicates the limitation of the work. At best, the study will be replicated and inadequacies learned first-hand. At worst, someone might design and conduct an extension of the work and waste some effort if the original study were not carefully reported or understood.

The control of personal destiny and the interpretation of knowledge is in the hands of researchers, who presumably subscribe to a minimum set of standards in design and interpretation of empirical work. On the other hand, teachers who are evaluated by use of the performance test and for whom decisions may have serious consequences are not in positions of control. Moreover, it is likely, that the educational p[illegible] conducting the evaluation, a school principal, for example, may not be sensitive to the inadequacies of the data. Thus, I would hope to see performance tests developed to a high level of precision when they are to be used for personnel decisions and to permit less well-refined tests to function in an exploratory role in instructional research.

Other contrasts between the attributes of tests for evaluation and research purposes may be drawn. For instance, one would expect that perceived relevance to the task of teaching would be at a higher premium in evaluation rather than research problems. Similarly, the need to permit adequate preparation time would vary. Certainly, in instructional research, a dependent measure with little variability is not desirable. In contrast, performance tests for teacher evaluation might not require as much variability and could be used to identify only the aberrant individual, one who has been given time and assistance, and still is unable to demonstrate influence over the outcomes of instruction.

Performance Tests as a Technology

Performance tests represent the beginning of a technology, and if the history of technology at large is repeated, they will not be used only or primarily for the purpose for which they were originally intended. The original experimenters with laser technology had some particular purposes

in mind; the present use of lasers is wide-ranging and continues to be explorative. Scotch tape was developed for a given purpose, but as users of the product experimented, the invention gained expanded functions. New uses were suggested and the technology was explored as changes in the product were made, so that scotch tape has a range of utility, from to providing a writing surface on inhospitable exteriors to pasting down durls on cheeks of teen-age girls.

The performance test may become an effective tool if it is considered as an invention, to be tested against various uses and fruitful modifications and not prematurely ossified. If the performance test proves adaptive to the broad requirements of the field, its utility as a dependent measure might be only one of its important contributions.

Appendix

Analysis of the data was pursued, using step-wise regression, and data are presented in Tables A and B from transformed scores.

Table A. Step-wise Regression Summary for Transformed Achievement Scores

Variable	Multiple R	F
Prompting	.307	
Practice	.345	5.062
Knowledge of Results	.438	4.128
Task Description	.475	.784
Individualization	.492	.952
Motivation	.497	.208

N = 42

Table B. Step-wise Regression Summary for Transformed Interest Ratings

Variable	Multiple R	F
Knowledge of Results	.351	
Motivation	.452	7.164
Task Description	.502	3.91
Practice	.528	1.388
Individualization	.539	.579

Because of the correlation among predictor variables, analysis were re-run for both sets of data. For the transformed achievement dependent measure, Knowledge of Results was dropped from the set of variables. The variables identified in the analyses were Prompting, Practice and Task Description.

For the transformed interest ratings, Practice was deleted, again because of the .94 correlation obtained with Knowledge of Results. The order of the predictors was as follows: Knowledge of Results, Motivation, Task Description, and Individualization.

Table C. Summary of Step-wise Regression with Transformed Achievement Scores Deleting Knowledge of Results

Variable	Multiple R	F
Prompting	.307	.914
Practice	.345	.862
Task Description	.370	.251
Motivation	.381	.579
Individualization	.398	.544

Table D. Summary of Step-wise Regression with Transformed Interest Scores Deleting Practice

Variable	Multiple R	F
Knowledge of Results	.351	
Motivation	.452	6.020
Task Description	.503	2.763
Individualization	.513	0.526

References

Baker, Eva L. "Relationship Between Learner Achievement and Instructional Principles Stressed During Teacher Preparation," The Journal of Educational Research, Vol. 63, No. 3, November, 1969, pp. 99-102.

Baker, Eva L. and Quellmalz, Edys. Research-Based Techniques for Instructional Design. U. S. Office of Education, Department of Health, Education, and Welfare: National Center for Educational Research and Development, 1972, 307 pp.

Keller, Fred S. "Goodbye Teacher...." Journal of Applied Behavioral Analysis, 1(1), pp. 79-89, Spring, 1968.