R E P O R T R E S U M E S

ED 015 144 SP 001 255

A PERFORMANCE TEST OF TEACHING EFFECTIVENESS.

BY- POPHAM, V. JAMES BAKER, EVA L.

PUB DATE 19 FEB 66

EDRS PRICE MF-$0.25 HC-$0.36 [illegible]P.

DESCRIPTORS- ACADEMIC PERFORMANCE, *EFFECTIVE TEACHING, INDUSTRIAL EDUCATION, *PREDICTIVE MEASUREMENT, RELIABILITY, SECONDARY EDUCATION, STANDARDIZED TESTS, STATISTICAL ANALYSIS, STUDENT ATTITUDES, STUDENT INTERESTS, STUDENT MOTIVATION, TABLES (DATA), *TEACHER EVALUATION, *TEST CONSTRUCTION, *TEST VALIDITY, KUDER RICHARDSON FORMULA 20, WONDERLIC PERSONNEL TEST,

THIS REPORT DESCRIBES THE INITIAL VALIDATION OF PERFORMANCE TESTS OF TEACHER EFFECTIVENESS--USING PUPIL GAINS AS THE CRITERION OF EFFECTIVENESS--AND THE STEPS TAKEN IN RECOGNITION OF THE PROPRIETY OF SUCH MEASURES ONLY IF ALL TEACHERS ARE TEACHING FOR THE SAME OBJECTIVES. AS A FIRST STEP, IT WAS HYPOTHESIZED THAT A VALID PERFORMANCE TEST OF TEACHER EFFECTIVENESS SHOULD DISCRIMINATE BETWEEN TWO EXTREME GROUPS-(1) NONTEACHERS AND (2) SUPERIOR EXPERIENCED TEACHERS-BEFORE IT COULD BE USED FOR ASSESSING TEACHERS WHO DIFFER IN SPECIFIED WAYS-(E.G., THOSE WHO ARE AND ARE NOT INTENSIVELY TRAINED TO BRING ABOUT BEHAVIOR CHANGE IN STUDENTS). SCORES ON STUDENT ACHIEVEMENT MEASURES ON TWO INDUSTRIAL EDUCATION TOPICS WERE ASSESSED FOR RELIABILITY AND INTERCORRELATED WITH MEASURES OF GRADE POINT AVERAGE, INTEREST IN THE SUBJECT MATTER, AND WITH WONDERLIC PERSONNEL TEST SCORES. THE OBJECTIVE HERE WAS TO DETECT VARIABLES THAT COULD POTENTIALLY BE USED TO CONTROL FOR STUDENT DIFFERENCES IN SUCH FACTORS AS "SET," INTELLIGENCE, ETC., IN ASSESSING TEACHER EFFECTIVENESS. KUDER-RICHARDSON RELIABILITY COEFFICIENTS OF .44 AND .78 WERE FOUND FOR THE ACHIEVEMENT TESTS. TEST SCORES CORRELATED .68 WITH GRADE POINT AVERAGE. A "PERPLEXING" FINDING WAS HIGHER TEST SCORES AMONG THOSE EXPRESSING LESS INTEREST IN THE INSTRUCTIONAL TOPIC. PRETEST SCORES WERE MORE HIGHLY CORRELATED WITH POSTTEST SCORES THAN WERE WONDERLIC SCORES. (PAPER PRESENTED AT THE 1966 AMER. EDUC. RES. ASSN. MEETING, CHICAGO, FEBRUARY 17-19, 1966). (AF)

Paper Presented at the 1966 American Educational Research Association Meeting
Chicago, Illinois, February 17-19, 1966

A PERFORMANCE TEST OF TEACHING EFFECTIVENESS*#

W. James Popham and Eva L. Baker
University of California, Los Angeles

Problem: One of the more enduring and distressing problems in education has been our inability to develop satisfactory measures of teacher effectiveness. Although the amount of attention which this problem has received during the past sixty years is considerable, few really promising advances have occurred. Generally, three classes of criterion measures have been employed in previous empirical studies: ratings, observations, and pupil gain. Most researchers in the field agree that the ultimate criterion of teacher competence is pupil growth, and usually ratings and observations of the teacher's behavior have been used as indications of the instructor's probable influence on pupils.

The prime difficulty in using pupil change as a measure of teaching proficiency is a consequence of the fact that different teachers often seek to accomplish different objectives. Turner and Fattu[1] have built a compelling argument that since teachers' objectives vary from situation to situation, it is impossible to use measures of teaching effectiveness which do not take account of such variability, and thus inappropriate to compare teachers on the basis of their students' growth toward dissimilar goals. These researchers[2] have attempted to resolve this dilemma by using as an index of teaching skill the teacher's ability to solve paper and pencil problems which represent selected teaching tasks. In their approach, the teacher is required to perform several different types of tasks, such as determining the order of materials according to their difficulty level for pupils.

The rationale for the research described in this paper is similar to that of Turner and Fattu, except that an actual performance test of teacher's ability to produce pupil achievement is used instead of a paper

*The research reported herein was performed pursuant to a contract with the U.S. Department of Health, Education and Welfare.

#The research reported herein was supported by the Cooperative Research program of the U.S.O.E., U.S. Department of Health, Education and Welfare.

[1]Richard L. Turner and Nicholas A. Fattu, "Skill in Teaching, a Reappraisal of the Concepts and Strategies in Teacher Effectiveness Research," Bulletin of the School of Education, Indiana University, Bloomington, Vol. 36, No. 3, May, 1960.

[2]Richard L. Turner, "Task Performance and Teaching Skill in the Intermediate Grades," The Journal of Teacher Education, Vol. 14, No. 3, September, 1963, pp. 299-307.

and pencil predictive test. The problem of different objectives is hopefully allayed by giving teachers identical goals to achieve. The assumption is that a teacher who is a successful goal achiever with given objectives will, other factors remaining equal, more probably be successful in achieving his own instructional objectives.

It is with this in mind that a series of performance tests of instructor competence are being developed at U.C.L.A. under provisions of two U.S.O.E. contracts. Two of the tests are in the field of vocational education and one is in the social sciences. These tests consist of (1) a set of operational instructional objectives stated in terms of specific pupil behaviors, (2) a collection of possible learning activities which the teacher may wish to employ, and (3) pre- and post- tests, not seen or administered by the teacher, which adhere closely to the operational objectives. The objectives and possible activities are given to the teacher one week in advance of instruction, and he is told to prepare plans for three weeks of teaching. By stipulating identical objectives to be achieved and tested but permitting teacher divergence in accomplishing these ends, a method of evaluating teacher performance without impinging on individual pedagogical style is provided. Teachers are comparing on the basis of their pupils' achievement rather than other, more idiosyncratic criteria.

Related Research. A review of teacher effectiveness research since the turn of the century finds a considerable number of studies, but none which might be classified as "breakthrough." Along with the periodic reviews of the field, such as those by Morsh and Wilder[3] and, more recently, Barr[4] and Ryans[5], we have witnessed a number of theoretical and empirical reports by individuals such as M. Cogan, N. Flanders, N. Gage, D. Medley, B. Smith, R. Wilk, W. Edson, and J. Withall[6], to mention but a few.

There is, fortunately, an increasing tnedency on the part of researchers to do away with such simplistic notions as the effective teacher and to replace these conceptions with a view that the teacher's instructional proficiency is the function of the particular setting in which the instruction takes place, including the particular students, the social and physical environment, and the instructional goals.

---

[3]Joseph E. Morsh and Eleanor Wilder, "Identifying the Effective Instructor: A Review of the Quantitative Studies, 1900-1952," Research Bulletin AFPTRC-TR-54-44, Lackland Air Force Base, Texas, 1954.

[4]A.S. Barr, issue eidtor, "Wisconsin Studies of the Measurement and Prediction of Teacher Effectiveness: A Summary of Investigations," Journal of Experimental Education, Vol. 30, September, 1961, pp. 5-156.

[5]David G. Ryans, "Assessment of Teacher Behavior and Instruction," Review of Educational Research, Vol. 33, No. 4, October, 1963, pp. 415-441.

[6]All of these researchers contributed articles to the Symposium on Classroom Behavior of Teachers, edited by Harry F. Silberman for the Vol. 14, No. 3, September, 1963, issue of The Journal of Teacher Education.

The present line of research is an outgrowth of the writers' studies of the influence of certain aspects of teacher education programs of the classroom instructional behavior of student teachers. In these projects, the problem of dissimilar instructional goals presented formidable measurement obstacles. The decision to develop performance tests of teaching proficiency was reached partly as a result of these difficulties.

Objectives. At the end of two years it is planned to test a specific validation hypothesis with each of the three performance tests. The hypothesis is that there will be significant differences in pupil gains achieved by (1) non-teachers and (2) expereinced teachers in a limited instrutional situation provided by the performance tests of teaching proficiency. Although this hypothesis relates to the validity of the performance tests, the prior question of their reliability will also be treated. The predicted outcome of the primary hypothesis test is that the experienced teachers will produce better student gains than the non-teachers. At present only two of these tests have been developed to the point where they have been field tested. This paper will describe the preliminary results with these two instruments.

It should be pointed out that this two year project represents only an initial effort to validate the performance tests. If the tests withstand this evaluation, more sensitive attempts to assess their validity must be undertaken. The validation of a teacher effectiveness measurement device is particularly difficult, of course, because of the lack of defensible criterion measures against which to evaluate teacher performance on the device. Thus, as a first step in the evaluation of these new instruments we will attempt to show that at least the tests discriminate between two extreme groups: (1) non-teachers and (2) experienced teachers previously judged superior by their supervisors with respect to promoting pupil achievement.

The writers are well aware of the notorious deficiencies in administrator ratings of teacher competence. Yet, for this first assessment of the performance tests' validity it seems reasonable that a group of teachers who have been judged excellent by administrators and supervisors should certainly be able to out-perform a group of non-teachers. This assumption, of course, is central to the rationale of the study.

If the perfomance tests withstand this initial trial, it can then be seen if they discriminate between other groups such as (1) teachers whose past records indicate great ability to produce pupil achievement (as reflected by student performance data) and (2) teachers whose records indicate the opposite. The tests should also discriminate between samples of (1) experienced teachers and (2) experienced teachers who have received special, intensive training in how to bring about student behavior change with respect to specified objectives. These studies, however, depend on the performance tests' first passing the validity hurdle represented by the presently planned investigation.

Description of the Tests. At this point initial versions of two tests have been developed in the fields of electronics and auto-mechanics. A

difficult early decision regarding these two tests involved the selection of topics for the unit. Ideally, it was judged that topics should be (1) sufficiently important so that teachers would be willing to include them in their curricula, and (2) sufficiently unique so that great student familiarity with the topic would not be common. It was also hoped that topics could be selected which could be inserted rather freely at various points during the academic year.

With these criteria in mind, the topic selected for the electronics unit was the "General Principles of Electronics Troubleshooting," and the topic selected for auto-mechanics was "Carburetion." Although development of the social science test has just commenced, the topic tentatively selected is "Research Methods in the Social Sciences."

Developmental work on the two units occurred in the following pattern: First, topics meeting the above criteria which might be covered in two or three weeks were selected. These were then submitted to several subject matter specialists who served as consultants during the project. From these tentative topics, two were selected and a series of instructional objectives were prepared which were also screened by consultants. A preliminary set of these objectives was agreed on and test items based directly on the objectives were developed. In addition, possible learning activities and reference materials were assembled. In some instances these learning activities were designed to be particularly pertinent to the given objectives. In other cases, the activities were planned to be "flashy" but not germane to the objectives. It was thought that less experienced insturctors might be attracted to the "flashy," irrelevant activities, but that the sophisticated teacher would tend to use the pertinent activities. These materials were revised several times prior to initial trial. It is, of course, possible that the teacher might choose to develop his own activities and not use any of the materials provided in the unit.

The early forms of the post-tests were given to several teachers for administration to classes of students currently taking electronics or auto mechanics courses. Such data underwent item analysis procedures which resulted in the improvement of many test items.

When ready for the first field trial both the carburetion and electronics unit consisted solely of objectives measurable by paper and pencil tests. The instructional time alloted to each was ten hours. Currently, the carburetion materials (with 49 objectives) consists of 22 pages while the electronics unit (with 23 objectives) consists of 41 pages. Both units were considered incomplete, for it is planned during the next year to add objectives demanding performance on actual carburetors and electronic circuit boards which will require the instructional period to be lengthened from 10 to 15 hours.

For each unit a pre- and post- test were prepared. The post-test for electronics consists of 52 multiple choice items and the post-test for carburetion consists of 97 multiple choice items. Both of these tests were drawn specifically from the objectives. Items for the pre-tests were randomly selected from the post-test items in order to provide measures which could be completed in approximately 20 minutes. The pre-test for electronics contained 17 items while the carburetion pre-test had 20 items.

Initial Tryout. Both performance tests were given their first test during January and February, 1966. The electronics test was tried out in eight electronics classes at Los Angeles Trade Technical College.[7] The carburetion performance test was tried out in two classes, one at the junior college level at Los Angeles Trade Technical College, and one at Fullerton Union High School.[8] Thirty-six students took part in the carburetion tryout and 108 students in the electronics tryout.

Each instructor was given a copy of the unit objectives and the resource materials approximately one week prior to the time he was to teach the unit. Each was instructed to attempt to accomplish the objectives stated in the unit, but to use any instructional techniques he wished. For purposed of this trial, participating instructors were also asked to make suggestions regarding ways in which the materials could be improved.

Arrangements were made with each teacher so that a member of the research staff administered (1) a twenty minute pre-test during the first day of the ten hours devoted to the unit and (2) a firty minute post-test at its conclusion. In addition, a questionnaire was administered to the students at the close of the unit. A questionnaire was also given to the teacher at that time soliciting his suggestions regarding the unit. Finally, in two of the electronics classes and one carburetion class the Wonderlic Personnel Test, a 12 minute test of "problem solving ability" was administered to the students at the time of the pre-test. In all, 25 different variables were represented by the two questionnaires and the Wonderlic. The trials were completed between the dates of January 17 and February 10, 1966.

Analysis. Two different types of analysis were conducted on the preliminary data. The first was to compute item analyses and coefficients of internal consistency (Kuder Richardson Formula 20) on the pre- and post-tests. The second was to compute intercorrelations among the several measures. Key interest in the latter analysis focused on the possibility of detecting variables which could be used, in part, to control for differences among the pupils due to such factors as "set" toward the unit's material, intelligence, etc. Further, of course, the responses of the instructor were carefully considered. The overall purpose of the initial analysis was essentially heuristic. We were attempting to find possible variables to be considered in subsequent trials of the materials.

It was fully expected that a great number of deficiencies would exist in the first experimental versions of both the performance tests. Procedural deficiencies regarding such details as tests administration, relations with instructors, etc., were also anticipated.

---

[7]Appreciation is expressed to the administration and staff of Los Angeles Trade Technical College for their participation in this investigation.

[8]Appreciation is also expressed to the administration and staff of Fullerton Union High School for their participation in this investigation.

Results. The performance of students on the pre- and post-tests is summarized in Table 1. Item analysis results revealed a considerable number of items, particularly in the electronics tests, which were in need of revision. The KR_{20} coefficients were markedly higher for the carburetion tests than for the electronics tests.

Table 1
Electronics and Carburetion Pre- and Post-Tests Results

	Electronics		Carburetion	
	Pre-Test	Post-Test	Pre-Test	Post-Test
Number of Pupils	108	98	36	33
Number of Items	17	52	20	97
Mean	9.68	23.68	10.41	51.64
Standard Deviation	2.71	9.57	3.48	17.86
KR_{20}	.56	.44	.71	.78

Of the 25 variables constituting the pupil and teacher questionnaires and the Wonderlic test, interest centered on those which might be of value in adjusting for initial differences among pupils and/or teachers. However, some of the questionnaire items were not constructed for this purpose. For instance, teachers were asked to list the number of instructional hours they actually used during the unit. While one might be interested in the possible correlation between instructional time invested and pupil achievement, there is no need to control for such instructional variables. On the other hand, if related to post-test scores, pupil entry behavior variables, such as grade point average, as well as teacher variables, such as sttitude toward the unit's objectives, might permit statistical covariance adjustments in analyzing data.

In the case of carburetion, the variables which were most strongly related to post-test performance were the following (all positive relationships): (1) pupils' overall grade point average, r=.68; (2) pupils' estimate of the pre-test's difficulty, i.e., the students who thought it easier tending to score higher on the post-test, r=.60; (3) pupils' expressed interest in the general field of auto mechanics, i.e., the students responding with less interest tending to score higher on the post-test, r=.59; (4) pupils' expressed interest in carburetion prior to beginning the unit, i.e., lower interest associated with high post-test performance, r=.57; and (5) pre-test socres, r=.56. The correlation between Wonderlic socres and post-test scores for only one class or 20 pupils was .26.

For electronics, the variables most highly related to post-test performance were: (1) pupils' estimate of the pre-testss difficulty, i.e., students who thought it easier tending to score higher on the post-test, r=.56; (2) pupils' overall grade point average, r=.49

(3) pupils' expressed interest in the general field of electronics, i.e., students responding with less interest tending to score higher on the post-test, r=.47; and (4) pupils' expressed interest in electronics troubleshooting prior to beginning the unit, i.e., lesser interest associated with higher test performance, r=.39. Negligible correlations existed between post-test scores and pre-test performance (r = .05) as well as Wonderlic performance (r = .05).

The chief suggestions from participating instructors concerned the addition and deletion of certain objectives for the units. A number of criticisms were made of the technical terminology employed in the objectives and reference materials. Many instructors thought that the topics could not be adequately treated in ten instructional hours. Many minor deficiencies in the quality of the reference materials were also noted.

Discussion. This first trial of the two performance tests yeilded the anticipated results, namely, a host of defects in the procedures and materials employed. The internal consistency coefficients for the electronics pre- and post-tests were particularly low and will warrant factor analytic treatment to see if (1) we are indeed measuring two or more relatively distinct dimensions or, as is more likely, (2) the test has too many poor items.

The strength of the relationships between the post-test measures and some of the pupil variables which might be used for control purposes was rather encouraging. Several of the relationships were rather perplexing, however, including the tendency for those expressing less interest in the general field and specific unit topic to score higher on the post-test. Common sense might suggest that the opposite would be true. The Wonderlic Personnel Test, apart from its ease of administration, seemed to offer little promise in this study as a predicter of post-test performance. In the case of carburetion, the pre-test score appeared to be a much more effective predicter. Hopefully, further analysis of these data and subsequent field trials will reveal a number of measures which can be used to reduce variation among different classes of pupils.

Even though this initial set of data must be scrutinized with much more care, it is apparent even now that a great deal more work must be expended in improving the materials and measures involved in these performance tests of teaching effectiveness.