Educational Leadership / RESEARCH SUPPLEMENT

EDITED BY ASCD RESEARCH COUNCIL • FREDERICK A. RODGERS, CHAIRMAN

Volume 3 **October 1969** Number 1

CONTENTS

A Reliable Measure of Teacher Effectiveness — *Thomas B. Justiz* **49**

The Relationship of Self-Supervision to Change in Selected Attitudes and Behaviors of Secondary Student Teachers — *Donald P. Johnston* **57**

A Call for Papers — *Frederick A. Rodgers* **63**

A Reliable Measure of Teacher Effectiveness

***THOMAS B. JUSTIZ** **

THE absence of reliable pupil performance criterion measures of teacher effectiveness is a well known fact in the state of the art. In the past 60 years, teacher effectiveness literature has been so clouded with arguments as to why pupil performance criterion measures are supposedly unworkable that at times it is difficult to see beyond this position. Consider the following excerpts from various research articles, as examples:

1. "Researchers, with few exceptions, have not been too successful in demonstrating that the method differentiates between more and less competent teachers."

2. "It is never possible to isolate the influence which can be attributed to a given teacher over a given period of time."

3. "The tests used in developing the pupil gain criterion will have varying degrees of operational validity, except as the teachers agree to pursue certain stated objectives which can be defined with sufficient clarity to provide like meanings to all the participants."

4. "The imperfections in tests used make it difficult for some pupils and classes to demonstrate satisfactory gains no matter how effective the teaching."

5. "The gaps in the criterion arising from inadequate tests with which to measure pupil gain will be found to be considerable."

6. "Pupil gains measures tend to have low reliability (to be inconsistent) and so to be of doubtful validity."

7. "Pupil growth varies with different pupils; poor English speaking ability on the part of some pupils may handicap some teachers."

8. "Performance tests are impractical to use in the school; they demonstrate low relia-

* *Thomas B. Justiz, Research Educationist, Los Angeles, California*

Copyright © 1969 by the Association for Supervision and Curriculum Development, NEA. All rights reserved.

bility, and therefore doubtful validity, and they cannot differentiate between more and less competent teachers."

9. "Finally, tests measure effects but not causes. The sources of the effects observed are not readily ascertained, even under carefully controlled experimental conditions."

This is a formidable list of obstacles. Most of them, however, were overcome in a single research design, and the findings were considered to be evidence of a valid, reliable, and practical measure for assessing general teaching ability in the public schools.[1]

Theoretical Frame of Reference

The study was designed for a university coordinator of student teachers or a high school principal who may be concerned with securing one of the brighter prospects from his internship training program. Either person, it was assumed, would be interested in collecting some evidence of the pupil-achievement-producing ability of the student teachers, as an indication of their overall effectiveness. The dependent variable (or criterion of teacher effectiveness) of the study was, therefore, operationally defined as: the mean of all post-test scores generated by a class of pupils, on content-validated tests of subject matter objectives.

So that no student teacher would have a subject-matter advantage, the range of objectives was restricted to subject fields which were unfamiliar to both the student teachers and their pupils. Furthermore, to enhance the reliability of the student teacher rankings (on the basis of their pupils' performance), the student teachers were required to teach twice, in succession, using objectives in two different subject fields.

Now we are ready to draw our first hypothesis. Will student teachers who teach effectively in one subject field teach as well in the second subject field (providing that all student teachers were unfamiliar with the subjects prior to the time they were confronted with the teaching situation and, providing of course, that all student teachers were able to begin instruction with equivalent learning abilities in their classrooms)?

Performance tests take time and careful planning, and any time we can validate some predictor of a teacher's effectiveness, we are all too grateful. Therefore, our second hypothesis creates the independent variable. Using the Minnesota Teacher Attitude Inventory questionnaire: Will student teachers who score well on the MTAI also produce high pupil achievement mean scores in their classes? (The independent variable is a teacher's attitude toward pupils and toward teaching in general.)

Procedure and Controls

In order to establish a reliable measure of teacher effectiveness based on pupil performance, four obstacles were to be overcome, as follows:

1. Samples of student teachers were to be made available, all teaching to the same pupil-performance objectives, over the same period of time. (Samples like these are usually unavailable in the public schools.)

2. The objectives and tests were to be such that no student teacher or pupil could come to the teaching situation with an advantage. (This is essential to the validity of the performance test.)

3. The objectives were to be at a level which would generate reliable differences between the class mean scores of the student teachers, in a test-retest sense.

4. The pupil-group learning abilities were to be arranged randomly so that no student teacher would have an initial advantage in the teaching situation, and so that gain scores could be supplanted by the post-test-only design. (The logistics of arranging equivalent pupil-group learning abilities are usually impractical in the public schools for long periods of time; the complications of adjusting gain scores for individual differences are just as distressing.)

The intervening variables previously listed (pupil and teacher tasks and pupil variables) were controlled, to afford each stu-

[1] T. B. Justiz. "A Method for Identifying the Effective Teacher." Doctor of Education Dissertation. Los Angeles: University of California, August 1968.

dent teacher an "equal opportunity" to produce effective results. These conditions also provided the researcher with control over all relevant variables other than the effects of the teaching itself, as follows:

• In order to establish a sample of student teachers, all teaching to the same pupil-performance objectives, over the same time period, the following procedures were implemented:

General teaching abilities were measured for 10 student teachers in the first of two senior high schools, with seven student teachers in the second school. All student teachers were enrolled in the UCLA internship training program. Student teachers were selected for their unfamiliarity with two different subject fields (News-Story Structure and Punched-Card Computer Concepts). Each student teacher was then supplied with a packet containing the two subjects in "Kit" form. Each "Kit" contained objectives, related subject matter (necessary resource information for teaching the lesson), and practice exercises (including distractors which had no clear relationships to the objectives). Student teachers were instructed to select practice exercises appropriate to each objective, to prepare the lessons overnight, and to use any means whatever for bringing about pupil achievement of the objectives. (The procedure was, therefore, to evaluate student teacher performance without restricting teacher style.) On the following day, all student teachers instructed (without observers in the classroom) for 30 minutes in each subject, then were given paper and pencil post-tests and 15 minutes for testing (under close supervision).

• In order to establish objectives and tests which are such that no student teacher or pupil comes to the teaching situation with an advantage, the following procedures were implemented:

Objectives were selected which were not currently taught in the senior high schools, reducing the likelihood of previous student or teacher exposure to the materials. The objectives required no specific entry behaviors and were not dependent on previous instruction. The objectives were content validated for historical time value, so that students and teachers could see purpose in the subject matter. The objectives were in the cognitive domain and were process validated, that is, operationally defined and classified at the same cognitive level of the Bloom taxonomy (the objectives, in the two subject fields, were selected from the comprehension level of Bloom's cognitive *Taxonomy*).[2]

Exhaustive prevalidation pilot studies were conducted to ensure the face validity of the instruments, the timing of each aspect of the procedures, and the ceiling effects of the post-tests. (The post-test ceiling is the difference between the maximum possible post-test score, and the distribution of the pretest scores. If a pretest score is high, a teacher cannot possibly produce much of a change on the post-test. On the other hand, if there are many maximum post-test scores, the true effectiveness of the teacher is not measured.) A 15-minute post-test was considered adequate for the purpose of the instrument, that is, to provide a teacher-training coordinator with an indication of the general teaching ability of his student teachers. Pupils were presented with identical post-tests in each subject field, and an equal amount of time for testing. The possible contamination of test results was controlled by presenting each teacher with explicit instructions for the distribution and use of the pre- and post-tests, and through close supervision of the testing procedures.

• In order to establish objectives at a level which would generate reliable differences between the class mean scores of the student teachers, in a test-retest sense, the following procedures were implemented:

Comprehension skill objectives were chosen for each of the two subject fields (according to the Bloom taxonomy definitions). Comprehension skills require problem-solving abilities, and were considered more of a

[2] B. S. Bloom. *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain*. New York: David McKay Company, Inc., 1966.

teaching challenge than lower-order store-recall skills. (This was an insurance factor, so that the objectives would generate differences between class mean scores.)

• In order to establish pupil-group learning abilities so that no student teacher has an initial advantage in each of the two teaching situations, the following procedures were implemented:

On the morning after the evening of lesson preparations, each student teacher randomly selected a group of 18 experimental pupils from his training-teacher's class (12 pupils in the second school), and escorted them to a testing area where the pupils were distributed (two by two) into as many different classrooms as there were student teachers. The pupil groups were, thus, reconstituted to reduce the initial differences between classes. In the first of two schools, each student teacher selected 18 pupils from his original class and distributed the pupils into nine classrooms (two pupils per classroom). The student teacher was assigned to the tenth classroom, which did not contain any of his own pupils. Thus, the bias of teaching to "known" pupils was eliminated, in addition to providing groups of randomized learning abilities. In the second school, each student teacher selected 12 pupils and distributed them into six classrooms, leaving the seventh classroom for his own occupancy. (Pupils were therefore chosen in multiples of one less than the total number of student teachers used in the experiment.)

The diagram in Figure 1 indicates the teacher-classroom-pupil combinations at the first (10-classroom) high school. Each square on the diagram, containing the digits 1 through 10, represents two pupils. The number in each square is the number of the student teacher who brought the pupils to that room. For example, student teacher 1 (assigned to room 1) deposited his pupils in rooms 2 through 10, then taught the News-Story Structure lesson to pupils from T4, T5, T6, T7, T8, T9, T10, T2, and T3 in room 1 during the first period. These pupils then moved to the rooms shown vertically for period 2, that is, the two pupils from T4 (shaded square) moved from room 1 to room 2, the two pupils from T5 moved from room 1 to room 3, etc. All student teachers remained in their own rooms. For example, student teacher 1 remained in room 1 (for period 2) and received students from T3, T4, T5, T6, T7, T8, T9, T10, and T2 (shown vertically across the top of the diagram). These conditions were set up to take out most of the errors due to initial class differences, prior to the experiment, without having to rely on statistical manipulations (such as analysis of covariance) usually performed after the data are collected. (Pupils were reconstituted a second time, between lessons, at the first school.)

Period 1 News-Story Structure Concepts										
	Room 1	Room 2	Room 3	Room 4	Room 5	Room 6	Room 7	Room 8	Room 9	Room 10
Room 1		3	4	5	6	7	8	9	10	2
Room 2	4		5	6	7	8	9	10	1	3
Room 3	5	6		7	8	9	10	1	2	4
Room 4	6	7	8		9	10	1	2	3	5
Room 5	7	8	9	10		1	2	3	4	6
Room 6	8	9	10	1	2		3	4	5	7
Room 7	9	10	1	2	3	4		5	6	8
Room 8	10	1	2	3	4	5	6		7	9
Room 9	2	5	6	8	10	3	4	7		1
Room 10	3	4	7	9	1	2	5	6	8	
Period 2 Punched-Card Computer Concepts										

Figure 1. Teacher-Classroom-Pupil Combinations in Two Consecutive Teaching Situations

(The present design also allows for post-experimental adjustment of error, by comparing the scores of all pupils taught by the first student teachers, against the scores achieved by these same pupils with the other student teachers during the second period.)

Given the randomized conditions above, each student teacher began instruction with equivalent pupil-group learning abilities. Gain scores and the usual adjustments for initial differences were, therefore, not required. The mean post-test scores were used as the cri-

terion of teacher effectiveness. The mean pretest scores of the control groups (those pupils who remained in the original classrooms after the 18 [or 12] pupils had been randomly selected out) were recorded only to provide evidence that learning had taken place, that is, that equivalent students had not possessed the desired behaviors prior to instruction. (This post-test-only position is supported by Campbell and Stanley, as follows: "The most adequate all-purpose assurance of lack of initial biases between groups is randomization. Within the limits of confidence stated by the tests of significance, randomization can suffice without the pretest." [3]

Pupil and Teacher Tasks and Pupil Variables were, therefore, controlled in this experiment, leaving the Teacher Variables to produce the differences in the pupil achievement scores. One Teacher Variable was selected out so as not to impact the criterion test scores. This was the teacher's knowledge of subject matter. (A screening questionnaire was administered so that no student teacher would have an advantage in either of the two different subject fields.) The criterion of teacher effectiveness was, therefore, considered a measure of general teaching ability, in that what each student teacher brought to the teaching situation was something other than a knowledge of subject matter.

Another Teacher Variable (teacher attitudes toward pupils and teaching, measured with the MTAI) was measured as an Independent Variable to establish a relationship between teacher attitudes and pupil achievement.

The specific research hypotheses were as follows:

Hypothesis 1: Given, a group of student teachers, all teaching to the same two subject matter objectives, each objective being in a different subject field, each objective being taught to a different group of pupils, and all student teachers being ranked (twice) according to the mean of their pupils' post-test scores for each subject field, the Spearman Rank-Difference Correlation Method, when applied to the two rankings, will produce a coefficient which will be significantly different from zero.

Hypothesis 2: Given, a group of student teachers, all ranked according to the mean of their pupils' post-test scores, in two different subject fields (as in Hypothesis 1), and also ranked according to their scores on the Minnesota Teacher Attitude Inventory (MTAI), the Spearman Rank-Difference Correlation coefficient for the MTAI and each subject field will be significantly different from zero.

(The criterion of general teaching ability was, therefore, pupil performance in subjects with which both the teachers and pupils were unfamiliar.)

Findings

After the data were collected, each student teacher was ranked according to the mean score generated by all pupils in his class, for each of the two different subjects taught. The two rankings were then correlated, using the Spearman Rank-Difference Correlation Method.[4, 5] The correlations were statistically significant at the .05 level of confidence (.632/.600 and .900/.900), at schools respectively, that is, most of the student teachers who were effective in one subject were as effective in the second subject as well. (See Figure 2 for student teacher rankings.)

Student teachers were also ranked according to their scores on the Minnesota Teacher Attitude Inventory. The MTAI rankings were then correlated with each of the two different subject rankings. The correlations were statistically significant at the .05 level of confidence (.690/.640, .717/.600,

[3] D. T. Campbell and J. C. Stanley. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally & Company, 1966. p. 25.

[4] J. E. Morsh, G. G. Burgess, and P. N. Smith. *Student Achievement as a Measure of Instructor Effectiveness*. Research Bulletin AFPTRC-TN-55-12. Lackland Air Force Base, San Antonio, Texas, 1955. p. 8.

[5] J. P. Guilford. *Fundamental Statistics in Psychology and Education*. New York: McGraw-Hill Book Co., 1965. p. 593.

1.000/.900, and .900/.900) respectively, at each school.

Conclusions

The following conclusions are considered relevant to the data collected in this study:

1. The ability of student teachers to produce pupil achievement can be measured reliably. It was demonstrated above that the student teacher rankings in two consecutive teaching situations were not the results of chance. The measure may, therefore, be used to identify student teacher ability to produce pupil achievement at the senior high school grade levels.

2. The general teaching ability (GTA) of student teachers can be reliably measured in terms of pupil achievement. (GTA is defined as that ability which the teacher brings to the teaching situation other than a knowledge of the subject matter.) It was demonstrated above that the student teacher rankings in two different subject fields were not the results of chance. The measure may, therefore, be used to identify general teaching ability (GTA) at the senior high school grade levels.

3. There is a relationship between student teacher attitude (as measured with the MTAI) and pupil-achievement-producing ability. It was demonstrated above that the student teacher MTAI rankings relating to the rankings of the student teacher in each of two consecutive teaching situations were not the results of chance. The performance tests of student teacher effectiveness may, therefore, be used to identify predictors of student teacher effectiveness (for example, the MTAI).

4. There is a relationship between student teacher attitude (as measured with the MTAI) and general teaching ability. The student teachers taught objectives in two different subject fields to two different groups of pupils, without prior familiarity with either subject, and it was demonstrated above that the MTAI and subject field rankings were not the results of chance.

Summary and Implications

The foregoing study appeared to generate the first reliable measure of general teaching ability (based on pupil performance in two different subject fields) in that all student teachers were selected for their unfamiliarity with the two different subject fields taught. The study also demonstrated a relationship between student teacher attitudes (as measured with the MTAI) and general teaching ability. The MTAI, thus, appeared to be a reliable predictor of the pupil-achievement-producing abilities of senior high school student teachers.

Whereas the foregoing study was designed as a practical method by which a teacher-training coordinator or high school principal could compare the general teaching ability of several student teachers at the same

		1st School									2nd School				
Student-Teacher Number		1	2	3	4	5	6	7	8	9	1	2	3	4	5
MTAI	Teacher Attitudes	42	87	63	65	44	43	39	1	29	79	76	70	54	−5
Experimental Groups Post-Tests	News Story Means	81½	71½	67½	66	61	57½	57.3	57.2	57	70	66	59½	56.7	51.5
	Punched Card Means	64½	70½	55	82	43	56	44	55½	30	57½	71½	51½	47½	47½
Control Groups Pretests	News Story Means	40	54	44	39	61	56	41	41	52	—	—	—	—	44
	Punched Card Means	19	7	17	—	24	16	13	9	—	—	—	—	—	11

Figure 2. Student-Teacher Rankings

time, the design is also adaptable for regular teachers, or any combination of regular teachers, student teachers, and/or non-teachers, etc.

The operational definition of general teaching ability was supported by six assumptions, as follows:

1. That student teachers differ with respect to their abilities to produce pupil achievement of subject matter objectives

2. That a student teacher's ability to produce pupil achievement can be measured using paper and pencil post-tests of subject matter objectives

3. That a student teacher who successfully accomplishes pre-stated objectives in subject fields other than his own will be likely to achieve at least equal success with his own objectives

4. That studei t teachers can be ranked on the basis of their pupils' scores on achievement tests, providing that no teacher is familiar with the subject prior to the time he is confronted with the teaching situation

5. That a positive and statistically significant correlation of the rankings of student teachers, for two subject fields, demonstrates that the rankings are reliable

6. That a reliable measure of student teacher effectiveness in two different subject fields is a measure of general teaching ability, and can be used to identify the effective student teacher, regardless of the teacher's background, subject specialty, or organizational status.

The findings in the foregoing study were based on the following general principles of teacher effectiveness:

1. Pupil change is the ultimate criterion of teacher ability.

2. Teacher ability may be assessed when all teachers in a group are provided with an equal opportunity to produce effective results. This includes identical objectives over time, reliable instruments, and equivalent pupil-group learning abilities (that is, control over all relevant variables other than the effects of the teaching itself).

3. The validity of the pupil achievement criterion depends on the discriminating power of the objectives, and the consistency of the teacher rankings (based on class mean scores), that is, reliability.

This study should tend to allay some of the time-honored mistrust of pupil performance criterion measures as the basis of teacher effectiveness. The following remarks relate to the nine excerpts from research articles which were critical of pupil performance criterion measures:

1. A reliable procedure for assessing the competency of the student teacher has been developed, regardless of teacher style or method.

2. The influence of the teacher over time may be measured when all relevant variables other than the effects of the teaching itself are controlled.

3. Pupil gain may be measured reliably using the post-test-only design, by arranging equivalent pupil-group learning abilities through randomization, and the elimination of the bias of teaching to known pupils.

4. Teachers can be induced to teach to identical objectives by providing identical "Kits," containing objectives, related subject materials, practice exercises, instructions, and related testing instruments.

5. Test imperfections may be reduced through prevalidation pilot studies, which can be used to modify objectives, post-test ceilings, the relevancy of practice exercises, procedure timing, and the possible contamination of test results.

6. Pupil growth, which varies with individual differences in pupils, tends to make pupil gain scores unreliable; however, mean class scores are only slightly affected by pupil growth when pupil groups are chosen randomly.

7. The post-test-only design, used as a substitute for pupil gain scores, can achieve very high reliability coefficients and considerable validity.

8. Teacher effectiveness may be assessed reliably in a period of only two hours. Set-up time will take an additional day (with this design).

9. Tests always measure effects, and causation is rapidly becoming an invalid term. The true importance of measured effects is revealed only when they show relationships with other measured effects. □

Copyright © 1969 by the Association for Supervision and Curriculum Development. All rights reserved.