Teacher Evaluation Via Videotape:

Hope or Heresy?

Philip Hosford and John Neuenfeldt

Continuing research may produce a way to evaluate teaching performance objectively using videotape.

Would you want this teacher for your son or daughter next year?

If this question refers to teachers we know and work with, we have ready answers. We know the "good" teachers, and we know they will make a difference to the child. But we need to examine the kinds of differences that we value.

There have been many studies during the past 50 years examining the characteristics associated with "good" and "poor" teachers. Besides a set of desirable characteristics associated with all "good" people in any profession—the "best" doctors, lawyers, and preachers, as well as teachers—an intuitively perceived x-factor helps principals, supervisors, and professors decide those who are "really good teachers."

Some colleges and universities provide administrators the opportunity to view a videotape of graduating candidates. Superintendents, directors of personnel, and principals indicate high degrees of satisfaction with the tapes and report that they can "tell much more about the candidate" by viewing a five-minute tape than they can in the usual interview process. Indications are that the x-factor operates importantly in the viewing and selecting process.

Most professionals working with student teachers know the uncertainties of evaluating performance. Sometimes, an uncomfortable admission is made that the x-factor influenced a final assessment. Certainly, if performance-based teacher certification programs are to be valid, the gap between all such professional judgments and the limited range of decisions permitted by research must be bridged. The trend toward more and earlier field experiences in formal teacher

preparation programs offers controls not previously possible. These controls permit serious study and analysis of the x-factor. Results of several pilot studies, a dissertation, and current research are reported here to support our contention that objectively gained judgments should play a major role in teacher evaluation processes.

Rationale

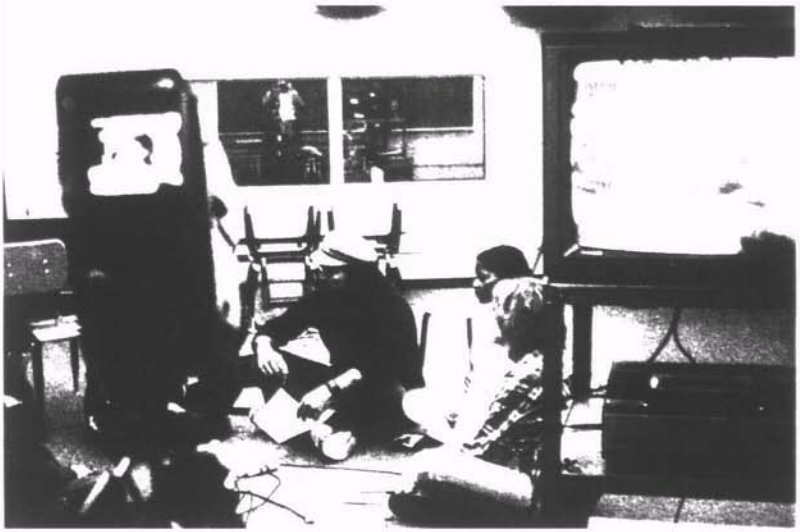
If any two teachers are selected with similar academic preparation—teaching in the same school, with the same subject matter, and with similar groups of learners—the achievement test scores of their groups ordinarily will not differ in any significant way. The scores will differ significantly when these two groups are compared to any other groups in that school who did not study the subject matter tested. That is, if the two teachers are chemistry teachers, then their classes will score significantly higher on a chemistry achievement test at the end of the year than will groups who did not take chemistry. The school and the teachers make a significant difference, and one that is valued. But to examine the test scores to determine which one of these teachers is superior to the other is an abomination in the eyes of teacher organizations in addition to being expensive and time consuming.

Still, we might agree that one teacher is a far better teacher than the other based on our anonymous responses to the question of which teacher we would want for our child next year. One teacher may be better because students of that teacher learn other things that we value in addition to the measured academic achievement, and the students end the year with a good feeling toward the school subject matter, themselves, and others. They have a vision of the structure and possible use of their knowledge. They may even elect to study the field further.

We know the other teacher is not equal to the first even though the students of this teacher do pass the tests. The trouble is that they also learned other things during the year. They may have learned to cheat, to distrust others, and to view learning as strictly an academic exercise unrelated to the world outside of school.

Perhaps this illustrates an extreme case, but it is one most of us have known in our own experience, and it speaks to our values. In Q-sorts run with over 300 teachers, aides, and administrators, four objectives of public school education are consistently selected as the most important: (1) achievement in the 3R's; (2) desire for learning; (3) healthy self-concept; and (4) respect for others (Hosford, 1975). Growth in the last three areas is the consequence of teacher performance in the process of instruction. This is the "silent curriculum" each of us creates in our own classroom during the year. Significant portions of learner differences in the "non-content" areas of improved

"Superintendents, directors of personnel, and principals indicate high degrees of satisfaction with the tapes. . . ."



self-concept, desire for learning, and respect for others can be explained by differences in teachers. These are the differences that dictate our response of "yes" or "no" to the question of wanting a given teacher for our child next year.

Performance abilities in these silent curriculum areas are not easily measured; therefore, we shy away from attempts to include judgments in these areas in any formal evaluation system. Most of us have long felt that others would be unwilling to render such judgments anyway. But this intuitive approach to the problem seems wrong.

The Pilot Studies

Six Practicing Teachers

In a 1973 pilot study, six teachers were filmed for five minutes each while teaching a fifth-grade class. The films were shown to fifth-grade groups from two other elementary schools, undergraduate teacher preparation students, and professional educators (teachers, supervisors, and professors).

The fifth-grade students were asked only: "Would you want this teacher to be your teacher next year?" Undergraduates and educators were asked: "Would you want this teacher for your child or little brother or sister?" Each rater had a choice of checking *yes*, *maybe*, or *no*. After examining the data, several conclusions seemed warranted:

1. All groups discriminated, but professional educators were the most definite in their consensus.

2. Groups of undergraduates and groups of professional educators achieved internal agreement. Fifth graders did not.

3. Group consensus of undergraduates and of professional educators was in agreement. Fifth-grade groups were not in agreement.

Viewer Consensus by Educational Level

A second study (Hosford-Schroder, 1974) was based on videotapes of 108 preservice undergraduates in microteaching situations. In this study, only two specific objectives were sought. The first was to determine to what degree people involved with teachers at different levels might achieve consensus in their reactions to the videotaped teaching segments. The second objective was to determine if the evaluations of prospective teachers made by five selected ninth graders would correlate highly with other groups and be reliable enough to serve as sufficient and representative evidence in the evaluation process of prospective teachers. Once again, a high level of agreement was found within each group of raters. Second, the five selected ninth-grade students were apparently sufficient to represent evaluations by large groups of secondary students. As shown in Figure 1, correlations of ratings by the five ninth graders with both ninth-grade and senior high classes were significant at the .05 level.

Figure 1. Correlations of Ratings by Five Ninth Graders With Other Groups

Other groups	Five ninth graders
Ninth grade class (N=24)	.80*
Senior high class (N=21)	.87*
Master's degree students (N=12)	.34
Teachers (N=13)	.41
Doctoral students and professors (N=5)	.37

*Significant at or beyond .05

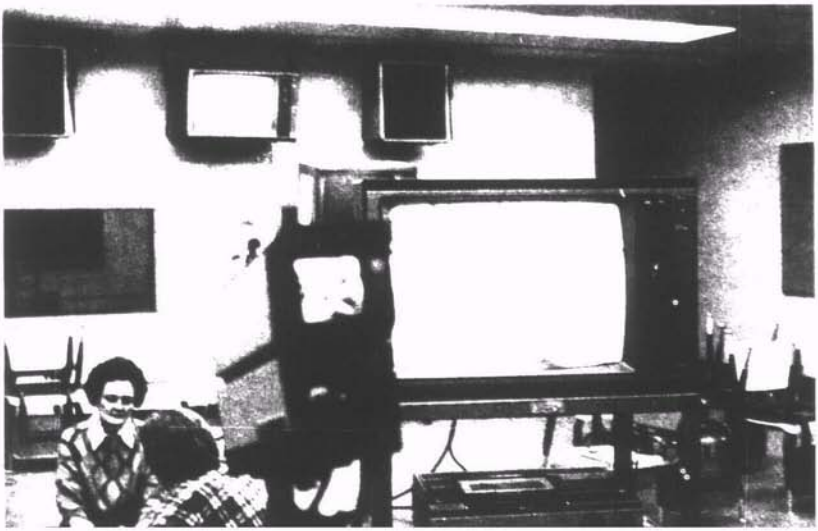
Many valuable questions were derived from these two pilot studies. Whose predictions of teacher performances are more accurate—students or professional educators? Are either predictions valid? Will either correlate with later performance? What characteristics are being valued in making the evaluations? What about variables such as sex, age, and educational level?

The Neuenfeldt Study

A third study (Neuenfeldt, 1977) was designed to answer many of the preceding questions. In the first stage of the study, two videotapes were edited so that each tape contained five teachers teaching four-minute segments. Both tapes contained experimental and control performances known to prompt strong "yes" or "no" consensus among viewer groups. These tapes were viewed by over 300 raters including teachers, student teachers, and public school students.

Stage two involved the use of videotapes made of eight junior and senior high school teachers in one of their regular classes, and videotapes of six college mathematics teachers each teaching a regular class.

"Our evidence indicates that four-minute videotape segments may provide the best, most practical diagnostic tool available to supervisors in both preservice and inservice programs."



Generalized findings of the study include:

1. Professional educators easily achieve high "yes" and "no" consensus in their evaluation of videotaped teaching segments. Significant agreement in evaluations is achieved by all such groups ($N>5$) whenever they view segments that prompt definite "yes" or "no" reactions from professional educator groups.

2. When strong "yes" or "no" consensus is attained by professional educators the variables of sex, age, and grade level play no significant part.

3. Student groups do not produce tally distributions comparable to professional groups. As the maturity of the student evaluators increases, their tally distributions become more like those of all professional groups.

As a final dimension of the Neuenfeldt study, ratings were obtained from supervisors of teachers independent of any and all videotape procedures. These nonvideo supervisor ratings were obtained through the regular procedures used by each school organization and then were compared with those generated from the videotape evaluations. Two more generalized conclusions resulted.

First, in all cases, the supervisor evaluations agreed with the evaluations previously obtained through videotape procedures. However, since some cases were not statistically conclusive, more study is needed relating video and nonvideo evaluations.

Second, consensus is less frequently achieved by all viewer groups when a videotape teaching segment fails to prompt mostly "yes" or "no" reactions.

Summary

Evidence supports the hypothesis that objectively gained judgments obtained via the videotape procedures described here will correlate with those judgments being rendered in the field and in teacher preparation programs. All such judgments are particularly clear and correlate most highly when they relate to videotape segments that commonly prompt definite "yes" or "no" evaluations. Professional educators all obtain high intergroup agreement in their evaluations of the videotape segments. Age, sex, and occupational grade level are not significant variables in either the viewing group or the person being viewed.

However, the following critical questions remain:

1. Will preservice undergraduates identified as definite "yes" or "no" types later confirm those evaluations in their overall college performances and in particular in their student teaching performance?

2. Will videotape diagnosis prove more pre-

dictive than other variables such as overall grade point average, competency tests in cognitive areas, ACT scores, or other available predictors?

3. Will viewer-viewee variables of ethnicity or physical traits significantly affect videotape evaluations?

4. Can videotape evaluations predict end-of-year evaluations of practicing teachers by their principals and supervisors?

5. Will videotape evaluations of teachers correlate with changes measured in the students of those teachers?

6. Will videotapes consistently evaluated as "yes" or "no" provide models valuable for use in preservice or inservice training programs? Will such tapes provide clues to the essential aspects of the x-factor in teaching? Can those aspects be extracted and taught to others?

Research is currently underway dealing with the first three questions. The last three questions are more difficult and will require time, patience, and effort to formulate any definitive answers. Question five is critical. The process-product research needed to answer it is now possible. Recent studies reported by Rosenshine and Berliner (1978), Medley (1977), and Aspy and Roebuck (1977) speak clearly to the issue. The more recent "vote counting" and "cluster analysis" procedures advocated by Gage (1978) will prove helpful.

Our evidence indicates that four-minute videotape segments may provide the best, most practical diagnostic tool available to supervisors in both preservice and inservice programs. Moreover, professional viewer reactions to videotape segments do serve as objective indicators of professional judgments regarding the x-factor of teaching.

References

D. N. Aspy and F. N. Roebuck. *Kids Don't Learn from People They Don't Like.* Amherst, Massachusetts: Human Resources Development Press, Inc., 1977.

N. L. Gage. "The Yield of Research on Teaching." *Phi Delta Kappan* 60:228-35; 1978.

P. L. Hosford. "Inservice Education and the Silent Curriculum." In: R. E. Wright, editor. *Inservice Education Programs to Improve Teaching Competence.* Association of Teacher Educators *Bulletin* 39:1-3; 1975.

D. M. Medley. *Teacher Competence and Teacher Effectiveness: A Review of Process-Product Research.* Washington, D.C.: American Association of Colleges for Teacher Education, 1977.

J. C. Neuenfeldt. "An Investigation of an Alternative Method of Evaluating Classroom Teaching." Doctoral dissertation, New Mexico State University, 1977. *Dissertation Abstracts International* 38:7121A-122A; 1978.

B. V. Rosenshine and D. C. Berliner. "Academic Engaged Time." *British Journal of Teacher Education* 4:3-16; 1978.

Philip L. Hosford (left) is Professor of Education, College of Education, New Mexico State University, Las Cruces; John Neuenfeldt is Assistant Professor, Department of Mathematics, University of Wisconsin, Stout, Menomonie.