DOCUMENT RESUME

ED 274 704 TM 860 581

AUTHOR	Cronin, Linda; Capie, William
TITLE	The Influence of Daily Variation in Teacher Performance on the Reliability and Validity of Assessment Data.
PUB DATE	18 Apr 86
NOTE	20p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Behavior Change; *Behavior Rating Scales; Classroom Observation Techniques; Evaluation Methods; Grade 7; Interrater Reliability; Junior High Schools; Secondary School Teachers; *Teacher Behavior; Teacher Effectiveness; *Teacher Evaluation; *Test Reliability; *Test Validity; *Time
IDENTIFIERS	Group Assessment of Logical Thinking; Middle Grades Integrated Process Skills Test; *Teacher Performance Assessment Instruments

ABSTRACT

The influence of day-to-day variation in teacher performance on the reliability and validity of teacher assessment was examined. An attempt was made to identify and quantify sources of score variation attributable to differences in teacher performance, day of observation, observers, and test subscales; and to determine their effects on reliability and validity of decisions made. Data were collected from a field test of the revised Teacher Performance Assessment Instruments (TPAI). Thirty-nine seventh grade science teachers in a large Georgia school district were observed twice: (1) by two observers at the same time on the same day; and (2) by two observers on different days. Each teacher taught a two-week unit on scientific problem solving; student achievement was measured with pretests and (the Group Assessment of Logical Thinking) and posttests (Middle Grades Integrated Process Skills Test. The estimated expected posttest score was substracted from the actual score to provide an index of teacher effectiveness. A four-facet fully-crossed design was used, including teachers, observers, day of observation, and performance indicators as sources of variation. Results suggested that additional observation time enhanced validity, and that daily variation was a greater source of error than observer differences. (GDC)

THE INFLUENCE OF DAILY VARIATION IN TEACHER PERFORMANCE ON THE RELIABILITY AND VALIDITY OF ASSESSMENT DATA

Linda Cronin
William Capie

Teacher Assessment Project
The University of Georgia
Athens, GA 30602

A paper presented at the annual meeting of the
American Educational Research Association
San Francisco, CA
April 18, 1986

THE INFLUENCE OF DAILY VARIATION IN TEACHER PERFORMANCE ON THE RELIABILITY AND VALIDITY OF ASSESSMENT DATA

In recent years, concern about the qualifications and credibility of teachers has been repeatedly expressed by the media, as well as politicians, policy makers and practitioners. In fact, on any given day, in any major periodical or newspaper, at least one article can be found which focuses on issues associated with educational reform, teacher training, teacher evaluation, and the like. This widespread national concern has sparked the development of teacher certification tests and/or teacher assessment programs in several states, including Florida, Georgia, Tennessee, and Virginia.

Most statewide teacher assessment programs have focused on the assessment of teacher performance for purposes of initial certification or promotion/reward. Although the expectations associated with the two contexts may differ, several concerns regarding instrumentation of performance assessment systems apply to both situations. Examples of pertinent questions include the following: (1) How many times should a teacher be observed?; (2) Who should do the observing?; (3) Should the observers be in the room at the same time?; (4) Should the teacher know when observers are coming?

Although common sense, as well as practical and political concerns, plays a role in decision making regarding these questions, studies of the reliability and validity of scores and decisions made using a particular assessment instrument under various conditions provide the soundest basis for determination of

appropriate answers to questions such as those posed above. As Caple and Cronin (1986) have indicated, "the bottom line in designing an assessment system is that the scores which are generated must be credible" (p.2).

Both common sense and political logic indicate that decisions regarding teacher certification and/or promotion should not be made on the basis of a single assessment. The general view is that a teacher should be observed by more than one observer on more than one occasion. In actuality, when multiple observers are used in teacher performance assessment programs, they do in fact visit the classroom on separate days. Doing so provides a larger sample of teacher behaviors than does a single visit. However, when variation occurs in the scores generated from occasion to occasion, uncertainty exists regarding the source of variation in these scores. The primary concern in whether the variation in scores is due to variation in teacher performance or due to error attributable to observer differences.

PURPOSE

The purpose of this study was to determine the influence of daily variation in teacher performance on the reliability and validity of assessment data. Specifically, the study was conducted in an attempt to identify and quantify sources of variation in scores attributable to differences in teacher performance, day of observation, observers, and subscales of the assessment instruments and to determine the effects of these sources of variation on the reliability and validity of decisions made.

DATA SOURCES

Data collected from a winter 1985 field test of the revised Teacher Performance Assessment Instruments (TPAI) provided an opportunity to investigate questions associated with the influence of different sources of variation on the validity and reliability of decisions made. The field test sample consisted of all seventh grade science teachers (N=40) in a large Georgia school district including schools in suburban Atlanta as well as more rural outlying areas. Thirty-nine of the participating teachers were observed by two observers at the same time on the same day as well as by two observers on different days. These two data sets provided the researchers with an opportunity to address the ever-ambiguous issue of occasion effect. In most previous studies involving the TPAI, it was not possible to separate occasion effects from observer effects since different observers went into classrooms on different occasions. Past analyses of the TPAI have been based on the assumption that all of this ambiguous variation is error, and as Yap and Caple (1985) have indicated, it is important to find out just how much of the apparent observer variance represents instability of teacher behavior.

CONTEXT

Each teacher taught a prepared two week unit focusing on science problem solving and experimenting. In addition, all participants administered a common post test after instruction. The use of teachers from a similar field, the common unit, and the common post test each represented an effort to reduce context variation. The lessons contained in the unit were intended to be

somewhat demanding for teachers, a fact which was expected to lead to increased variation in teacher TPAI scores.

INSTRUMENTATION

The revised version of the TPAI used in this study consists of eight subscales or competencies. Each competency is a decision making unit used to make pass-fail decisions about a teacher's performance. Each competency is defined by three or four subordinant items, called indicators, which are in turn broken down into specific descriptions of behavior, called descriptors. The complete revised TPAI contain 30 indicator statements and 120 descriptor statements. However, since plans and formal assessment materials were not prepared by the teachers participating in this study, the seven indicators related to these areas were not included in the analyses.

When evaluating teacher performance using the TPAI, each observer responds to each descriptor statement indicating whether or not it was present to an acceptable level in the observed lesson based on criteria specified in the instrument and learned during training. Descriptor data are aggregated into indicator scores which are then aggregated to form competency scores for decision making. The logic of the relationship among descriptors, indicators and competencies has been confirmed in an extensive content validation study (Cronin & Capie, 1985). A list of the eight competencies and their 30 constituent indicators is displayed in Figure 1.

INSERT FIGURE 1 ABOUT HERE

ANALYSES

The analyses had four critical components.

TPAI Scores. The TPAI were scored using procedures specified in the instruments. Descriptor scores of Ø (no credit given) or 1 (credit given) were assigned by each observer. These scores were then aggregated to form indicator scores which had values ranging from 1 (no descriptors scored acceptably) to 5 (all four descriptors scored acceptably). Competency scores were then computed which reflected the portion of indicators scored at or above minimum acceptable level by each trained observer.

Learner Achievement. All learners were assessed with the Group Assessment of Logical Thinking (GALT) prior to instruction and with the Middle Grades Integrated Process Skills Test (MIPT) after instruction. Regression techniques were used to generate expected post-test scores for each learner based on his/her ability as measured by the GALT and the correlation between GALT and MIPT (r=.62). For each learner, the expected post-test score was subtracted from the actual observed post-test score. This difference was considered to be a teacher effect on the learner. The means of these "teacher effects" for each class were used as a teacher effectiveness index. Thus, a variable reflecting class means was available for use in subsequent analyses where classes were considered to be the sampling units. In subsequent analyses the teacher effectiveness index was considered to be the criterion variable.

Validity Indices. Simple correlations were computed between indicator scores and the teacher effectiveness index as well as between competency scores and the teacher effectiveness index.

These correlations were computed for data sets composed of observations by two observers on the same day and two observers on different days.

Generalizability Analyses. Generalizability theory was used to plan the analyses of the TPAI data. Five factors were identified as important sources of variation: teachers, individual observers, observer type, day of observation, and performance indicators. The five facet design with individual observers nested within observer types is arithmetically identical to the simpler four facet fully-crossed design with teachers, observer types, day of observation, and indicators as sources of variation. As a consequence, the simpler four facet model was used. For each analysis, teachers were considered to be the facet of differentiation and the other facets were treated as random facets of generalization. Values of rho squared and phi(lambda) were computed to assess the suitability of the scores for differentiating teachers from each other and from the ideal standard of having all indicators at or above the minimum level.

Generalizability analyses were conducted with a subset of twenty teachers observed by one external and one school system observer at the same time on two different days. In order to determine the influence of the number of days and/or the number of observers on the dependability of decisions made, the convenient D-study feature of the GENOVA program was used to simulate the effects of observations made on 1, 2, 3, 4, and 5 days with 1 and 2 observers. Total instrument scores for each teacher were used in these analyses.

RESULTS

Correlations between achievement and mean raw indicator scores are displayed in Table 1. For data obtained when two observers were in the classroom on the same day, seven of the 23 correlations were significant ($p < .05$). Data obtained from observers visiting the classroom on different days yielded six significant correlations ($p < .05$) out of 23. For five of the significantly correlated indicators, correlations were higher when observers visited the room on different days. However, three significantly correlated indicators had higher correlations when observers were in the room at the same time. Only four indicators were significantly correlated with achievement with both same day and different day sets of observation data.

Insert Table 1 about here

Correlations between achievement and mean competency scores are displayed in Table 2. When observers visited classrooms on the same day, two of the seven competencies were significantly correlated with achievement ($p < .05$) while five of the seven competencies were significantly correlated with achievement when observers visited classrooms on different days. The two competencies not significantly correlated with achievement in either data set were Competency 3 (Demonstrates acceptable written and oral expression and command of subject matter) and Competency 5 (Communicates with learners).

Insert Table 2 about here

The mean portion of indicators scored acceptably by the twenty teachers included in the generalizability analyses was .63.

The variance components associated with the analyses are included in Table 3. Both day effects and observer effects were near Ø while the indicator effect (.Ø22) was one sixth the size of the residual variance and the teacher effect (.Ø15) was approximately one eighth the size of the residual variance. Teacher by day effects and observer by day effects were also zero or near zero.

Insert Table 3 about here

Reliability coefficients generated from simulations involving different numbers of days of observation and different numbers of observers are summarized in Table 4. Values of rho squared ranged from .38 to .63 when data from one observer were used. When data from two observers were used, rho squared valued ranged from .53 with one day of observation to .76 with 5 days of observation. Values of phi(lambda) were higher when two observers were used than when one observer was used, and all values were above .9Ø when three days of observation were involved, regardless of the number of observers. For both one and two observer combinations, values of phi(lambda) were virtually identical when four and five days of observation were involved.

Insert Table 4 about here

DISCUSSION

The principal purpose of this study was to document the extent to which differences among scores generated by observers on different days may be due to instability of teaching performance or due to observer differences. Since observers visit the classroom on different days during assessments and this pattern is followed in most field testing, there has been no way to identify

the relative magnitudes of these effects. Multiple days of observation are planned to enhance the validity of observations as well as the reliability. The potential of observer differences and of day to day variations has been acknowledged in TPAI reliability studies where both of these effects have been treated as error (Yap and Capie, 1985).

The validity coefficients for the TPAI competencies support the contention that additional observation time enhances validity. When the indicator scores and achievement were compared, the number of significant correlations and their magnitudes were similar for both data collection models. For competencies, however, there was a substantial difference. Five of seven correlations were significant ($p < .05$) when observations were made on two days. Only two values were significant for two observers on a single day. These differences could easily be due to increased reliability and could be anticipated by examining the results of the generalizability analyses.

The variance components for the four main effects were essentially as expected. The effects due to indicators and teacher were relatively large, indicating that there was variance associated with each of these facets. The day effect was zero, suggesting that there was no systematic change in the group of teachers from day one to day two. There was a small observer effect, reflecting that external observers were slightly more rigorous in their scoring.

The interactions involving teacher and days are most pertinent to the general question of the study because they reflect the extent to which teachers' performances vary (on

certain indicators or in the eyes of certain observers) from day today. Thus, for example, the variance components show that the teacher x day x indicator effect is nearly three times as large as the teacher x day x observer type effects. Also, the teacher by observer types by indicator effect is substantially smaller than other three way interactions involving teachers. This set of findings suggests that the variation from day to day on particular indicators by particular leaders is a greater source of "error" than are observer differences.

The various reliability coefficients reflect these differences. Of course, increasing either the number of days or number of observers enhances reliability. Values of rho squared fro two observers on a single day or on two independent days are similar (≈ .5). This finding would be modified if each four indicator competency were considered rather than a single total instrument coefficient. With competency scores, the day to day variation in indicator scores would have more impact than it does with the total instrument.

The magnitude of day to day variations in instrument scores must be a concern to program managers, particularly as new uses of these types of instruments are contemplated. Observations of different classes under different conditions (day of week, time of day, drop in visits, etc.) are bound to influence the qualities of the scores substantially. While this study with two days is sufficient to show that there is a problem, it does not speak to many of the types of variation that exist in the measurement context. Clearly, much more should be known about the factors--if only there were time.

CONCLUSION

Single observations made on two separate days were better predictors of achievement than were two observations made on a single day. Variation on indicator scores from day to day was greater than variation from observer type to observer type. Consequently, data collection models involving multiple days of observation are more credible than those based on "one snap shot." However, the number of other potentially important factors influencing scores is so large that much more work should be done before "open" drop in visit systems can be structured.

References

Caple, W., & Cronin, L.C. (1986). A career ladder begins at the bottom. Science Education (in press).

Yap, K., & Caple, W. (1985). The influence of same day or separate day observations on the reliability of assessment data. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April.

Summary of TPAI Organization

Planning

I. PLANS INSTRUCTION TO ACHIEVE SELECTED OBJECTIVES

1. Specifies or selects learner objectives for lessons.
2. Specifies or selects learning activities.
3. Specifies or selects materials and/or media.
4. Plans activities and/or assignments which take into account learner differences.

II. OBTAINS INFORMATION ABOUT THE NEEDS AND PROGRESS OF LEARNERS

5. Specifies or selects procedures or materials for assessing learner performance on objectives.
6. Uses systematic procedures to assess all learners.

III. DEMONSTRATES ACCEPTABLE WRITTEN AND ORAL EXPRESSION AND KNOWLEDGE OF THE SUBJECT

7. Uses acceptable written expression.

Observation

IV. ORGANIZES TIME, SPACE, MATERIALS, AND EQUIPMENT FOR INSTRUCTION

8. Attends to routine tasks.
9. Uses instructional time efficiently.
10. Provides a physical environment that is conducive to learning.

11. Assesses learner progrsss during the lesson observed.

12. Uses acceptable written expression with learners.
13. Uses acceptable oral expression.
14. Demonstrates command of school subject being taught.

V. COMMUNICATES WITH LEARNERS

15. Gives explanations related to lesson content.
16. Clarifies explanations when learners misunderstand lesson content.
17. Uses learner responses or questions regarding lesson content.
18. Provides information to learners about their progress throughout the lesson.

Observation

VI. DEMONSTRATES APPROPRIATE INSTRUCTIONAL METHODS

19. Uses instructional methods acceptably.
20. Matches instruction to learners.
21. Uses instructional aids and materials during the lesson observed.
22. Implements activities in a logical sequence.

VII. MAINTAINS A POSITIVE LEARNING CLIMATE

23. Communicates personal enthusiasm.
24. Stimulates learner interest.
25. Demonstrates warmth and friendliness.
26. Helps learners develop positive self-concepts.

VIII. MAINTAINS APPROPRIATE CLASSROOM BEHAVIOR

27. Maintains learner involvement in instruction.
28. Redirects learners who are off-task.
29. Communicates clear expectations about behavior.
30. Manages disruptive behavior.

Figure 1.

Table 1

Correlation of Achievement with Mean Raw Indicator Scores
(N=39)

Indicator	2 Observers Same Day	2 Observers Different Days
8	.02	.10
9	.17	.25
10	.53*	.54*
11	.00	-.17
12	.14	.11
13	.05	.06
14	-.06	-.10
15	-.03	.07
16	.29*	.15
17	.11	-.01
18	.00	.16
19	.13	.18
20	.32*	.34*
21	.29*	.33*
22	.01	-.20
23	.18	.28*
24	.08	-.10
25	.30*	.34*
26	.16	.19
27	.34*	.20
28	.37*	.21
29	.23	.27*
30	.21	.06

* $p < .05$

Table 2

Correlation of Achievement with Mean Raw Competency Scores (N=39)

Competency	2 Observers Same Day	2 Observers Different Days
2	.Ø8	.26*
3	.Ø5	.16
4	.43*	.28*
5	.17	.14
6	.28*	.33*
7	.22	.26*
8	.2Ø	.31*

* p < .Ø5

Table 3

Variance Components for Fully Crossed Design
(T=20, D=2, O=2, I=23)

Source	Variance Component
T (Teacher)	.015
D (Day)	.000
O (Observer)	.003
I (Indicator)	.022
TD	.000
TO	.004
TI	.007
DO	.003
DI	.000
OI	.000
TDO	.012
TDI	.033
TOI	.003
DOI	.001
TDOI	.134

Table 4

Reliability Coefficients for Simulated Data Collection
(T=20, I=23)

Number of Days	Number of Observers	ρ^2	$\phi(\lambda)$
1	1	.38	.82
1	2	.53	.89
2	1	.50	.63
2	2	.66	.93
3	1	.57	.90
3	2	.71	.94
4	1	.60	.91
4	2	.74	.95
5	1	.63	.91
5	2	.76	.95