DOCUMENT RESUME

ED 117 164 TM 005 019

AUTHOR Fanslow, Alyce M.; Wolins, Leroy
TITLE Assessment of Observational Items for Use in Competency-Based Teacher Education Programs.
PUB DATE 75
NOTE 15p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Washington, D.C., March 31-April 2, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 Plus Postage
DESCRIPTORS *Classroom Observation Techniques; College Supervisors; Cooperating Teachers; *Evaluation; Higher Education; Home Economics Teachers; Item Analysis; *Performance Based Teacher Education; *Student Teachers; Teacher Education; Test Reliability

ABSTRACT

A 50-item observational instrument was used by cooperating teachers and college supervisors to evaluate the competencies of 77 home economics student teachers from two universities at two time intervals during the student teaching experience. Two analyses of variance (AOV) for each of the 50 items were used to identify items which judges could rate reliably. Intra-class correlation coefficients were utilized to ascertain if different ratings between student teachers were due to differences in their performance. The AOV analyses suggested that judges could reliably rate 18 items. Of these, five appeared to discriminate between student teacher's performance whereas eight did not. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

ED117164

Assessment of Observational Items for Use in Competency-Based Teacher Education Programs[1]

Alyce M. Fanslow and Leroy Wolins
Iowa State University

One of the major problems in developing competency-based teacher education programs (CBTE) is the validation of appropriate assessment measures. While this type of research demands large investments of financial and other resources, Rosner and Kay (1974) contend that a most urgent need is for the educational community to recognize and accept vigorous and analytic research as a top priority problem for CBTE programs.

The research summarized in this paper presents one such investigation associated with a fledgling CBTE program--a procedure for assessing the reliability of types of judges to rate student teachers during the student teaching experiences of a CBTE program. The device assessed is a 50 item observational device and types of judges included cooperating teachers and college supervisors.

Instrumentation

The observational device[2] designed for this study involved the measurement of student teacher competencies based on four aspects of the teaching-learning process: classroom performance,

M005 019

[1] A paper presented at the National Council on Measurement in Education Annual Meeting, Washington, D. C., March 31-April 2, 1975.

[2] Observational device developed by faculty members and graduate students in the Department of Home Economics Education, Iowa State University. Selected items were adapted from devices by Thatcher (1969) and Menne (1972).

relationship skills, evaluation skills, and management and professionalism.

The device is one of several proposed within the total professional CBTE program and represents competencies within one of the major areas identified by the Iowa State Home Economics Education faculty as necessary for beginning home economics teachers. When possible, items for the observational device associated with the teaching-learning process were selected from other studies that had previously been used in departmental research or that indicated promise for discriminating between teachers. While some of the items used in the classroom performance, relationship skills, and the management and professionalism sections were adapted from items included in Thatcher's (1969) and Menne's (1972) studies, no instrument was found to assess evaluation skills and this section was developed by Hausafus (1973). Emphasis in the items on evaluation was on the identification of evaluation skills desirable for student teachers.

The completed 50 item observational device included 32 items measuring classroom performance, 11 items assessing relationship skills, 14 items devoted to evaluation skills, and 4 items evaluating management and professionalism.

A 99 point scale was selected for use in responding to the items. The directions explaining the use of the 99 point scale instructed the evaluator to determine if the student teacher observed was functioning below or above average on each specific item and to record the degree of certainty related to each decision. If the student teacher was above average, a number between 51-99

was recorded; if below average, a number between 1-49 was recorded. A 50 indicated that the evaluator was uncertain about the behavior or that there was no opportunity to observe the behavior.

Preliminary Reliability Data

Preliminary reliability assessments based on the total score for the classroom performance section were obtained by having three groups of observers view three 15-minute videotaped micro-lessons taught by home economics student teachers (Gilbert, 1974). Each of the groups viewed one tape, responded to the instrument, and discussed their responses. The observers then viewed and assessed two additional 15-minute videotaped micro-lessons. Reliability coefficients of .87, .89 and .90 were obtained (Winer, 1971, pp. 283-287).

Five training sessions were held to orient the 60 cooperating teachers and the 11 college supervisors who evaluated the home economics student teachers during 1973-1974. A procedure similar to that in the preceding example was followed, i.e., three videotapes of teachers teaching 45-minute lessons were viewed and evaluated using the classroom performance section. After the first videotape, item ratings were discussed. Subsequently, the other two 45-minute videotapes were viewed and rated. A hierarchal analysis of variance for three variables, teacher, judges, and items was computed. A study of the F ratios indicated that while judges used some items differently, overall the judges could discriminate between teachers using the designated items. An analysis on the same data including orientation as an additional source of

variance showed no difference in ratings between orientation sessions.

Even though the three sections on relationship skills, evaluation skills, and professionalism and management were not pretested, assessments by college supervisors indicated the device appeared useful. From these two procedures, the observational device was judged suitable for use in the study.

Sample

The potential sample in this study included 107 home economics education students who were enrolled in the teacher education program at Iowa State University (ISU) and South Dakota State University (SDSU) and who were student teaching during the 1973-1974 academic year. Of these, 68 were enrolled at ISU and 39 at SDSU. Student teachers from both universities were selected because of the potential of a larger sample and basic similarities in the two programs. Similarities between the two programs included the basing of both programs on the objectives and generalizations designated by a representative group of home economics teacher educators as common to all home economics teacher education programs (Kreutz and Anthony, 1966), the same cumulative quality grade point averages as a prerequisite to admittance into the teacher education program, and an eight-week off-campus student teaching experience in the public school at the junior or senior high school levels.

As originally planned, ratings of the student teacher were to be made by two types of judges, the cooperating teacher and the college supervisor, at four, six, and eight week time intervals.

However, due to the energy crisis and the resulting gasoline shortage, it was not possible to visit each student teacher three times as originally planned. Further, the observational plan was confounded by such things as illness of one of the raters or by inclement weather which prevented ratings at the designated time period.

Consequently, data resulting from the first and third visits to the teaching centers were collected from 77 student teachers from Iowa State University and South Dakota State University. Of these 77 student teachers, data from 45 included observations during three visits to the cooperating schools. A breakdown by university indicated that 44 of the observations of student teachers were at ISU, 33 at SDSU.

Analysis of Data

The reliability between types of judges was determined by computing two analyses of variance (ANOV) for each of the 50 items in the observational device. The first ANOV was based upon the mean of the raw scores obtained for each of the 77 student teachers; the other analysis of variance was computed from the mean of the differences between the first and the third observations of each of the 77 student teachers.

The model upon which the analyses were based was (Winer, 1971, p. 365):

$$Y_{ijk} = \mu + C_i + T_{ij} + J_k + CJ_{ik} + \varepsilon_{ijk}$$

where μ= overall mean, C = teaching center, T = teacher within center (Error A), J = type of judge, CJ = teaching center by type of judge interaction, and ε = teachers within centers by types of judges (Error B).

The expected mean squares are designated in Table 1.

Table 1. Expected values of mean squares in the ANOV design.

Source of variation[a]	df	Expected values of mean squares
Centers (C)	43	$\sigma_\varepsilon^2 + 2\sigma_T^2 + 2\bar{t}\kappa_C^2$
Teachers within centers (T/C) (error)	33	$\sigma_\varepsilon^2 + 2\sigma_T^2$
Types of judges (J)	1	$\sigma_\varepsilon^2 + \bar{t}c\kappa_J^2$
Centers by types of judges (CxJ)	43	$\sigma_\varepsilon^2 + \bar{t}\kappa_{CJ}^2$
Teachers within centers by types of judges T/CxJ (error)	33	σ_ε^2

[a]Teachers are considered random effects while centers and types of judges are considered fixed.

The first analysis, based on the sum of scores from the visits, pertains to overall performance. The second analysis, the difference between the ratings given on the first and third time period, reflects change in performance. Thus, the reliability index from the first analysis reflects overall judged performance whereas the reliabilities derived from the second analysis reflect the reliability of a change score.

The items which were judged to have the most potential for reliable ratings for rating student teachers were ascertained by studying both analyses of variance. Specifically, F ratios for both type of judge and judge by center interaction sources of variance were inspected for nonsignificance or marginal significance. If both sources of variance were nonsignificant for both analyses,

it suggested not only no differences between ratings by types of judges but also that judges were not rating student teachers differently because of characteristics of a teaching center. These items were judged as potentially useful for reliably rating student teachers in a CBTE program.

The items found to have the most potential from the above analyses were further studied to determine their ability to discriminate between student teachers. Intraclass correlation coefficients were computed for both analyses using the formula (Winer, 1971, p. 286):

$$r_I = \frac{\sigma_T^2}{\sigma_\varepsilon^2 + \sigma_T^2}$$

where σ_T^2 = student teacher variance and σ_ε^2 = error variance. These coefficients were studied to ascertain if the item: 1) discriminated between student teachers, i.e., $r_I > .15$[1], 2) did not discriminate between student teachers, i.e., $r_I < .15$ on one analysis, or 3) appeared to discriminate more because of teaching center differences than because of perceived student teacher differences, i.e., $r_I < .15$ on both analyses.

Reliability estimates for each item for the average of two judges were calculated using the Spearman Brown procedure; r_I was used as the estimate of the correlation coefficient.

Items which were judged to be least reliably rated had significant F ratios for the J and CJ effects on at least one analysis of variance.

---

[1]This numerical value for r_I was judgementally selected based upon considerations of higher reliabilities when the item was combined with similar items to represent a broader competency.

A third group of items remained which did not follow either of the patterns previously described in that the items either had significant J effects, nonsignificant or marginal CJ effects, or a significant C effect.

Results and Discussion

Using the analysis of data method described, 18 items were identified as most promising; 22 items were identified at least promising; and 10 items fell into a category representing ambiguous results. Example of items in each category are presented in Table 2.

The most promising items

Of the 18 items identified as most promising, i.e., had no significant J or CJ effects, subgroup I is representative of items that clearly differentiated between student teachers as represented by the numerical value of the intraclass correlation coefficient. These items have the most potential for use in a CBTE program since it is desirable that before minimum performance levels can be set, the item needs to differentiate between the performance of student teachers.

For eight of the most promising items, one analysis suggested the difference was due to center differences and the other suggested the differences were due to student teacher differences. These items are illustrated in subgroup II. Therefore, these items were interpreted as not discriminating between teachers. If one is willing to impose the criterion that items only need to be reliably rated and do not need to discriminate between student teachers

Table 2. Illustrative examples of analysis of variance components of items in various classifications

Item	F ratios C[a]	F ratios J[b]	F ratios CJ[a]	r_I	r_j
I. Most promising items					
A. Subgroup I-items reliably judged; student teacher differences identified					
1. The teacher indicated the objectives of the lesson and their importance to students.	1.31[c]	.05	.80	.19	.29
	1.43	1.01	1.44	.17	.40
2. The teacher is well prepared for class.	2.04*	4.07	2.13*	.54	.70
	1.37	.11	.76	.24	.39
B. Subgroup II-items reliably judged; but item does not discriminate well between student teachers					
3. The teacher used meaningful examples or illustrations for conveying ideas during the lesson.	2.71**	.19	1.61	.02	.04
	2.30**	.52	.92	.29	.45
4. The teacher used questions to elicit thinking and student response consistent with instructional goals.	1.64	1.48	1.61	.33	.50
	1.72*	1.07	.99	.02	.04
C. Subgroup III-items reliably judged but items appear to be rated more on teacher center differences than student teacher differences					
5. The teacher emphasized reasons and relationships concerning the facts.	2.43*	.83	1.52	.03	.05
	1.35	1.26	1.17	.10	.10
6. The teacher followed through with her plans and yet remained flexible enough to adjust as needs became evident.	2.39**	.87	.91	.00	.00
	1.71	.38	1.00	.07	.13

II.	Least promising items					
7.	The teacher constructs well defined test items which reflect observance of the principles of item writing.	4.16** 2.65**	74.04** 1.97	2.28** .99	.28 .00	.43 .00
8.	The teacher encourages pupil's own evaluation of his/her work in both specific and informal ways.	4.22** 2.28**	29.22** 1.16	1.81* 1.58	.04 .36	.08 .53
III.	Other items					
9.	The teacher assisted the students in synthesizing, summarizing, and drawing conclusions.	.99 .97	17.41** 6.85*	.89 1.02	.27 .47	.43 .63
10.	The teacher provided an opportunity for the students to participate actively and/or to apply their learnings in different ways (verbal response, written work, etc.)	1.40 1.35	13.68** .96	1.38 .88	.49 .04	.66 .08
11.	The teacher encouraged the students to describe or show how the learning affects them personally.	3.03** 2.10*	8.65** 2.32	.84 .79	.04 .00	.07 .00
12.	The teacher tries to find things that students are "good at" instead of things they are "poor at."	4.32** 2.81**	45.64** 3.16	1.58 1.26	.14 .00	.24 .00

[a]Degrees of freedom for F are 43,33. Table values for F are 1.75 at 5 percent and 2.23 at 1 percent.

[b]Degrees of freedom for F are 1,33. Table values for F are 4.14 at 5 percent and 7.47 at 1 percent.

[c]The first line of data represents results from the analysis of variance "average" analysis; the second line, results from the "difference" analysis.

*Significant at $P<0.05$.

**Significant at $P<0.01$.

before a given preset criterion of performance is established in a CBTE program, these items have the potential for assessment of a teaching behavior at a designated mastery level.

The third group of items labeled subgroup III consisted of five items. While the items could be reliably rated, because $r_I < .15$ it appeared that perceived differences between student teachers are more attributable to differences between centers. Hence, even if these items can be adequately observed by types of judges, it appears that differences obtained are more attributable to center differences than student teacher differences. Because the source of variance producing differences between student teachers is largely beyond their control, these items appear to be least fair for use in rating a student teacher in a CBTE program.

The least promising items

The 22 least promising items for rating student teachers were those that had significant F ratios for the J and CJ effects on at least one analysis of variance. Since the significant J effect indicated that types of judges used the response pattern differently and the significant CJ effect suggested that judges ordered the centers differently, these items appeared to be the least optimum for observation of teaching behaviors.

Eleven of these items involved evaluation skills which had generated questions from the judges during the orientation sessions. This suggests that in future investigations more than one session for the judges is needed.

The remaining items

The 10 remaining items were those items which had significant J effects, nonsignificant or marginal CJ effects, and in some instances a significant C effect. The items fall into two major combinations which are presented in Table 2.

Three items had significant J effects with all other effects nonsignificant or marginal; two items (9 and 10) are illustrated. This pattern indicated that although judges are differing in their responses, one type of judge is consistently rating the student teacher higher than the other type of judge. Therefore, since the item appears to be more subjected to level differences on the response pattern by judge than different ordering of centers by judges, these items have potential usefulness for further investigation.

Four items showed significant differences for C and J effects on the average analysis and a significant J effect on the difference analysis; two items (11 and 12) are provided. All other effects were nonsignificant and the r_I was low for both analyses. These significant effects suggested that not only type of judge used the response pattern differently but also that center differences contributed to the variance. Therefore, these items need to be reworded or the implication of the items clarified to both types of judges if they are to be considered for further investigation.

No pattern could be discerned for the remaining three items in this group of 10.

Summary

To summarize, several items were identified which appear promising for reliably rating student teachers in CBTE programs. The results of the analyses used in making these judgments indicated the necessity to examine the patterns or combinations or significant and nonsignificant effects of the sources of variance in conjunction with the intraclass correlation coefficients in order to determine the item's usefulness for a CBTE program. This paper delineates one method by which additional observational items useful in a CBTE program could be identified.

Literature Cited

Gilbert, Ardyce. Changes, over time, in judged competencies of home economics student teachers. Unpublished doctoral dissertation, Iowa State University, 1974.

Hausafus, Cheryl. Evaluation competencies of student teachers. Unpublished term paper. Ames, Iowa: Home Economics Education, Iowa State University, 1973.

Kreutz, Shirley, and Anthony, Hazel. Home economics education objectives and generalizations related to selected concepts. A report on U.S. Office of Education Contract OEG 3-6-062205-1929, 1966.

Menne, John W. Teacher evaluation - performance of effectiveness? Unpublished paper distributed at a teacher evaluation conference held at Iowa State University, Ames, Iowa. Nov. 26-27, 1972. Mimeographed.

Rosner, Benjamin, and Kay, Patricia. Will the promise of C/PBTE be fulfilled? Phi Delta Kappan, 1974, 55, 290-296.

Thatcher, Sandra S. The development of a critique form for teaching for concept development. Unpublished master's thesis. Ohio State University, 1969.

Winer, R. J. Statistical principles in experimental design,. Second Edition. New York: McGraw Hill Book Company. 1971.