DOCUMENT RESUME

ED 240 103 SP 023 809

AUTHOR	Defino, Maria E.
TITLE	The Evaluation of Student Teachers.
INSTITUTION	Texas Univ., Austin. Research and Development Center for Teacher Education.
REPORT NO	RDCTE-9041
PUB DATE	Apr 83
NOTE	40p.
PUB TYPE	Information Analyses (070) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Behavior Rating Scales; Cooperating Teachers; *Evaluation Criteria; *Evaluation Methods; Higher Education; *Measurement Techniques; Personality Assessment; Preservice Teacher Education; Student Behavior; *Student Evaluation; *Student Teachers; Student Teacher Supervisors; Teacher Effectiveness; *Teacher Evaluation
IDENTIFIERS	*Student Teacher Evaluation

ABSTRACT

A review of pertinent literature on the evaluation of student teachers is presented. Descriptions are given of differing evaluation approaches and techniques, and several recent works discussing how to evaluate student teachers are appraised. One of these works focuses on what to avoid in teacher evaluation, while others describe different processes and perspectives (for example, horizontal evaluation or competency-based evaluation). Research on the scope and quality of variables employed for particular approaches to evaluating student teachers is examined. The validity of data obtained from these different evaluation methods is analyzed, with particular focus on the tendency of cooperating teachers and supervisors to inflate grades of students well known to them. An outline is presented of data and findings about student teacher evaluations, resulting from a multi-site, multi-method investigation of the student teacher experience. This examination of the process includes descriptions of the strategies used to analyze the various types of data collected. Implications of the conclusions for further research and improvement of student teacher evaluation, drawn from this study, are offered for consideration. (JD)

Research and Development Center for Teacher Education
The University of Texas at Austin
Austin, Texas 78712

THE EVALUATION OF STUDENT TEACHERS

Maria E. Defino

Report No. 9041

Research in Teacher Education Program
Gary A. Griffin, Program Director
April 1983

The Evaluation of Student Teachers

Preservice clinical teacher education, or student teaching, has been conceptualized as an experience embedded in several contexts at a multiplicity of levels, e.g., from that of individual differences among participating triads, up to that of interacting organizations in a broad social domain (Griffin, Hughes, Barnes, Carter, Defino, & Edwards, Note 1). Thus, the overall nature of the experience may be influenced by these multiple sources. One particular influence is that exerted by teacher education programs through required evaluations of student teachers (Defino, Barnes, & O'Neal, Note 2). In this manner, students often receive final clearance for receipt of their degrees and subsequent professional certification.

Purpose and Perspective

Given the apparent importance of final evaluations to student teaching at the personal and professional levels, this paper intends to serve three major purposes: first, a review of pertinent literature will be presented to establish the state of the art; second, findings about student teacher evaluations resulting from a multi-site, multimethod investigation of the experience will be outlined; and third, possible implications of the first two sections for the research and improvement of student teacher evaluation will be offered for consideration.

Before addressing each of these purposes in turn, however, it is necessary to inform the reader of a central premise: that is, teacher education institutions want and need to produce certified teachers who are also capable of promoting the academic achievement of their students, i.e., are "effective." Consequently, one would expect teacher education programs to have or to develop standards of evaluation which would enable them--as well as potential employers--to determine relatively objectively the degree to which student teachers behave in ways most likely to be effective.

Reflecting this premise, one section of this review of literature discusses some articles which describe possible evaluation techniques and methods for use with student teachers. The second section examines research of variable scope and quality which employs particular approaches of evaluating student teachers. It should be noted that, at present, little if any research exists to document which (if any) approach has more merit than any other.

Descriptions of Evaluation Approaches

Several recent works discussing how to evaluate student teachers summatively are available for review and discussion. One of these focuses primarily on what to avoid in teacher evaluation (Walker, 1981), while others describe processes to follow from different perspectives (for example, horizontal evaluation versus competency-based evaluation). Each of these will be reviewed briefly in turn.

Regarding what one should attempt to do when evaluating teachers and student teachers, both Goldhammer (1981) and Walker (1981) have offered scholarly works for practitioners. Goldhammer focuses upon the severe inadequacy of current evaluation techniques in teacher education (p. 27); he acknowledges that there is little hope for improvement until teacher educators apply extant research-based knowledge about teacher education to their tasks, and then subject the results of their efforts to objective scrutiny.

Conversely, Walker engages in a debunking of myths in teacher evaluation which he claims apply to the evaluation of student teachers as well. He decries the beliefs (among others) that 1) the same forms must be used to evaluate all teachers and 2) the focus of an evaluative observation is necessa. nd exclusively the teachers' behaviors. However, practical considerat for teacher education institutions must enter into the evaluation . neir students. For instance, it may very well be a legal necessity to use equivalent forms in the final evaluation of student teachers within a given teacher education program (from the point of view of equal protection).

Several tools for evaluating student teachers have been described in the literature, and are a reflection of particular assumptions or philosophies. Thus, they may be roughly categorized into two groups, those which are "vertical" in nature (skill-based and ranking individual achievement relative to that of the group) and those which are "horizontal" in nature (reflecting intraindividual development). Within the latter group, one article by Gitlin (1981) describes the steps one ought to follow in employing his version of horizontal eval ation both formatively and summatively with student teachers. Benefits of following the model (which is essentially one of clinical supervision) include an information yield over the course of a semester that supervisors should rely upon in the summative evaluation: the clarity and scope of goals/intents generated by student teachers; the student teachers' degree of success in translating their goals/intents into practice; and the quality of self-diagnosis acquired by the student teachers.

Within the group of vertical competency-based (CBTE) evaluation methods, Johnston and Hodge (1981) describe an approach which claims to further the student teachers' self-evaluation and professional development, termed "Self-Evaluation through Performance Statements" (SEPS). Not only do the student teachers actively participate in their own evaluation, they collect data used to make evaluative judgments. These are in the form of written performance statements (brief, low-inference statements without adverbs) about their activities in the classroom. The larger the performance statement data base, the stronger the conclusions one may draw about the quality of a given student teacher's work. Also, performance statements may be organized according to extant formal evaluation criteria. Therefore, supervisors are justified in requesting them, and are likely to obtain pertinent information for summative evaluations.

Two other means of assessing student teacher competencies are the Teacher Performance Assessment Instrument (TPAI) developed through Georgia (see Reiff, 1980) and the Classroom Observation Keyed for Research instrument (COKER; see Dickson, Note 3). The former set of instruments provides the observer-evaluator with a list of competency indicators and corresponding sets of descriptors; his/her task is to judge how well a given teacher's performance "fits" the competency described (e.g., none of the teacher's behaviors are recorded per se). The latter instrument, the COKER, requires the simple recording of specified behaviors which the observer has seen the teacher demonstrate. In this sense, it probably represents a lower-inference schema than the TPAI. What separates both the TPAI and the COKER from other available evaluation systems is the fact that they have been utilized in various research efforts in teacher and/or student teacher evaluation, and have therefore demonstrated some level of generalizabilty. This is in contrast to other evaluation forms and procedures used in the restricted contexts of given teacher education programs here and abroad.

The reader is reminded that the systems just described are purely a sampling of all those available. They were selected for discussion from the point of view of revealing variety and choice, rather than any endorsement of validity. In the next section of this text, research-based information about the quality of various evaluation/observation systems will be reviewed and discussed.

Research Using Particular Evaluation Strategies

The literature reporting findings pertaining to the evaluation of student teachers appears scattered, perhaps reflecting the variety of assessment forms and techniques employed in teacher education programs. One way to organize the literature which seemed functional was the following: first, to consider research pertaining to the assessments themselves (e.g., reliability, item discrimination, cut-off scores, etc.); and second, to consider evidence on the particular issue of grade inflation. Each of these will be discussed in turn.

Instrument properties. Several articles describe instruments which rely upon listings of competencies as the basis for evaluating the performance of student teachers. Reiff (1980) and Mitsakos (1979) are two prime examples of research in this category. The former author utilized items from the TPAI, whereas the latter derived ten teacher performance "standards" from the fairly substantial body of research on teacher effectiveness (each subsuming numerous performance/competency statements). Reiff's thrust involved the establishment of minimum rating scores for the TPAI items as well as their reliability across different raters; Mitsakos' concern was the establishment of measurement properties of the ten standards presented as items upon a rating scale (e.g., interrater agreement; discriminatory power of the items; etc.).

Both authors report favorable results. Mitsakos obtained high reliability coefficients (.87-.97) and adequate item-total correlation coefficients (.610-.931) for the ten research-derived standards. Also, Reiff was able to establish certain minimum scores for the competencies listed on several TPAI scales. However, the latter author observed that in all data analyses, cooperating teachers and classroom pupils consistently rated student teachers higher than did a team of trained data collectors.

Several questions for future evaluation research emerge from these findings. For example, are most supervisors of student teachers trained in data collection methods appropriate to the formal evaluations utilized by their teacher education programs? How often are student teacher evaluations dependent upon a single rating source? (Howey, Yarger, & Joyce, 1978, indicate that in formally stated procedures this is rare.) Can one assume comparable quality of performance in student teachers who have received equally high (quantitative) final evaluations from institutions using qualitatively different evaluation systems?

Two articles have directly addressed the problem of dependency of results upon evaluation method (Dickson, Note 3, and Irvine, Note 4). Dickson contrasted findings of two observational systems used to assess student teaching experiences in a CBTE program. The first system was the TPAI, where observers were required to judge how well a given student teacher's performance "fit" the indicators listed. Second was the lower-inference COKER system, whereby instances of specific student teacher and classroom pupil behaviors were observed and recorded. Overlap occurred on 18 of the competencies either described or targeted for observation across the TPAI and COKER. Dickson observed that for these 18 competencies, there was no significant correlation between the scores obtained by student teachers on the two different instruments. This reinforces the concerns expressed above regarding the validity and interpretation of student teacher final evaluations.

The second study employing the TPAI was also one which contrasted method, specifically, rating source: student teacher self-ratings with university supervisor ratings. Irvine (Note 4) reported findings drawn from the implementation of an Integrated Model for Training and Supervision (IMTS) of student teachers. They received instruction on the meaning of TPAI competency statements, their corresponding descriptors, and their relationship to teaching effectiveness research. Unlike Reiff's (1980) findings, Irvine reported that highly significant correlations ($p = <.01$ or $<.001$ for all variables) were obtained between preservice teachers' self-ratings and supervisor ratings of their (student teachers') performance. However, the magnitude of the shared variances ranged from 9% to 67%, with the majority of figures in the 19%-29% range. Therefore, even with the academic intervention, there generally remained more combined unique and error variance than shared variance across the ratings. The implication is that student teachers and supervisors in this study may each have been responding to distinct considerations when assigning ratings (assuming that specific variance may have been greater than error variance).

Other researchers have examined relationships between student teachers' self-ratings and the performance ratings assigned them by different role groups, as well (e.g., Grafton, Walters, & Magitti, Note 5; Hoffman & Gellen, 1981; McEwing, 1981). Briefly, Hoffman and Gellen (1981) compared ratings made by teachers of their pre-student teaching aides with the aides' self-ratings. Certain patterns emerged, in which overall means appeared similar across raters but differences within groups existed. For example, female elementary aides rated themselves significantly lower than did their classroom teachers; male elementary aides' self-ratings were not significantly different from their teachers' ratings. Also, classroom teachers rated female elementary level aides higher than male elementary level aides. Further research and greater mutual understanding of the objectives of the evaluations were recommended.

McEwing (1981) arrived at a similar recommendation as a function of his investigation of the perceived importance and perceived mastery of 27 identified teaching skill areas. Six groups responded to the items, ranging in experience from students entering the teacher education sequence to supervisors of graduates in teacher education. Among the experienced student teaching respondents, McEwing observed that his sample consistently reported their mastery at levels exceeding the self-reports of students entering the teacher education program. However, the student teachers also tended to rate their own skill mastery lower than their university supervisors had rated them. Unfortunately no statistical analyses were run to index the significance of this mismatch in perceptions of mastery.

Grafton, Walters, and Magitti (Note 5) investigated interjudge reliability on competency ratings of 111 student teachers; in addition to the descriptive study, an intervention designed to enhance the assessment skills of a small group of cooperating teachers was implemented. Briefly, their findings indicated that: ratings of student teachers' competence generally increased over time, regardless of who was making the judgment; all grand means (times 1, 2, and 3) were above the midpoint of the rating scale; significant levels of agreement were obtained across judges on nearly all items; and the training intervention did not serve to substantially enhance interjudge reliability within the targeted group of teachers.

In sum, then, it would seem that adequate reliabilities can be obtained with various rating scales of student teaching performance. Validity data are less abundant than reliability data. In particular, "mismatches" in perceived levels of student teaching mastery have been reported. Evidence regarding the utility of interventions designed to augment assessment skills of those assigning performance/evaluation ratings is minimal, also.

One attempt at behavioral validation of student teacher evaluations has been reported by Denton and Kazimi (Note 6). Their work paralleled the reasoning in teacher effectiveness research. That is, an attempt was made to relate classroom students' achievement to the teaching performance of student teachers, as assessed by university supervisors on final evaluation forms. The supervisors utilized forms with 20 five-point items related to instructional skills and eight items related to personal competencies. Ratings actually assigned on the final evaluations reflected the consensus achieved among student teaching triads (supervisors, cooperating teachers, and student teachers) through final conferences. The achievement of classroom pupils was indexed through the percentage of objectives they attained for a particular unit taught and evaluated by the student teachers. Denton and Kazimi's description of the research did not make it clear to this writer whether or not all student teachers were trying to compare their pupils' performance against a uniform set of goal and objective statements, yet this would seem to be a significant piece of information when trying to interpret the data.

The zero-order correlation coefficients obtained between student teachers' final evaluation ratings (item by item) with the attainment of objectives by their pupils were generally low in magnitude. Only one of the 28 items obtained significance at the .05 level ($r = .19$); it assessed the student teachers' ability to develop lesson plans. Three other correlations were of marginal statistical significance, one of these being in a negative direction: use of different levels of classroom questions ($r = .16$, $p = .08$); overall rating of performance while teaching two-week units (r = -.18, p = .06); and personal energy (r = .15, p = .10). Note that, except for the item with the negative correlation, these seem to parallel classroom teacher behaviors which have already been shown to be "effective" or associated with pupil achievement (see, e.g., Barnes, Note 7). However, Denton and Kazimi (Note 6) conclude that student teachers' facility on final evaluation scale items is not necessarily associated with their pupils' academic achievement. Further research, which draws upon sources external to student teaching triad members for indices of mastery and/or effectiveness, seems desirable in spite of logistic difficulties (see, for example, the framework outlined by Schalock, 1979). One major advantage would be the avoidance of possible positive response bias on the part of university supervisors and cooperating teachers. Articles which address this problem of "grade inflation" and/or positive response bias will be discussed next.

Grade inflation. As noted by Chiarelott, Davidman, and Muse (1980, p. 297), grade inflation among student teachers can present problems for administrators who need to select the best qualified individuals for employment in their districts. Chiarelott et al. therefore attempted to develop a new evaluation scale to be judged by administrators for its utility. Among the changes in the pilot which administrators generally requested were to keep the scale short, and to provide space for teacher comments about the ratings given as well as about the types of classrooms in which student teachers had been placed.

Funk, Hoffman, Keithley, and Long (1982) also endeavored to develop a new, more informative evaluation scale for the teacher education program at their university. Yet they encountered the same problem: "...Raters tend to overrate those student teachers with whom they are familiar. When a scale employs only four descriptive choices...the room for this error is even greater." (p. 319) They concluded that ratings of "good" (the third choice on their scale) had to be considered reflective of only "average" preparation. Program areas responsible for preparation where student teachers had been rated "good" by the cooperating teachers, therefore, were thought to be areas which might need improvement.

Two groups of researchers have attempted to address the question of grade inflation directly (Blackmon, Andrews, & Mackey, Note 8, and Haviland & Haviland, 1981). Blackmon et al. studied ratings obtained by 442 student teachers over a two-year period on nine five-point items taken from their university's evaluation form. These items were selected because research literature generally indicated their importance to teaching. Blackmon et al.'s findings included the following: 1) the grand means for all nine items were consistently above the conventional Likert mean of 3.00; 2) student teachers with low cumulative grade point averages (GPAs), as a group, also received the lowest mean ratings across the board--yet on only two variables did their means equal or fall lower than 4.00 (classroom control, 3.87; teaching skills, 4.00); 3) all student teachers, regardless of sex, cumulative GPA, or level (elementary/secondary) received their lowest ratings on classroom control and teaching skills; 4) elementary level student teachers, on the average, received higher ratings across all items than secondary level student teachers; 5) all items intercorrelated to the .0001 level; and 6) one factor accounted for 62% of the variation in the correlation matrix for the evaluation items, which they termed an "overall evaluation" factor. That is, supervisors in this study were thought to be assigning ratings consistent with both the student teachers' cumulative achievement and their own global or overall judgment of them.

Haviland and Haviland (1981) utilized a different approach in attempting to determine whether or not grade inflation was significant at their university. They chose to examine changes in the distribution of mean ratings given to student teachers by their cooperating teachers and university supervisors over the last 1½ decades. Chi-square analyses on cooperating teacher and supervisor ratings of student teachers for the sampled years in that time span (1964, 1969, 1974, and 1979) revealed that the supervisors were largely consistent over time in their ratings. However, the cooperating teachers' ratings of student teachers had increased on 11 of the 15 evaluation form items. Haviland et al. speculate that any of three reasons may have been responsible for the change: improved pre-student-teaching field experiences for the more contemporary students; increased humanism and concern for accentuating the positive among teachers-in-service; and/or, concerns over the increased availability of (and consequent liability for) evaluation forms and personnel files. They recommend that practitioners become increasingly aware of the grade inflation problem, and periodically review evaluation practices at their teacher education institutions as a check against it.

In conclusion, two research efforts located by the present writer that focus directly on grade inflation indicate tentatively that it is a problem which has increased over time, and that it is associated with global or general positive views of the student teachers. Inconsistencies in the perceptions of student teachers' performance across various role groups (e.g., self, supervisor, and so on) seem to indicate that one's professional vantage point while assigning ratings may be of significance, also.

Summary

While the variety of possible evaluation approaches ranges from individualized (Newport, 1982) to state-mandated competency checklists such as the TPAI, few articles in this review explicitly address the linkage between the process of student teacher final evaluations and broader contextual and professional issues, such as certification (see Schalock's framework, 1979, p. 401). Articles originating from teacher preparation programs often seem to take an insulated or self-contained view of possible goals for student teachers, and/or the content of the evaluations. Thus, goals, competencies, and/or items considered appropriate for final evaluation forms appear to be infrequently linked to extant research-based knowledge regarding generally desirable teaching practices (as opposed to particular skills or desired end-points established by each program through craft-oriented knowledge).

It was also observed that reliability data for the various quantitative measures were more readily located in the literature than validity data. One problem with student teacher evaluations, reported by more than one source, was a discrepancy in ratings dependent upon role group of the raters. In an effort to update this picture of the state of the art of student teacher evaluation, data from the Clinical Teacher Education-Preservice study (Griffin, Barnes, Hughes, O'Neal, Defino, Edwards, & Hukill, Note 9) was analyzed to provide a comprehensive description of student teacher evaluation in two contrasting institutional contexts. The methods and results of the inquiry are described next.

Methods

Subjects

A total of 88 cooperating teachers, 93 student teachers, and their 17 university supervisors from two sites participated in the CTE-P study (N = 198). Ten triads of student teachers-cooperating teachers-university supervisors at each site participated more extensively than the remainder, and were termed the "intensive sample." The other triads comprised the "general sample."

Data Collection

Both qualitative and quantitative data were collected from the participants over the course of the semester. All 88 cooperating teachers and 17 university supervisors agreed to share with the investigators copies of completed final evaluation forms on their student teachers. In addition, each triad member completed a Performance Rating (Hughes & Hukill, Note 10) of the student teacher at the end of the semester (thus, student teachers rated their own performance over the course of the semester, while their cooperating teachers and university supervisors rated them on parallel forms). Also, nine intensive sample university supervisors and their 20 cooperating teachers were interviewed near the end of the student teaching semester by members of the Research in Teacher Education (RITE) staff. In this manner, qualitative data pertaining to the perceived strengths and weaknesses of their student teachers were obtained. Last, three of the intensive sample triads were able to audiotape their three-way final evaluation conferences. These tapes were shared with the investigators.

Instrumentation

The evaluation forms at each site (referred to as State University and Metropolitan University) consisted primarily of five-point Likert-type rating items (24 items in one case, 11 in the other). Both forms provided for or requested prose comments from the persons completing the form (either cooperating teachers or university supervisors). In order to facilitate discussion of the results across the two distinct forms, items on each were grouped by the RITE staff to create two approximately parallel subscales. Thus, five items on the State University form were labeled the "Teaching

Competency" subscale, while 11 items on the Metropolitan University form served as the parallel. Some examples are: "Demonstrated skillful implementation of learning plans," "Presents lessons clearly and effectively," or "Demonstrated skillful choices of instructional methods based on children's needs and interests." Ten other items on the Metropolitan University form were labeled the "Professional Competency" subscale, as were four apparently parallel items on the State University evaluation form. Examples of the items included here are "Demonstrated ability to profit from feedback," "Attends to schedules and commitments," or "Handles situations with poise, self-control." The remaining three items on the Metropolitan form pertained to the student teachers' "Personal Characteristics," and the one remaining item on the State University form required an overall judgment of the student teacher.

The Performance Rating scale was developed by members of the RITE staff as an independent means of assessing student teachers' performance. It consisted of 29 behaviorally oriented five-point Likert-type items. The language on the three parallel versions of the instrument (self, cooperating teacher, university supervisor) was identical except for the use of the first person on the self-rating form. Roughly half of the items were stated negatively in an attempt to avoid creation of a response set on the part of the raters.

Data Analysis

Several strategies were utilized to analyze the various types of data collected. Each of these will be described in turn.

Final evaluation data. A variety of descriptive statistics was calculated on the mean evaluation ratings (by subscale) given the student teachers by their supervisors and cooperating teachers. Grand means, standard deviations, and indices of kurtosis and skewness for the final evaluation data were calculated and are displayed in Table 1. Table 2 shows the same descriptive data broken down according to site, role group making the evaluations (referred to as "evaluator type"), and subscale. As is evident from the tables, the substantial indices of skewness and kurtosis obtained (while not tested for significance as per the procedures in Snedecor & Cochran, 1967) point out the possible lack of validity in the use of standard inferential statistics with these data (due to possible violation of the assumption of normality). However, hierarchical ANOVAs for site (entered first) and evaluator type were calculated; results are presented in Tables 3 and 4.

Performance Rating data. Descriptive data for the Performance Ratings were calculated but have been reported elsewhere (Griffin, et al., Note 9). Of greatest interest for the present purposes was the calculation of correlation coefficients between the Performance Ratings (self, supervisor, and cooperating teacher) and the final evaluation ratings given to student teachers. The resulting correlation matrix is displayed in Table 5.

Interview data. Responses of intensive sample university supervisors and cooperating teachers to the interview questions, "What are the strengths and weaknesses of your best student teacher? Of your weakest student teacher?" or "What are the strengths and weaknesses of your present student teacher?," respectively, were analyzed for themes or unifying, underlying constructs. This was done in the following manner: lists of the descriptors of strengths and weaknesses generated by interviewees were compiled and typed onto individual slips of paper (e.g., one key phrase per paper). These were sorted out twice; in the first sort, exact duplicates were placed together; in the

Table 1

Descriptive Statistics for Student Teachers'

Mean Ratings on Their Final Evaluations

Subscale	Grand Mean	Standard Deviation	Kurtosis	Skewness
Teaching Competency	4.251	.901	7.137	-2.276
Professional Competency	4.448	.840	12.396	-3.107

All evaluations were made on five-point Likert-type rating scales.

Table 2

Descriptive Statistics by Site, Evaluator Type, and Subscale for Student Teachers' Mean Final Evaluation Scores[a]

e	Evaluator Type	Subscale	Mean	Standard Deviation	Kurtosis	Skewness
ropolitan versity	Supervisor	Teaching Competency	4.521	.748	5.977	-2.229
		Professional Competency	4.612	.639	5.991	-2.354
	Cooperating Teacher	Teaching Competency	4.542	.589	.606	-1.266
		Professional Competency	4.615	.537	1.485	-1.447
te versity	Supervisor	Teaching Competency	4.021	.698	1.961	-1.131
		Professional Competency	4.413	.645	10.499	-2.641
	Cooperating Teacher	Teaching Competency	4.165	.723	1.090	-1.335
		Professional Competency	4.429	.642	2.878	-1.662

1 evaluations were made on five-point Likert-type rating scales.

Table 3

Summary of ANOVA of Student Teacher Ratings on the Teaching Competency Subscale by their Cooperating Teachers and University Supervisors at Two Sites

Source	SS	df	MS	F
Site	8.759	1	8.759	18.317**
Evaluator Type	.290	1	.290	.606
Site X Evaluator Type	.173	1	.361	.549
Residual	85.120	178	.478	
Total	94.341	181	.521	

** p < .001

Table 4

Summary of ANOVA of Student Teacher Ratings on the Professional Competency Subscale by their Cooperating Teachers and University Supervisors at Two Sites

Source	SS	df	MS	F
Site	1.685	1	1.685	4.435*
Evaluator Type	.003	1	.003	.005
Site X Evaluator Type	.002	1	.002	.005
Residual	67.651	178	.380	
Total	69.342	181	.383	

* p < .05

Table 5

Correlations between Student Teaching Performance Ratings by Self, Supervisor, and Cooperating Teacher, with Supervisor or Cooperating Teacher Mean Final Evaluation Ratings of Student Teachers

	University Supervisors		Cooperating Teachers	
	Teaching Competency Subscale	Professional Competency Subscale	Teaching Competency Subscale	Professional Competency Subscale
Self-Ratings of Performance	.3662** (N = 83)	.3613** (N = 83)	.4389** (N = 82)	.4330** (N = 82)
Supervisors' Ratings of ST Performance	.4920** (N = 32)	.5718** (N = 32)	.4864** (N = 31)	.4266** (N = 31)
Cooperating Teacher Ratings of ST Performance	.5580** (N = 76)	.6666** (N = 76)	.7432** (N = 76)	.7365** (N = 76)

** _p_ < .01

second, semantically similar descriptors were sorted, with groups of exact duplicates treated as a single response. At least three RITE staff members participated in this sorting process.

Final conference data. The three final conferences were analyzed through an adaptation of Weller's (1971) MOSAICS (see O'Neal, Note 11 for a comprehensive discussion of this procedure). Information pertaining to these three research questions was obtained: 1) who did most of the talking in the conference; 2) who was the recipient of most comments; and, 3) what were the topics discussed. The reader is referred to Appendix A for a sample grid used to tally conference statements.

Results

Final Evaluation Ratings

As displayed in Tables 1 and 2, several findings pertaining to student teachers' mean evaluation ratings upon the Teaching Competency and Professional Competency Subscales are apparent. First, the values for the standard deviations are relatively low while the means (both grand means and cell means) are high for having been made on five-point Likert-type scales, where the conventional mean approximates 3.00. To concretely illustrate the skewness and kurtosis indicated by the numerals in Tables 1 and 2, Figure 1 charts the percentages of student teachers at either site whose lowest individual item rating on the final evaluation form had a value of 1, 2, 3, 4, or 5, exclusively. Thus it is clear that the great majority of student teachers in this study received item ratings of three or above on their final evaluations.

The hierarchical ANOVAs yielded statistically significant differences for the site variable on both the Teaching ($p < .001$) and Professional Competency

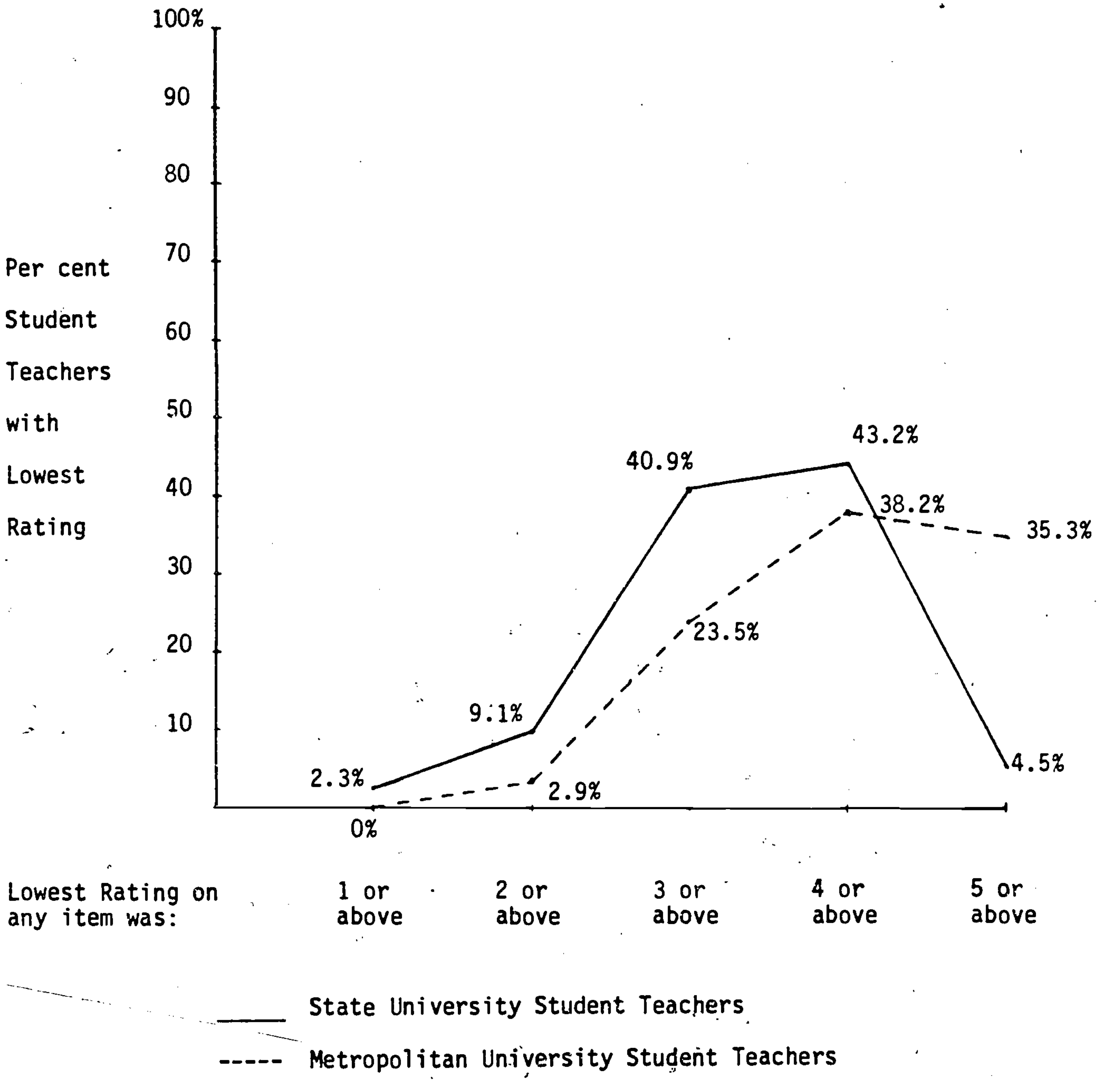

Figure 1. Graph of Percentages of Student Teachers whose Evaluation Item Ratings had Lower Limits of 1, 2, 3, 4, or 5, respectively.

($p < .05$) subscales of the final evaluation ratings (see Tables 3 and 4). The effect was more pronounced on the Teaching Competency subscale.

Performance Ratings

As displayed in Table 5, the correlations between the final evaluation ratings assigned student teachers by their cooperating teachers and university supervisors, and the performance ratings of student teachers made by themselves, their supervisors, and their cooperating teachers, were all highly significant ($p < .01$ in all cases). These results are interesting in view of the high means and low standard deviations characterizing both the Performance Ratings (see Table 6) and the evaluation ratings (Tables 1 and 2).

Interview Data

When supervisors and cooperating teachers in the intensive sample were asked to describe the strengths and weaknesses of their student teachers (best and weakest, or present ones, respectively), responses were prompt and tended to contain references to inferred characteristics. Ninety-four responses to the questions about weaknesses were studied for themes, while 111 responses about strengths were considered. Identified themes will be described in descending order of prominence.

Strengths. A major category among the student teachers' assets, reported by cooperating teachers and supervisors, involved an inferred characteristic such as social sensitivity or being child-oriented. That is, the student teachers were thought to have "wonderful rapport" with the classroom pupils, were able to "communicate well with the kids," and were "sensitive to individual needs" and "concerned about students."

A second common theme to the cooperating teachers' and supervisors' responses included various inferred intellectual characteristics. Student teachers were most admired for their "flexibility," "creativity," and

Table 6

Means, Standard Deviations, and Ranges of

Performance Rating of Student Teachers

Type of Rater	Mean	Standard Deviation	Range	N[a]
Self	4.36	.40	3.10 - 5.00	86
CT	3.99	.68	1.92 - 4.76	81
US	4.12	.64	2.41 - 4.73	33

[a] Discrepancies in N reflect variation in N of supervisors and/or missing data.

"openness" or "objectivity." Also, several remarks seemed to revolve around a sense of readiness or security in assuming the teaching role: "she's looked to as the other teacher," "has a professional air," or "her manner is laid-back, and it's nice, she's at a different pace than I am."

The next most common theme also reflected an inferred personal characteristic of the student teachers, and might be termed their motivation. Phrases subsumed by this theme included "wanting to be a good teacher," "dedication," "enthusiasm," "hard worker," and the like.

The next identifiable theme among the student teacher strengths described by cooperating teachers and supervisors dealt with various teaching skills. For instance, student teachers' ability to "follow through" with plans, or to be "good in timing," "questioning them (the classroom pupils)," and to "design interesting methods" were all viewed as strengths in the classroom setting.

Thus, the four dominant themes among the perceived strengths of participating student teachers were: (1) social sensitivity or being child-oriented; (2) intellectual flexibility and/or security in the student teaching role; (3) positive motivation; and, (4) particular teaching skills.

Weaknesses. Several themes were discerned among the perceived weaknesses of student teachers, also. Two equally common but very discrete themes involved either a lack of confidence or a sort of intellectual rigidity. Examples of the former include "being over-cautious," "being unsure," "insecurity," and "she's not totally comfortable in that class yet." Among phrases supporting the existence of a theme of intellectual rigidity or narrowness were these: "he didn't think there was anything wrong," "her inability to see what's too much noise (sic)," "judgment," "inability to see what's happening around her," or "lacks ability to diagnose effectively."

Management of instruction and/or the pupils' classroom behavior was another commonly mentioned perceived weak area among student teachers. Among the evidence for this theme were comments such as these: "monitoring noise level and keeping kids on-task," "management of transitioning," "not communicating...procedures to the students," "still uses the chalkboard continually," "control," or "she had a hard time keeping up with the planning and the paperwork."

Various forms of communication problems were cited as weaknesses, also. At the most general level, one cooperating teacher spoke of a broad failure in "communicating with the students." Other participants spoke of "dealing with this age child, the vocabulary, the concepts at hand;" "being able to relate on an equal level to parents, a lot of parents complained...;" "her interactions with the children;" and "her voice level, that intangible sameness."

In sum, cooperating teachers and university supervisors at both sites were able to specify what they felt were weaknesses among their student teachers. Among the more readily identifiable constructs or themes were these: (1) lack of confidence; (2) intellectual rigidity; (3) management of teaching tasks such as instruction, classroom behavior, and planning; and, (4) communication problems.

Conferences

While these results may be described, it is essential for the reader to bear in mind the extremely constricted sampling from which they were drawn: only three of the 20 intensive sample student teaching triads audiotaped their final evaluation conferences. The data can in no way be considered representative.

Use of the coding system with these three conferences indicated the following bits of information. In two of the three conferences, supervisors dominated the conversation while student teachers spoke the least. Identified recipients of most of the statements varied with each conference. In all three cases, cooperating teachers made mostly evaluative statements, as did supervisors; however, the latter offered a greater variety of statement types. In two of the three conferences, discussion centered around teaching events (particularly social or disciplinary interactions) which had occurred in the classroom. The "objectives and content" category as a focus of discussion was never used in the conference coding.

Discussion

The results from descriptive data on the student teachers' final evaluation forms are relatively straightforward: at both sites and on both subscales, the means were high for five-point scales; the standard deviations were fairly small; and the skewed nature of the distribution is apparent. Several explanations for these results may be considered. For example, it may be that student teachers in these two sites were relatively superior in their clinical experience, and the results are a reflection of this sampling. It could also be that teachers and supervisors in the present study were assigning evaluation ratings on the basis of poorly differentiated, generally positive views of the student teachers, as Blackmon et al.'s (Note 8) participants were. The fact that the Performance Ratings showed parallel characteristics--high means, low standard deviations--yet were highly correlated with the evaluation ratings tells us that the respondents were fairly stable or consistent in the way they rated the student teachers, but does not clarify the reasons for this convergence. However, other information about the student teachers in the CTE-P study, such as their low mean scores on standardized measures like the Quick Word Test (about the 15th percentile; see Hughes & Hukill, Note 10), tend to indicate that the ratings may have reflected at least a degree of response bias.

Although a statistically significant result was obtained through the hierarchical ANOVA for site and sample conducted on the student teachers' final evaluations, with major differences occurring for the site variable, very little can be said about the meaning of this result. For example, each institution utilized a different five-point rating scale on its evaluation forms. In addition, one institution had the 24 items on its evaluation form already grouped into the Teaching Competency and Professional Competency subscales; 10 of the 11 items on the other form were divided into two "parallel" subscales by the RITE staff--thus, neither number of items nor wording of items were equivalent across sites. Therefore, it is impossible to determine whether the obtained significant difference across sites was a function of the instruments employed; the persons at either site; the programs at either site; and so on. The significant difference across sites, then, is difficult to interpret.

The similarity in the interview responses across sites further substantiates the general portrayal of similarity in the student teaching participants at either location. Evidence for the existence of a global, poorly differentiated view of the student teachers on the part of their supervisors and cooperating teachers is also present. Note that, among the student teacher strengths perceived and reported by these teacher educators, the three most prominent themes pertained to inferred intellectual or personality characteristics. Three of the four most prominent themes pertaining to student teacher weaknesses were similarly inferred characteristics. Also, as a unique cross-check to the content validity of the themes, it was observed that the independently derived themes of strengths and of weaknesses represent near opposites: lack of confidence versus security in the role; flexibility versus rigidity; social sensitivity and rapport versus communication problems; and so on.

These results should not be taken to mean that supervisors and cooperating teachers cannot describe behaviorally the strengths and weaknesses of student teachers. At least one theme in each of these domains did pertain to observable tasks of teaching, such as management of instruction. However, it seems reasonable to speculate that specific, behaviorally-defined strengths and weaknesses are lower in the supervisors' and teachers' response hierarchies than the global inferential ones described above. If this is plausible, then it would seem likely that more specific kinds of information will not be revealed unless the person seeking it--researcher, administrator, or whomever--asks for it directly through specific questions and/or highly structured evaluation items. When one stops to consider, for example, that the entire student teaching semester at one of the study sites is to be summatively judged on 11 rating items, it would seem that one is being asked to respond at a relatively general level. Hence, it seems logical for cooperating teachers and supervisors to respond in a global, poorly differentiated manner, as per the data obtained.

All of this points to a single, core question which bears on the nature of student teaching evaluation forms and the kinds of information which they elicit: what purpose or purposes, both formal and informal, are to be served through final evaluations of student teachers? Are they intended to screen student teachers for entry into the profession, and if so, at what organizational level? Are they intended primarily to provide feedback to the student teacher about broad skill areas which might need strengthening? Or, are they a form of protection for the teacher education program, in the sense that they furnish evidence that the student teacher was doing what he/she was supposed to be doing (however defined) during the clinical experience? Any or all of these purposes, among others, might be legitimately served through final evaluation ratings. However, one cannot help but suspect that clarification of formal and informal evaluation purposes, together with clarity of teacher education programs goals/intents, might serve to alleviate problems such as so-called "grade inflation" (through problem redefinition, if nothing else). Closer linkage of the content of evaluation forms to objectively, generally desirable teacher behaviors--as identified through research--should also be beneficial (as per Goldhammer's 1980 recommendation).

One last set of data, the conference data, remain to be discussed. In view of the highly restricted, nonrepresentative sample obtained, it does not seem reasonable to try to reach conclusions about the content, process, or products of three-way final evaluation conferences. Though firm conclusions cannot be drawn from the final evaluation conference data, a potentially useful coding system was developed in the process of analyzing the conferences (O'Neal, Note 11). This system would seem to open the way for studying numerous research questions about the nature and results of such conferences within the broad context of supervision, and is worthy of further attention.

Conclusions

In conclusion, if it can be said that one hallmark of quality descriptive research is that more questions are raised than are answered, it would seem that the description of student teacher evaluation obtained through the Clinical Teacher Education-Preservice study is a rich one. First, a substantial amount of evidence indicating a tendency among participants in the clinical experience to view student teachers in a general, favorable light was obtained; it can be added to earlier literature in this vein. What makes this addition so valuable is that it is one of the first which collected both qualitative and quantitative data simultaneously from more than one location at the end of the clinical semester, resulting in greater generalizability of its findings. Second, it has brought to attention a new method for use in analyzing final conference data which is not bound to the goals of a particular program or site, but, rather, is bound to theory. Of course, the ultimate value of the coding system can only be judged with its use over time and across locations. Finally, the data have given rise to some major questions about the intended purposes and therefore the content of student teacher final evaluations. Clearly a shared future effort by researchers and practitioners in response to these questions is needed.

Reference Notes

1. Griffin, G.A., Hughes, G.R., Jr., Barnes, S., Carter, H., Defino, M.E., & Edwards, S. The student teaching experience (Research proposal submitted to National Institute of Education). Austin, TX: The University of Texas at Austin, The Research and Development Center for Teacher Education, June 1981.

2. Defino, M., Barnes, S., & O'Neal, S. The context of clinical preservice teacher education: The student teaching experience (Report No. 9022). Austin, TX: The University of Texas at Austin, The Research and Development Center for Teacher Education, September 1982.

3. Dickson, G.E. Basic education: Fundamentally a concern for competent, effective teachers. Paper presented at the World Assembly of the International Council on Education for Teaching, Cairo, Egypt, August 1981.

4. Irvine, J.J. The effects of the Integrated Model for the Training and Supervision of teachers on the self-assessment skills of preservice teachers. Paper presented at the annual meeting of the American Educational Research Association, New York, March 1982.

5. Grafton, J.H., Walters, S.A., & Magitti, P.J. Teaching competency assessment: An interjudge reliability study for student teachers in Early Childhood and Elementary Education at West Chester State College. Paper presented at the annual National Conference of the Association of Teacher Educators, Dallas, TX, February 1981.

6. Denton, J.J., & Kazimi, E. *Relations among final supervisor skill ratings of student teachers and cognitive attainment values of learners taught by student teachers*. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX, February 1982.

7. Barnes, S. *Synthesis of selected research on teaching findings* (Report No. 9009). Austin, TX: The University of Texas at Austin, The Research and Development Center for Teacher Education, September 1981.

8. Blackmon, C.R., Andrews, J.W., & Mackey, J.A. *Evaluation of student teachers: Ratings by supervising teachers on nine performance variables*. Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, LA, November 1978.

9. Griffin, G.A., Barnes, S., Hughes, R., Jr., O'Neal, S., Defino, M.E., Edwards, S.A., & Hukill, H. *Clinical preservice teacher education: Final report of a descriptive study* (Report No. 9025). Austin, TX: The University of Texas at Austin, The Research and Development Center for Teacher Education, February 1983.

10. Hughes, R., Jr., & Hukill, H. *Participant characteristics, change, and outcomes in preservice clinical teacher education* (Report No. 9020). Austin, TX: The University of Texas at Austin, The Research and Development Center for Teacher Education, July 1982.

11. O'Neal, S.F. *Supervision of student teachers: Feedback and evaluation* (Report No. 9047). Austin, TX: The University of Texas at Austin, The Research and Development Center for Teacher Education, February 1983.

References

Chiarelott, L., Davidman, L., & Muse, C. Evaluating pre-service teacher candidates. The Clearing House, 1980, 53, 295-299.

Funk, F.F., Hoffman, J.L., Keithley, A.M., & Long, B.E. Student teaching program: Feedback from supervising teachers. The Clearing House, 1982, 55, 319-321.

Gitlin, A. Horizontal evaluation: An approach to student teacher supervision. Journal of Teacher Education, 1981, 32, 47-50.

Goldhammer, K. Teacher education: Reality, hope, and promise. Journal of Teacher Education, 1981, 32, 25-29.

Grossman, G.C. A comparison of the effectiveness of student teachers who have had extensive early field experience with those who have not. Ellensburg, WA: Central Washington University, 1980. (ERIC Document Reproduction Service No. ED 207 943)

Haviland, M.G., & Haviland, C.P. Student teacher evaluations and inflation. Journal of College Placement, 1981, 42, 67-69.

Hoffman, R.A., & Gellen, M.I. A comparison of self-evaluations and classroom teacher evaluations for aides in a pre-student teaching field experience program. The Teacher Educator, 1981, 17, 16-21.

Howey, R., Yarger, S., & Joyce, B. Reflections on preservice preparation: Impressions from the national survey. Journal of Teacher Education, 1978, 29, 38-40.

Johnston, J.M., & Hodge, R.L. Self-evaluation through performance statements: A basis for professional development. Journal of Teacher Education, 1981, 32, 30-33.

Mitsakos, C.L. Teacher evaluation programs. 1979. (ERIC Document Reproduction Service No. ED 189 828)

McEwing, R.A. Skills related to teaching: Perceived importance and mastery. 1980-81 Teacher Education Program evaluation study. Pocatello, ID: Idaho State University, College of Education, August 1981. (ERIC Document Reproduction Service No. ED 209 189)

Newport, J.F. Users approve of a new way to evaluate student teachers. The Clearing House, 1982, 55, 414-416.

Reiff, J.C. Evaluating student teacher effectiveness. College Student Journal, 1980 14, 369-372.

Schalock, D. Research on teacher selection. In D.C. Berliner (Ed.), Review of Research in Education (Vol. 7). New York: American Educational Research Association, 1979.

Snedecor, G., & Cochran, W. Statistical methods. Ames, IA: The Iowa State University Press, 1967.

Walker, R.T. Myths in student teacher evaluation. English Education, 1981, 13, 10-16.

Weller, R.H. Verbal communication in instructional supervision. New York: Teachers College Press, 1971.

Appendix A

Conference Summary
028-029-030

Final (3-Way #1)

PROCESS			
Who US		32%	
CT		29%	
ST		39%	
Direction			
CT-ST		CT-ST= 15%	ST-US= 28%
CT-US		CT-US= 14%	US-CT= 8%
ST-US		ST-CT= 12%	US-ST= 23%
Type		CT	US
CT-1		1%	13%
CT-2		0	34
CT-3		28	15
CT-4		10	3
CT-5		1	2
CT-6		27	13
CT-7		11	5
CT-9		22	15
ST-3		17%	
ST-4		10	
ST-6		44	
ST-7		16	
ST-8		1	
ST-9		12	

CONTENT	ALL	CT	ST	US	ALL	CT	ST	US
Teaching Events					57%	57%	64%	50%
Generality								
Specific					31%	24%	42%	22%
General					69	76	58	78
Focus								
O-1					0%	0%	0%	0%
M-2					15	13	13	22
I-3					36	31	49	19
N/A					49	56	38	59
Domain								
C-1					16%	14%	21%	10%
A-2					0	0	0	0
D-3					9	8	12	6
N/A					75	78	67	84
Organization of events					29%	28%	22%	36%
Activities					36	46	26	35
Protocol					64	54	74	65
Other					14%	15%	14%	14%