Validation of a teaching performance device

Colleen Cluff Caputo

Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Teacher Education and Professional Development Commons

Recommended Citation

Caputo, Colleen Cluff, "Validation of a teaching performance device " (1975). Retrospective Theses and Dissertations. 5620. https://lib.dr.iastate.edu/rtd/5620

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

Xerox University Microfilms
300 North Zeeb Road
Ann Arbor, Michigan 48106
CAPUTO, Colleen Cluff, 1935-
VALIDATION OF A TEACHING PERFORMANCE DEVICE.

Iowa State University, Ph.D., 1975
Education, home economics

Xerox University Microfilms, Ann Arbor, Michigan 48106

THIS DISSERTATION HAS BEEN MICROFILMED EXACTLY AS RECEIVED.
Validation of a teaching performance device

by

Colleen Cluff Caputo

A Dissertation Submitted to the Graduate Faculty in Partial Fulfillment of The Requirements for the Degree of DOCTOR OF PHILOSOPHY

Major: Home Economics Education

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa

1975
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Purpose of the Study</td>
<td>5</td>
</tr>
<tr>
<td>Definition of Terms</td>
<td>6</td>
</tr>
<tr>
<td>PBTE/CBTE/CBE</td>
<td>6</td>
</tr>
<tr>
<td>REVIEW OF LITERATURE</td>
<td>7</td>
</tr>
<tr>
<td>Competency Assessment</td>
<td>8</td>
</tr>
<tr>
<td>Issues and concerns</td>
<td>8</td>
</tr>
<tr>
<td>Measurement considerations</td>
<td>17</td>
</tr>
<tr>
<td>General aspects</td>
<td>19</td>
</tr>
<tr>
<td>Criteria for evaluation</td>
<td>21</td>
</tr>
<tr>
<td>Characteristics Desired in Evaluative Devices</td>
<td>24</td>
</tr>
<tr>
<td>Criteria for using and developing evaluative devices</td>
<td>27</td>
</tr>
<tr>
<td>METHOD OF PROCEDURE</td>
<td>29</td>
</tr>
<tr>
<td>Instrument Development</td>
<td>29</td>
</tr>
<tr>
<td>Preliminary instrument</td>
<td>29</td>
</tr>
<tr>
<td>Revised and expanded instrument</td>
<td>35</td>
</tr>
<tr>
<td>Revised instrument</td>
<td>36</td>
</tr>
<tr>
<td>Expanded instrument</td>
<td>42</td>
</tr>
<tr>
<td>Population and Data Collection</td>
<td>45</td>
</tr>
<tr>
<td>Data Analysis</td>
<td>48</td>
</tr>
<tr>
<td>Revised student teacher evaluation</td>
<td>48</td>
</tr>
<tr>
<td>Establishment of cluster composition</td>
<td>55</td>
</tr>
<tr>
<td>An assessment device for human relations skills</td>
<td>57</td>
</tr>
<tr>
<td>FINDINGS AND DISCUSSION</td>
<td>Page</td>
</tr>
<tr>
<td>---------------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Assessment of RSTE Items</td>
<td>60</td>
</tr>
<tr>
<td>The most promising classroom performance items</td>
<td>60</td>
</tr>
<tr>
<td>The least promising classroom performance items</td>
<td>70</td>
</tr>
<tr>
<td>The remaining classroom performance items</td>
<td>72</td>
</tr>
<tr>
<td>The most promising cumulative over classes items</td>
<td>74</td>
</tr>
<tr>
<td>The least promising cumulative over classes items</td>
<td>79</td>
</tr>
<tr>
<td>The remaining cumulative over classes item</td>
<td>82</td>
</tr>
<tr>
<td>Cluster Composition of RSTE Items</td>
<td>86</td>
</tr>
<tr>
<td>Assessment of Human Relations Skills</td>
<td>94</td>
</tr>
<tr>
<td>SUMMARY</td>
<td>102</td>
</tr>
<tr>
<td>LITERATURE CITED</td>
<td>110</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>117</td>
</tr>
<tr>
<td>APPENDIX A: HOME ECONOMICS TEACHER EDUCATION MODEL</td>
<td>120</td>
</tr>
<tr>
<td>APPENDIX B: STUDENT TEACHER EVALUATION (STE) AND EVALUATOR'S GUIDE</td>
<td>123</td>
</tr>
<tr>
<td>APPENDIX C: REVISED STUDENT TEACHER EVALUATION (RSTE) AND EVALUATOR'S GUIDE</td>
<td>130</td>
</tr>
<tr>
<td>Part A</td>
<td>130</td>
</tr>
<tr>
<td>Part B</td>
<td>132</td>
</tr>
<tr>
<td>APPENDIX D: RSTE ITEM DESCRIPTORS AND CONFERENCE QUESTIONS</td>
<td>135</td>
</tr>
<tr>
<td>APPENDIX E: NUMERICAL CALCULATIONS FOR PRELIMINARY INTERRATER RELIABILITY ASSESSMENTS</td>
<td>139</td>
</tr>
<tr>
<td>Appendix</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>F</td>
<td>APPENDIX F: AN ASSESSMENT DEVICE FOR HUMAN RELATIONS SKILLS (HRS)</td>
</tr>
<tr>
<td>G</td>
<td>APPENDIX G: CORRESPONDENCE</td>
</tr>
<tr>
<td>H</td>
<td>APPENDIX H: MOST PROMISING ITEMS IN STE</td>
</tr>
<tr>
<td>I</td>
<td>APPENDIX I: RSTE ITEMS NOT USED IN CLUSTERS</td>
</tr>
<tr>
<td>J</td>
<td>APPENDIX J: ITEMS RECEIVING X RATINGS</td>
</tr>
<tr>
<td>K</td>
<td>APPENDIX K: CHI-SQUARE FOR HRS ITEMS WITH X RATINGS INCLUDED</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Illustration of non-significant teacher by judge interaction (TxJ) effect: Item 6</td>
<td>65</td>
</tr>
<tr>
<td>2.</td>
<td>Illustration of highly significant teacher by judge interaction (TxJ) effect: Item 16</td>
<td>68</td>
</tr>
<tr>
<td>3.</td>
<td>Illustration of ratings which resulted in a highly significant teacher (T) effect and non-significant judge (J) effect: Item 14</td>
<td>78</td>
</tr>
<tr>
<td>4.</td>
<td>Illustration of ratings which resulted in a highly significant teacher (T) and judge (J) effect: Item 13</td>
<td>81</td>
</tr>
<tr>
<td>5.</td>
<td>Illustration of ratings which resulted in a non-significant teacher (T) effect and significant judge (J) effect: Item 12</td>
<td>85</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Actual and expected mean squares of the three factor analysis of variance design</td>
<td>41</td>
</tr>
<tr>
<td>2.</td>
<td>Expected mean squares of the analysis of variance design</td>
<td>51</td>
</tr>
<tr>
<td>3.</td>
<td>Expected mean squares of the analysis of variance design</td>
<td>52</td>
</tr>
<tr>
<td>4.</td>
<td>Expected mean squares of the analysis of variance design</td>
<td>54</td>
</tr>
<tr>
<td>5.</td>
<td>Frequency distribution of tallies from human relations skills device</td>
<td>57</td>
</tr>
<tr>
<td>6.</td>
<td>Reduced frequency distribution of tallies from human relations skills device</td>
<td>58</td>
</tr>
<tr>
<td>7.</td>
<td>F ratios of analysis of variance components of most promising items in classroom performance section of RSTE</td>
<td>63</td>
</tr>
<tr>
<td>8.</td>
<td>F ratios of analysis of variance components of least promising items in classroom performance section of RSTE</td>
<td>71</td>
</tr>
<tr>
<td>9.</td>
<td>F ratios of analysis of variance components of remaining items in classroom performance section of RSTE</td>
<td>73</td>
</tr>
<tr>
<td>10.</td>
<td>F ratios of analysis of variance components of most promising items in cumulative over classes section of RSTE</td>
<td>75</td>
</tr>
<tr>
<td>11.</td>
<td>F ratios of analysis of variance components of least promising items in cumulative over classes section of RSTE</td>
<td>80</td>
</tr>
<tr>
<td>12.</td>
<td>F ratios of analysis of variance components of remaining item in cumulative over classes section of RSTE</td>
<td>83</td>
</tr>
<tr>
<td>Table</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>15. Intercorrelations among items in Cluster L, evaluation methods</td>
<td>92.</td>
<td></td>
</tr>
<tr>
<td>16. Intercorrelations between teaching performance clusters</td>
<td>93a.</td>
<td></td>
</tr>
<tr>
<td>17. Significant chi-squares for human relations skills behaviors</td>
<td>96.</td>
<td></td>
</tr>
<tr>
<td>18. Non-significant chi-squares for human relations skills behaviors</td>
<td>100.</td>
<td></td>
</tr>
<tr>
<td>19. Analysis of variance components of most promising items (Gilbert 1974, pp. 50-51)</td>
<td>157.</td>
<td></td>
</tr>
<tr>
<td>20. RSTE items receiving X ratings</td>
<td>161.</td>
<td></td>
</tr>
<tr>
<td>21. HRS items receiving X ratings</td>
<td>162.</td>
<td></td>
</tr>
<tr>
<td>22. Significant chi-squares for human relations skills behaviors with X ratings included</td>
<td>164.</td>
<td></td>
</tr>
</tbody>
</table>
INTRODUCTION

Educational research on teacher effectiveness has resulted in little agreement on a definition of a good teacher or how to evaluate teachers. Literature on the subject is overwhelming with more than 10,000 published studies (Dunkin and Biddle 1974). Not only is the literature substantial, but even bibliographies on the subject have become unmanageable. Generally, this research has provoked poor reviews. As the Committee on Criteria of Teacher Effectiveness of the American Educational Research Association (1953) commented:

The simple fact of the matter is that, after 40 years of research on teacher effectiveness during which a vast number of studies have been carried out, one can point to few outcomes that a superintendent of schools can safely employ in hiring a teacher or granting him tenure, that an agency can employ in certifying teachers, or that a teacher-education faculty can employ in planning or improving teacher-education programs (p. 657).

Early studies avoided looking at the actual process of teaching in the classroom. The typical study would be an investigation of causative factors such as classroom size, curriculum innovation or teaching method against some criterion of teacher effectiveness, without any attempt to assess what was actually going on in the classroom. Gage (1963) described such approaches as treating the classroom as a "black box" into which were fed teachers, pupils, hard-
ware and software and out of which came various results and more or less pupil learning. The crucial events within the classroom, the point at which teachers, pupils, tasks and equipment come together were ignored.

Most of the research on teacher effectiveness took place prior to 1950. In the past two decades, research on teaching has undergone considerable reorientation. Researchers have turned from the study of teacher qualities to the observation of actual instances of instruction in the classroom. Present research focuses more often on the processes of teaching. These research efforts are also more likely to reflect theory and to take into account contextual effects. Many of these changes have come about due to the demand for educational accountability and in response to criticism by research reviewers.

Although reviewers have been critical of results of teaching research, they have also noted the underfinancing of research efforts (Rosenshine and Furst, 1971). Very few of the funds available for education are spent on research, and of these funds most are committed to investigating areas other than teaching. In comparison with expenditures for teacher training, curricular innovations, performance contracting, development and promotion of new educational devices, or laboratory studies of human learning, support of
teaching research is lacking. Adjusting these appropriations and encouraging rigorous analytical research on the teaching process should improve future understanding of the process and its effect on pupils.

Education for accountability has not only influenced research on teaching but has also caused a rapid growth in competency-based education. The concept of performance or competency-based teacher education emerged in the latter part of the sixties spurred on by increasing demands for accountability, relevance and cost-effective schooling. It is difficult to assess accurately the spread of the movement. Precise figures on the number of teacher education programs currently attempting the transition to a competency-based approach are not available. However, approximately 30 states have mandated such a change, either through legislative action or through the certifying power of the state education agency (Potter, 1975).

The literature contains numerous articles on competency-based teacher education (CBTE) as one promising alternative way to prepare teachers. However, a major dilemma exists in implementing a CBTE program and that is the development and validation of appropriate evaluative devices. The major effort of developers up to now has been directed toward organizing and administering the program, and developing
materials. McDonald (1974) found in consulting with individuals engaged in or interested in the performance-based movement, that the most pressing need of the movement was a research and development effort to describe and measure teaching competency. Other surveys have also reaffirmed the need for better assessment and evaluation.

Fanslow (1974) observed that the issue of assessment in a CBE program necessitates a critical look at the basic concept of CBE and suggests that since a mastery level of performance is expected of students, the traditional statistical techniques for validating the measurement devices may be ineffective. Fanslow proposes the need to investigate experimental designs and use a variety of statistical procedures in attempts to determine the effectiveness of evaluation devices that can be used in a CBE program.

There has also been an investigation by the home economics profession on the feasibility of CBE programs, not only in the teacher education component but also in home economics subject matter. The Home Economics Education Department at Iowa State University has been involved in a study of program and curriculum revision, part of which involves developing and testing an assessment device to use in evaluating the competencies of home economics student teachers related to the teacher-learning process. As part of the
departmental study, this investigation will focus on the re-
vision of the instrument and the validation of items within
the instrument that observers can use to reliably assess
teaching performance and that discriminate between teachers.

Purpose of the Study

First year home economics teachers teaching in Iowa and
South Dakota and graduating from Iowa State University and
South Dakota State University participated in the present
study, which was designed to validate a teaching performance
device. Specific objectives for this study were:

1. To revise the Student Teacher Evaluation: Assessing
   the Teaching-Learning Process (STE).

2. To identify items in the Revised Student Teacher
   Evaluation: Assessing the Teaching-Learning Process
   (RSTE) that reliably assess classroom teaching per-
   formance and discriminate among teachers.

This study is limited geographically to first year home
economics teachers in Iowa and South Dakota and the sample
will include only graduates from Iowa State University and
South Dakota State University who participated as student
teachers in the study by Gilbert (1974).

Assumptions of the study are:
1. Subject matter areas taught within the respective areas of home economics will not influence overall teaching performance.

2. School environment and type of pupils in the classroom will not greatly influence the ability of the teacher to perform in the classroom.

3. Of the 77 student teachers in Gilbert's study, the 40 selected to participate in this study are not different from the other 37 student teachers.

4. Classroom observations obtained during class sessions are representative of the teacher's performance.

5. The teachers' performance will be stable and not changing from February to April, 1975.

**Definition of Terms**

**PBTE/CBTE/CBE**

Performance-Based Teacher Education/Competency-Based Teacher Education/Competency-Based Education are defined to include knowledge, performance, and values related to initial and continuing professional education of teachers and are used interchangeably. The terms will be used in this study as they were used in the articles and studies reviewed.
REVIEW OF LITERATURE

Within the past few years, educational literature has abounded with discussions of teacher competence and competence assessment throughout the writings in competency based teacher education and/or performance based teacher education. The focus of this study was the validation of items that observers can use to reliably assess teaching performance and that discriminate between teachers. Although the study was not directly concerned with competency based teacher education, to a degree it is addressed to a major dilemma in current as well as future CBTE programs, that of the lack of appropriate evaluative devices.

This study was part of a larger departmental study directed at the problem of evaluating the competencies of home economics student teachers. A historical review of competency based education and the development of CBE in home economics was done by Gilbert (1974). Due to the amount and scope of assessment literature, this review will be limited to competency assessment, issues, concerns and measurement considerations and characteristics desired in evaluative devices.
Issues and concerns

Researchers investigating assessment of teacher performance and behavior indicate that competency assessment has always presented a problem in teacher evaluation. Generally, the reports on assessment are inconclusive, and the lack of adequate instruments for evaluation noted repeatedly. McDonald (1972) attributed part of the problem to the connotations that have been attached to evaluation and assessment. For many people evaluation means testing and then ranking individuals with respect to one another. For others, evaluation has punitive connotations of selection and rejection, or approval and disapproval. Individuals have been tested at various stages in life, and the results of these tests have been used by others as a basis for important decisions about their future. Therefore, it is not surprising that the very words "assessment" and "evaluation" engender animosity for individuals as well as educators.

Schalock (1974) perceives the problem as not a matter of connotation, but rather one of massive conceptual confusion. The language that has emerged on assessment is confusing. The terms measurement, evaluation, and assessment seem to be used interchangeably and efforts are rarely made to tie a term to the operations that give it meaning. All three terms
represent different sets of operations and uses of the terms should reflect these differences.

Measurement as defined by Stevens (1951) technically refers to the assignment of numerals to objects or events according to rules. Evaluation refers to the assignment of worth or value to objects and events according to standards as defined by Glass (1971). Unlike measurement and evaluation, assessment has not received formal technical definition in the education literature.

Schalock (1974) proposes that assessment assume a technical meaning that complements the existing terms of measurement and evaluation. Specifically, he suggests that the term be used to refer to the identification, collection, reduction, analysis and use of information in targeted, adaptive decision-making. Womer, as cited in Schalock, defined assessment as a program to improve educational decision making by securing information about the outcome of education. Incorporating the preceding definitions into a framework, assessment becomes a more inclusive term than either measurement or evaluation although it includes both concepts.

The literature supports that generally three criteria are recognized for assessing teacher competency: knowledge, performance, and product (or consequence) criteria. At the
conference on CBTE held in Houston, Texas in May, 1971, two major criteria for assessing teacher competency were discussed, the criterion of teacher classroom performance, and the criterion of pupil performance. Both were recognized as significant, but the general agreement was that pupil performance constituted the ultimate criterion. This issue continues to be debated in the literature with Potter (1975), Weber (1974), Wolfe and Bugg (1974) among those supporting the pupil outcome criterion as a measurement of teacher competence. Authors supporting teacher assessment of skills and knowledge in a performance setting are Broudy (1967), Cheong (1970), and McDonald (1972). An eclectic point of view appears repeatedly in the literature and is sustained by Turner (1972b), Merwin (1973), Soar (1973) and Hatfield (1974). Each author presents convincing rationale for his viewpoint. The major objection to the use of the pupil product criterion stems from the difficulty in measuring outcomes. Added to this obstacle is the difficulty in measuring pupil needs, interest, motivation, and availability of resources, and the even greater difficulty in measuring the immediacy of the teacher effect on pupil achievement.

The short-term effects of teaching behavior can often be determined when assessment sessions are conducted under controlled conditions. An example of this method is analyzing
a microteaching episode or a short unit under simulated conditions. Several CBTE programs employ this technique.

The lack of conceptual tools for identification of long-term effects has made the analysis of these effects more difficult and complicated. Using the pupil product criterion as a measure of teacher competence would necessitate the differentiation of limited pupil performances from the long-term performances that may occur at intervals.

Because of the numerous difficulties in assessing pupil outcomes, many educators have turned to the teacher performance criterion as a measure of teacher competency. Data can be collected while the teacher or prospective teacher is actually performing the teaching act, but the methods of collecting data are primarily observational. The lack of trained observers and the lack of a standard of teacher performance cause most of the problems associated with the application of this criterion. Roth (1974) reported that at a conference on CBTE and related research a distinction needs to be made between using measures of student gain as a way of validating an observational measure of teacher competence in contrast to using student gain as a direct measure of teacher competence. The recommendation of this group was not to use student learning measures for evaluating individual teachers, but rather, to use the relationship as a learning device and
not as an evaluative criterion at this time.

In the last few years, some research efforts have gone into developing taxonomies of teaching behaviors and observation instruments. McDonald (1972) points out that in addition to providing descriptors, the taxonomy also provides for the systematic ordering of behavioral observations. He warns that many taxonomies of teaching behavior lack this ordering principle. McDonald advocates a taxonomy of teaching behaviors modeled after Guilford's (1967) structure of intelligent behavior. This model is three dimensional and includes the dimensions of teaching operations, organization of content, and type of content. Each dimension represents a way in which information is processed and the cells in the taxonomy combine to give a representation of a teacher's performance.

Turner (1972a) used the concept of teacher competency and developed a taxonomy in which he identifies six criterion levels related to professional preparation of teachers. Succinctly, Turner's levels can be viewed as six different criteria or several levels within the three major criteria of knowledge, performance, and product. Level six is a cognitive or knowledge criterion while levels three, four, and five are performance criteria and levels one and two are related to product criteria. Some educators believe levels one and two
are the most valid measures for assessing the effectiveness of teacher preparation programs. Rosner (1972) points out that teacher education does not address itself directly to the modification of pupil behavior but does accept the responsibility for modifying teacher behavior. Therefore, he argues that teacher education should introduce evaluative practices at level three, classroom application, to measure competence of students in the program.

Home economics educators have been involved in determining competencies of home economic student teachers. This endeavor was initiated at a Home Economics Education Research Seminar held at Iowa State University in 1962. The most recent of several subsequent workshops was sponsored by the American Home Economics Association and held at Iowa State University in 1974. The group, representing all professional facets of home economics, defined competencies specifically related to the professional educational components in the various areas of home economics. Participants advocated that competency development must continue after the baccalaureate degree to ensure professional growth as a continuous process for a home economist. The competencies and criteria at both the preprofessional and professional levels identified by the seminar participants were published by the American Home Economics Association (1974).
Although the pupil gain criterion has been acclaimed as the ultimate measure of teacher competence, a survey of one thousand teacher assessment studies made during the early part of the 1950's included only 20 which used teacher effectiveness as measured by pupil achievement as a criterion of evaluation. The other 980 studies rated teachers on judgments made by a supervisor or principal (Burkhart, 1969). Burkhart reported another teacher-assessment study made in 1966-67 which showed no increase in the use of pupil achievement as a criterion of teacher evaluation. Rosenshine and Furst (1971) concluded from their comprehensive review of empirical research that there were 11 "most promising" teaching variables which related to desired pupil outcomes and further stated that there were very few teaching skills which were related empirically to pupil gains.

Heath and Nielson (1974) reviewed the 50 studies from which Rosenshine and Furst (1971) identified the 11 "most promising" teaching variables. The investigators found 26 of the 86 operational definitions of teacher behavior did not correspond to the variable cited. The teacher was used as a unit of sampling in 59 or 75% of the variables; only 10 or 13% used the pupil. In the other 9 cases or 11%, the information was not sufficient to determine the unit of sampling. The investigators concluded that the analysis of
the research on the relation between specific teacher skills and student achievement fails to reveal an empirical basis for performance-based teacher education. The conception, design, and methodology of the studies preclude their use as such a basis. Operational definitions of teacher behavior that have little in common are often combined as examples of a single teaching variable. Operational definitions generally reflect the teacher in the traditional lecture-discussion role and are usually defined so vaguely as to be of little use in training teachers.

The literature contains numerous articles on competency assessment, teacher effectiveness, and teacher behavior. The number of studies that have been done on competency assessment is relatively small, and the number of reviews of this literature is even smaller. In a brief summary review of the research literature by Burdin and Mathieson (1972), they stated,

...the concept of performance-based teacher education (PBTE) is relatively new, and although there are a number of papers dealing with it, they consist mainly of opinions, discussions, and descriptions. They report very little research on PBTE or its companion term of competency (p. 61).

The investigator found the majority of research studies compared the effects on classroom performance of teachers trained in a traditional methods course or program with those trained in a competency-based methods course or program.
The purposes of these studies did not apply to this research.

A shortcoming of research in the area of teacher performance, as identified by Pankratz and Williams (1973), is ignoring the attempts to systematize a program of research in favor of independent research efforts to relate those variables fancied by the investigator. Most of these studies lack evidence of thought to theoretical or experiential bases for the research or significance the research might have for teaching or learning.

The American Home Economics Association has tried to encourage communication and cooperative research. At the competency-based workshop held prior to the 1974 national convention, participants who were interested in research on CBE made plans for sharing procedures and findings. As a part of that effort, Hughes, Fields, and Crawford (1975) conducted a survey of 61 home economics educators, including the original AHEA-CBE workshop participants and others who had expressed an interest in participating. Research in progress was reported in the areas of assessment of competency-based education, identification and validation of competencies and criteria, and development of competency-based curriculum. The report showed 9 studies in assessment, 12 on identification of competencies and criteria, and 11 studies on competency based curriculum. These findings look encouraging for future
reviews of competency-based programs and related research.

Several major concerns were expressed repeatedly in the materials reviewed. Statements involved aspects of competency identification and assessment. Expressions identified a need for:

- a firm knowledge of which teacher behaviors cause or facilitate educational growth in their students
- identification of interconnections among the various teaching behaviors (or skills or performances)
- simultaneous research on the relationships between teaching and student performance carried out across subject areas
- statements of teacher competencies and criteria by which they are to be judged, explicit enough to delineate assessment procedures to be used
- validation of teacher competencies through rigorously designed experimental studies
- measurement techniques that lend themselves to objective assessment of teacher competencies.

Measurement considerations

Several authors have acknowledged the lack of evaluation devices and difficulty in measuring competencies. However, few have actually identified and discussed the specific problems.

McDonald (1972) in looking at the state of the art in evaluation of teacher behavior, specifically measurement of subject matter knowledge, measurement of teaching skill, and measurement of teaching performance, recognized that good
indicators of the knowledge requisite to effective teaching are available. These measures are achievement tests which measure what a teacher knows about the subject to be taught. Although knowledge measures are important, they tell little about a person's ability to teach the subject. Techniques for measuring teaching skill are practically non-existent, yet few educators would disagree that the really significant aspect of teaching is the ability to behave in such a way that another person learns. Teaching skills contain many components the definition of which will only be possible when we have available data indicating a set of students' responses that are elicited by each teacher behavior. Instruments exist for observation of teaching performance, although most observations of such performances are mediocre and desultory. Teaching performance is viewed as a complex of knowledge and teaching skill extending over time, involving many specific items of knowledge, and usually involving several specific skills.

It appears that obtaining evidence regarding the teacher's demonstration of consequence competencies is the most difficult to collect. Evidence regarding performance competencies is somewhat less difficult to obtain, while evidence about subject matter knowledge competencies is generally easiest to collect. Cooper, Jones, and Weber (1973) caution educators
not to specify and measure certain competencies simply because they are easily measured while avoiding others which are more difficult.

**General aspects** Merwin (1973) identifies general measurement aspects involved in the assessment of a teacher's performance as errors and sampling. Error sources of concern are those due to lack of comparability of conditions under which the "measure" was taken, errors in observing and recording behavior and inaccuracies in the matching of the observed behavior against a criterion behavior. Potential errors due to the fallibility of humans as observers should also be considered.

Any assessment of performance involves only a sampling of a behavior repertoire. It is assumed that investigators are interested in the "performance" of teachers in a large population rather than the specific sets of circumstances at the particular time and place that the assessment takes place. A matter of major concern revolves around sampling which will permit defensible generalizations. Representative of the more relevant variables that must be considered in looking at the sampling problems related to assessment of performance in the classroom are, elements of time, environmental factors surrounding the performance under observation, and characteristics of both the pupils and the type of learning
involved.

An assessment program requires criteria by which to sample the domain of teaching behaviors. Lack of standardized assessment procedures causes each investigator to select those aspects of teaching performance that he considers significant. McDonald (1972) advocates overcoming the deficiency by the teaching profession stating precisely what is to be learned as a consequence of being educated in a teacher training program. He proposed that first a modest list of objectives and criteria for measuring their attainment be developed; second, those behavioral events that are reliably measurable be identified; third, that the intercorrelations among these events be determined; and finally, that the variation in teaching performances on a variety of skills be assessed.

Another measurement concern expressed by McDonald (1972) is the lack of agreement on the unit to be measured. Questions posed are: What is the basic unit of teaching behavior? Should it be a single question asked by a teacher or a pattern of a group of related questions? Should it be a dyad of teacher behavior and student behavior? Although there is no inherent contradiction between these various levels of analysis a great deal of effort is spent debating the virtues of each rather than bringing together investi-
gators working at each level and expanding ways to think about two different levels of analysis.

A frequent practice in CBTE programs is designating a certain cutoff percentage as the passing score for performance objectives. Quirk (1974) believes this makes little sense from a measurement point of view. The percentage of items related to an objective which a candidate answers correctly is a function not only of the content of the items but also of the difficulty of the items.

Ward (1973) suggests certain types of teaching skills must be judged according to whether a particular form of the skill does or does not occur rather than according to the number of times the skill is used in a lesson or the percentage of use as compared to total occurrences of it and other related skills. However, Messick (1971) supports monitoring the "unintended" as well as intended outcomes of each learning experience. Rather than assuming only the specific skills under development, one should also consider the need to monitor retention of all skills.

**Criteria for evaluation** Two kinds of criteria are accepted for evaluating teacher performance; one is the criterion of teacher performance, and the other the criterion of pupil performance. Merwin (1973) believes if pupil performance is used as the base for assessment there is a
need to state the competency in terms of pupil behavior. An assessment must be made in terms of a behavioral change based on a minimum of two observations, one before and the other after intervention by the teacher. The investigator observes and records the performance relevant to the teacher competency under consideration. Finally, the observer must accurately identify the teacher's contribution to the change observed.

Problems associated with the pupil performance base include the difficulty in reliably measuring changes in behavior; the attribution of causation; and the requiring of all pupils to attain the same objective irrespective of the needs of the pupils. If the competencies are written in terms of ability to bring about change in pupils, the process must involve separation of those changes as to those coming from the teachers efforts and those coming from other sources. A choice of pupil performance as the criterion doubles the problems of observation and the recording of behavior (Merwin, 1973).

There will be measurement errors involved at both ends of the process, making the assessment of what has changed and the degree of change less reliable than the pre- or post-status measures used to determine it (p. 14).

If teacher performance is used as the base for assessment there is little change in the steps involved under pupil
performance according to Merwin (1973). Teacher performance assessment necessitates research which provides evidence of the relationship between teacher performance and changes in pupil behavior. Thus, still required as in pupil performance assessment is someone to adequately measure change in pupil behavior and in addition, in teacher performance under conditions controlled in such a way as to establish the relationship.

Classroom observation is both complicated and expensive to obtain. Therefore, teacher performance is often measured in a simulated environment. Researchers are questioning the reliability of performance measures obtained in a micro-teaching or other simulated teaching situation. Questions asked by Quirk (1974) are: How consistent is the teacher's behavior over time? What is the effect of known versus unknown pupils on the behavior of the teacher? What is the effect of pupil practice on teacher behavior? Other questions of concern posed by Heath and Nielson (1974) are: What is to be taught and who is to be taught? Is there a possible interaction between teacher-variables and content? Is teacher behavior different for different groups? Quirk also points out that the influence of reliability on the standard error of measurement of a test also holds true with performance tests.
McDonald (1972) suggests a method to evaluate teaching performance in terms of actual classroom performance is to pair the data obtained under controlled conditions, such as micro-teaching, with the performance of the teacher "on the job". This approach would serve to validate the testing technique employed in the controlled environment.

Thus, while two approaches for evaluating teaching performance appear repeatedly in the literature, each presents some major problems in the resolution of the approach. Pupil performance includes the problems of adequate measures of pupil achievement while teacher performance involves the delineation and validation of what teaching behaviors really make a difference in pupil performance.

Characteristics Desired in Evaluative Devices

In a speech at the national meeting of the American Home Economics Association, Fanslow (1974) discussed desirable characteristics of evaluative devices for assessing teacher performance as content validity, reliability, objectivity and usability. Content validity was further subdivided into sampling validity and face validity. The characteristics were defined as:

content validity - the degree to which the evaluative device assesses the desired competencies.
sampling validity - a judgement that the evaluative devices not only represent content but that the components within the device are further weighted in proportion to the importance of the competency being assessed.

face validity - a judgement that items in the device represent appropriate content.

reliability - the degree to which scores on the evaluative device tend to repeat themselves and will produce consistent results.

objectivity - the degree to which the evaluative scores are unaffected by biases of the test scorer.

usability - represents a variety of considerations related to the ease of use of the evaluative device and includes concerns such as the conditions necessary for data collection, types of individuals for which the device is suitable, availability of reliability and content validity data, and time and money costs associated with the evaluation devices.

She concluded that if evidence is available which suggest that these characteristics are present at acceptable levels, decision makers can be confident of the resulting evaluation. If no or limited evidence is available, care must be exercised in making judgements on the data and effort expended in developing evaluative devices which meet these minimum criteria. She viewed data as perhaps coming from several sources, i.e., achievement tests, observational devices and perhaps pupil achievement.

Brun (1970) reviewed and discussed qualities desired in observational instruments to describe and evaluate behaviors that occur in the educational setting of the class-
room. Qualities identified as desired in any measurement procedure were objectivity; content, criterion-related, and construct validity; and reliability.

Further discussion of the list of minimum characteristics, content validity, objectivity, reliability and usability, needed in the various evaluative devices was made by Hughes and Fanslow (1975). The dilemma of reliability levels approaching zero if reliability estimates are calculated from an achievement test in which all individuals meet or approach a predetermined level of competence were recognized. It was suggested that while a judgemental decision could be made as to the adequacy of the ideas a preferred method would be to administer the test in a setting in which a spread of scores could occur. Reliability coefficients could then be calculated by traditional methods, such as Kuder-Richardson formula 20 or 21. If acceptable coefficients were obtained the objective test could be confidently used to assess the level of competency an individual has obtained. If the coefficient proved too low, item analysis data would provide information as to which test items should be revised in order to improve the test.

The issue of reliability related to observational devices also has the dilemma of approaching zero when individuals meet or approach a predetermined level of competence.
Again, this problem could be resolved by administering the observational device in a setting in which a spread of scores could occur.

Criteria for using and developing evaluative devices

Statistical procedures for analyzing data have far surpassed procedures for analyzing the means by which the data were obtained. Researchers often omit informing the reader regarding methodological problems experienced in the study or how the problems were approached. Methodological problems are defined as including the making of decisions about issues such as sampling, the use of context and inference by observers, observer training and selection, and questions of validity and reliability of observational measures. A computer search of the literature of observation studies conducted by Herbert and Attridge (1975) found although observation techniques had been used in the investigation little or no emphasis was given to discussing the methodological problems. Of the 832 articles reviewed only eleven included discussions on observation and reliability, five dealt with observation and validity, one with observation and sampling, and none with observation and inference.

In an effort to confront the problems associated with observational devices, Herbert and Attridge (1975) developed a criteria to guide users and developers of these devices.
Thirty-three criteria which devices should meet were divided into three main types: identifying, validity, and practicality criteria. Identifying criteria (6) solicit information which enables the user to select the correct instrument for his purpose and application. Validity criteria (15) specify the provision of evidence which allows both developer and user to decide whether the instrument represents accurately and consistently the events it claims it describes. Practicality criteria (12) pertain to the administration of the instrument and to the ease with which it is used and the results disseminated.

After a consideration of the above criteria and recognition of the need for a more adequately refined observational device of teaching performance, the purpose of this study was defined to further the refinement of an observational device. The device selected to be refined was the Student Teacher Evaluation: Assessing the Teacher-Learning Process (STE), a device that had been developed by department faculty members and graduate students.
METHOD OF PROCEDURE

The major purposes of the study were to revise the Student Teacher Evaluation: Assessing the Teaching-Learning Process (STE) and to identify items that reliably assess classroom teaching performance and discriminate among teachers. A secondary purpose was to investigate the stability of teaching performance over classes. This chapter includes reports of the preparation of the preliminary instrument (STE), the revision and expansion of the STE (RSTE and HRS), the collection of data using the revised and expanded instrument, and the analysis of data.

Instrument Development

Preliminary instrument

The research was initiated in the Fall of 1972, as a project funded by the Home Economics Research Institute (HERI), College of Home Economics, Iowa State University. The title of the project was Evaluation Devices for Assessing the Competency of the Teaching-Learning Process, Hughes, 1972.

Specific objectives of the research project were to:

1. Develop an evaluation matrix for the concept of teaching learning process.

2. Refine objectives of the home economics teacher education program developed in 1971-72 by Home Economics Education Department faculty.
3. Determine the aspects of attainment for each competency related to the objectives of the teaching-learning process.

4. Select or develop evaluation devices.

As part of a departmental seminar, faculty and graduate students refined the objectives of the home economics teacher education program and developed an evaluation matrix for the undergraduate teacher program (Appendix A). As a result of the seminar, interested faculty and graduate students continued work on the model, focusing on and developing evaluation devices for one of the five major competencies of the model -- the teaching-learning process.

A pilot study was conducted at the beginning of Spring Quarter, 1973. Six cooperating teachers and their principals were invited to Iowa State University for a one day workshop to introduce them to the preliminary form of the teaching-learning process evaluation device and to help prepare them for use of the device with their student teachers during Spring Quarter, 1973. Concurrently, college supervisors who would be working with the pilot group student teachers were trained in the use of the device. Data from the pilot study suggested refinement of the device and the feasibility of being able to obtain reliability across many judges.

During summer 1973 both the evaluation model and the
device assessing competencies related to the teaching-learning process were refined. Hausafus (1973) developed the evaluation component of the device from the objectives of related courses in the undergraduate home economics education curriculum.

A second project to continue objectives of the study, was funded by the Home Economics Research Institute (HERI). The project titled, Assessing the Competency of the Teaching-Learning Process Prior to and During Student Teaching, Hughes and Fanslow, 1973, contained the specific objectives to:

1. Evaluate the competencies of home economics education student teachers at ISU related to the teaching-learning process at the completion of the fourth, sixth, and eighth week of student teaching.

2. Ascertain the perceptions of cooperating teachers and principals toward the performance-based teacher education system implemented in their schools.

3. Further revise and develop an assessment device for the competencies related to the teaching-learning process for use with undergraduate students prior to student teaching.

Gilbert (1974) who had been involved with the project from its initiation continued the investigation for her dissertation topic. The instrument used in the pilot study was expanded to include in addition to items from the adaptation of Thatcher's instrument, selected items from an instrument by Menne (Gilbert 1974). The 14 items which identified competencies in evaluation skills desired for
student teachers were also included. The evaluation items were the only items in the instrument that had not been pretested for the ability to discriminate between teachers.

The 50-item instrument, Student Teacher Evaluation: Assessing the Teaching-Learning Process (STE), included 32 items measuring classroom performance, 11 items assessing relationship skills, 14 items devoted to evaluation skills, and 4 items evaluating management and professionalism (Appendix B). A 99-point scale was used to record the responses to the instrument. Directions for use of the 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. A number between 51-99 was recorded if the teacher was above average, and a number between 1-49 was recorded if below average. A 50 indicated that the evaluator was uncertain about the behavior or that there was no opportunity to observe the behavior.

Preliminary reliability assessments based on the total score for the classroom performance section were established by submitting the instrument to three different groups composed of home economics faculty members, high school teachers and graduate students. Each of the 64 observers viewed a 15 minute videotape lesson taught by home economics student
teachers. After responding to the instrument and discussing the responses, two additional 15 minute videotapes were viewed and rated. Interrater reliability coefficients of .87, .89 and .90 were established.

Five training sessions were held for the 60 cooperating teachers and 11 college supervisors involved in the study. A procedure similar to that used to establish preliminary reliability assessment was followed. From the data collected during these sessions an 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 the orientation session as an additional source of variance showed no difference in ratings between orientation sessions.

Student teachers from Iowa State University and South Dakota State University participated in the study. Data were collected from three observations of 45 student teachers and from two observations of 32 additional student teachers. Two types of judges, cooperating teachers from the public schools and college supervisors from the university faculty, made the observations.

Two analyses of variance were computed for each of the
50 items in the instrument. The purpose of the two analyses of variance was to identify the most promising items for potential use in CBE programs and to project reliability coefficients for two judges on the items. Intraclass correlation coefficients were computed on each item for both analyses. This combination of statistical techniques assisted in determining if the competencies discriminated between student teachers, if the teaching centers affected the student teacher's performance, and if the two types of judges could reliably rate the items (Gilbert, 1974, and Fanslow and Wolins, 1975).

These investigators concluded that 24 items were promising for future research and that, in general, student teachers achieve given competencies at various levels during the student teaching experience. The F ratios of the 24 items indicated that while judges used some items differently, overall judges could discriminate between teachers using the designated items.

Analysis of the 22 least promising items revealed contradictory results. Inspection of the items indicated they were not observable until the second or third observation, which suggested the items should not be discarded until further investigation had been made using a different system of recording the observations. Gilbert recommended that a
variety of analyses be used for interpreting data collected in future research involving assessment of student teaching competencies.

Revised and expanded instrument

Findings and recommendations from the first phase of the research project were studied and used as a basis for planning objectives and submitting additional proposals. The proposals were submitted to two funding agencies for the needed resources. Funding for the project was provided by research funds from the HERI project, Evaluation Devices for Assessing the Competency of the Teaching-Learning Process, Hughes and Fanslow, 1974, and from the Graduate Dean's office for the proposal, Validation of an Evaluative Device through Assessing First Year Teaching Performance, Hughes, Fanslow, Adams, Caputo, 1974.

The objectives specific to this project were to:

1. Identify items that can reliably assess classroom teaching performance and identify those which discriminate among teachers in the classroom.

2. Investigate the usability of the observation device, The Student Teacher Evaluation: Assessing the Teaching-Learning Process (STE) in assessing the teaching performance of first year Iowa State University and South Dakota State University home economics education graduates teaching in the states of Iowa and South Dakota.

3. Discover if teaching performance, as measured by the STE, exhibits different levels of quality for first year teachers as compared to their quality of classroom performance during student teaching.
Revised instrument. The first step in the study was
to revise the STE in accordance with results and recommenda-
tions of the Gilbert study, and suggestions from cooperating
teachers and college supervisors. Items were reviewed for
clarity, objectivity and observability. Several items were
eliminated, others were reworded and a general reorganization
of the instrument occurred. Descriptive statements were
established for some items in order that the same elements
would be considered by each judge as they rated the teacher.
Item descriptors are found in Appendix D.

The 47 item Revised Student Teacher Evaluation: Assessing the Teaching-Learning Process (RSTE) is divided into two
parts. Part A, Classroom Performance, includes 25 items to
be observed each class period. Five items are assessed
through preconference observation of lesson plans and con-
fERENCE questions. The remaining twenty items are assessed
during the regular class period. Part A of the RSTE is
found in Appendix C. Since it was decided to observe three
class periods the instrument was reproduced on three different
colors of paper to facilitate class period identification.

Part B of the RSTE, Cumulative Assessment Over Classes,
includes 22 items to be observed over several class periods.
Some items in the Evaluation section are assessed through
evaluation devices, both formal and informal, prepared by the
teacher for the courses observed in the Classroom Performance section. Other items are assessed through specific conference questions regarding item content. Part B of the RSTE is also found in Appendix C and the conference questions in Appendix D.

A 99-point scale was selected by the judges to use in rating the items. This choice was based on two considerations. Individuals respond to rating scales with varying degrees of response variability, resulting in a particular response set. Brown (1970, p. 94) defines response set as the systematic biases in responses of the individual regardless of the content of the instrument. For example, some persons use the extreme categories and others limit their responses to the middle categories. Liu (1971) stated "if response set is to be controlled by assessing it as part of the measurement procedure, it should be given opportunity to occur" (p. 28). The use of a 99-point scale should provide opportunity for response set to occur.

The second reason pointed out by Liu (1971) for use of the 99-point scale is that small differences in the responses in the middle of the scale are not likely to be reliable, however, small differences at the extremes of the scale are more apt to be reliable. These differences in the extremes of the scale are given more weight when the original responses
on the 99-point scale are non-linearly transformed to normal deviates before data are analyzed. Hendricks (1975) further found that the transformation made the scale more linear and therefore produced scores which more nearly meet the assumptions of classical scaling theory. Further discussion of the transformation is contained in this chapter under data analysis.

The evaluation guide used by the judges instructed them to determine if the teacher observed was functioning below or above average on each specific item and to record the degree of certainty related to each decision (Appendix C). If the 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 judge was uncertain about the behavior. An X was recorded if there was no opportunity to observe the behavior and an 0 if the task or behavior were appropriate but not attempted by the teacher.

Training sessions were conducted by selecting from the department library videotapes of first year home economics teachers teaching a 30 to 45 minute class. During the sessions a tape was observed and rated using the Classroom Performance section of the RSTE. Items that had conflicting responses were discussed as well as the interpretation of the items.

Four videotapes and two classroom observations were used
to establish preliminary inter-observer reliability between
the two judges for the average correlation for a single item
and for the total score for the 20 items in the classroom
performance section (Part A) of the observational device.
Adams and the investigator independently rated videotapes
not used in the training sessions. All raw scores were
transformed to normal deviates prior to the reliability
calculations.

Reliability estimates for a single item and for the
total score were calculated. The formula used was:

\[ r_I = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_\epsilon} \]

where

- \( r_I \) = reliability estimate of a single item
- \( \sigma^2_T \) = teacher variance
- \( \sigma^2_\epsilon \) = error variance

Variance components for each calculation were estimated
from a three factor analysis of variance design using
teacher, judges, and items as the sources of variance. The
model used was:

\[ Y_{ijk} = \mu + T_i + J_j + I_k + TJ_{ij} + TI_{ik} + JI_{jk} + \epsilon_{ijk} \]

where:
\[ Y_{ijk} = \text{transformed score assigned ith teacher by jth judge for the kth item} \]

\[ i = 1,2,3\ldots6 \text{ teachers} \]

\[ j = 1,2 \text{ judges} \]

\[ k = 1,2,3\ldots20 \text{ items} \]

\[ \mu = \text{overall mean} \]

\[ T = \text{teacher effect} \]

\[ J = \text{judge effect} \]

\[ I = \text{item effect} \]

\[ TJ = \text{teacher by judge interaction effect} \]

\[ TI = \text{teacher by item interaction effect} \]

\[ JI = \text{judge by item interaction effect} \]

\[ \varepsilon_{ijk} = \text{random error} \]

The actual and expected mean squares for the ANOV are shown in Table 1.

The interrater reliability coefficient between two judges for a single item was .093 while the reliability coefficient between two judges over the total items was .58. Numerical calculations for the formulas are found in Appendix E.

The unusually low reliability estimate for a single item is not surprising because only one item was being observed and the number of items being observed is a well-established
Table 1. Actual and expected mean squares of the three factor analysis of variance design

<table>
<thead>
<tr>
<th>Source of variation&lt;sup&gt;a&lt;/sup&gt;</th>
<th>df</th>
<th>Mean Squares</th>
<th>Expected values of mean squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teachers (T)</td>
<td>5</td>
<td>.4141</td>
<td>( \sigma_e^2 + 40\sigma_T^2 )</td>
</tr>
<tr>
<td>Judge (J)</td>
<td>1</td>
<td>.4932</td>
<td>( \sigma_e^2 + 20\sigma_{TxJ}^2 + 120\sigma_J^2 )</td>
</tr>
<tr>
<td>Item (I)</td>
<td>19</td>
<td>.7079</td>
<td>( \sigma_e^2 + 2\sigma_{TxI}^2 + 12\sigma_I^2 )</td>
</tr>
<tr>
<td>Teacher by Judge (TxJ)</td>
<td>5</td>
<td>.1143</td>
<td>( \sigma_e^2 + 20\sigma_{TxJ}^2 )</td>
</tr>
<tr>
<td>Teacher by Item (TxI)</td>
<td>84</td>
<td>.2231</td>
<td>( \sigma_e^2 + 2\sigma_{TxI}^2 )</td>
</tr>
<tr>
<td>Judge by Item (JxI)</td>
<td>19</td>
<td>.1159</td>
<td>( \sigma_e^2 + 6\sigma_{JxI}^2 )</td>
</tr>
<tr>
<td>Error ((\varepsilon))</td>
<td>82</td>
<td>.0780</td>
<td>( \sigma_e^2 )</td>
</tr>
</tbody>
</table>

<sup>a</sup> Teachers random, judges and items fixed.
factor known to influence the size of the reliability coefficient. The low reliability estimate for the total score is probably a function of the sample size of six. Again, the size of the coefficient is not unusual since it is also known that variability within the sample affects the size of the reliability coefficient. A wide or small range of variance is possible for any given sample of six selected teachers.

Since all the competencies included in the instrument were not observable from viewing the videotapes, the analysis of variance was computed on only the 20 items in the Classroom Performance section. The two classroom observations provided an opportunity to pretest the Cumulative Over Classes section. Further testing of this section was provided by rating evaluation packets prepared by department faculty. From these two procedures, the observational device was judged suitable for use in the study.

**Expanded instrument** During 1973-74, Huyck, a faculty member in the Home Economics Education Department, responsible for teaching the human relations component of the undergraduate program developed an assessment device. The purpose of the instrument was to help students receive feedback on their human relations skills. The work on the device began by reviewing department objectives and compiling those
pre-service objectives related to human relations skills. From objectives, a competency was written and criteria for measuring the competency established.

Additional refinements were made on the device using resources from the HERI project *The Building and Testing of a Matrix for Performance Based Evaluation of Human Relationship Skills*, Huyck, 1974. Objectives of the project were to:

1. Determine whether a selected group of taught relationship skills can be expressed in behavioral criteria which:
   a. validly reflect the more complex relationship eliciting or resulting from the behavior.
   b. can be reliably interpreted by an evaluator with minimal training in these criteria.

2. Create a matrix which can be used to make a performance based evaluation of a teacher's human relationship skills.

3. Determine whether relationship skills taught in a classroom setting are transferred to the behavior of the learner,
   a. within the setting,
   b. beyond the learning setting to the student teaching experience.

4. Determine whether the behaviors taught and executed are ones which correlate with the enhancement of the pupil's freedom to learn.

The first objective of the Huyck project was accomplished by including the HRS device as an expansion of the RSTE project.
Further discussions between Huyck and the investigators led to the following agreements: the items in the device would not be altered, the teachers would be rated on a six point scale, and that the assessment of the teachers would be cumulative over classes.

Directions for use of the scale instructed the observer to place a numerical rating each time the behavior was used. However, this was adjusted to one rating of the ability to perform the behavior over classes. A rating of 1 was given if the teacher attempted the behavior with difficulty, a 2 rating if the behavior was performed with some competence, and a 3 rating if performed with ease and proficiency. A 4 was used if the raters were uncertain the behavior was performed. A rating of X indicated there was no opportunity for the teacher to exhibit the behavior, and an 0 indicated the behavior was appropriate but was not attempted by the teacher.

The 20 item instrument, An Assessment Device for Human Relations Skills (HRS), is divided into five sections. The three components of self, others and the interaction are expressed within the five sections. The element of self is conveyed in the Creating An Open, Positive Atmosphere section, and others in the Active Listening section. The interaction component is part of the Giving Positive Feedback, and Giving and Receiving Negative Feedback. The
HRS device is found in Appendix F.

Preliminary reliability assessments were established on the instrument by 28 cooperating teachers and faculty members attending a workshop on the material contained in the instrument. Participants were asked to rate the items for relevance using a four point scale. A 1 was assigned to the item if the behavior was unrelated to effective teaching. A 2 rating was to be given if the behavior on occasion might have a positive effect on a student's learning. A rating of 3 indicated the behavior would more than likely have a positive effect on student's learning. A 4 rating indicated the behavior was definitely related to effective teaching.

The average rating of scores for all 20 items by the workshop participants ranged from 3.15 to 4.0. The average score for one item across the 28 raters ranged from 3.03 to 3.96. The average interitem reliability was .86.

Population and Data Collection

Subjects in the study were 40 first year home economics teachers. Twenty four teachers were teaching in Iowa and graduated from Iowa State University. Sixteen teachers were teaching in South Dakota and graduated from South Dakota State University. All of the subjects were participants in the study by Gilbert (1974) and were chosen to meet the criteria
of the follow up study being investigated by Adams (1975).

During Fall Quarter, 1974, letters were sent to the 24 first year Iowa teachers and their school principals explaining the basic purpose of the study and requesting their cooperation. Copies of this correspondence are included in Appendix G. The 16 first year South Dakota teachers and their principals were contacted by the Dean of the College of Home Economics at South Dakota State University during January, 1975. All agreed to participate.

Enclosed with the letters to the first year teachers was a form requesting the teacher's class schedule and content being taught within the classes. From this form, three classes were selected for observation. A tentative visitation schedule was planned for each teacher to include a 10-15 minute preconference prior to the first of the three classes to be observed and a postconference of approximately 30 minutes during a preparation period or after school.

The teachers were mailed a second letter informing them of the scheduled visitation day, the classes to be observed and specific details of the research. A card was enclosed with the letter for the teacher to indicate if the date was acceptable. If the date was not acceptable she was asked to list a date when she could be contacted to arrange for another visitation day.
From the returned cards and telephone conversations a visitation schedule was prepared. The 24 Iowa teachers were visited between February 3 and March 5, 1975. The 16 South Dakota teachers were visited between March 6 and April 11, 1975. Each teacher was scheduled for a full day visit. Five days were not scheduled during the time period to allow for adverse weather conditions or illness.

Two weeks prior to the visit the teacher received a letter reminding her of the visitation date and classes to be observed. An additional purpose of this letter was to request that a copy be available on the visitation day of the lesson plan for the classes to be observed and a set of evaluation techniques that had been or were being used in the selected classes. The evaluation techniques included formal devices (tests, quizzes), rating scales, checklists, student assignments, projects and performance tests and informal devices such as verbal descriptions of observations, classroom questions or conversations with students.

The teachers were visited on the scheduled days by Adams and the investigator. Data were collected by the two judges using direct observation of classroom performance, written materials and conference questions. The school day spent with each teacher included a conference with the principal and/or superintendent whenever possible.
Data Analysis

As indicated by the objectives for this study, there were two major purposes. The first was to revise the observation device, Student Teacher Evaluation: Assessing the Teaching-Learning Process (STE). The second was to identify items that could reliably assess classroom teaching performance and discriminate among teachers in the classroom. A secondary objective of the study was to investigate the stability of teaching performance over classes.

Raw data from 400 instruments completed by the two trained judges were prepared for key punching by recording code numbers on each instrument and hand coding the class characteristics and school characteristics on the appropriate devices. Code numbers for each of the teachers were assigned to each of the devices, the Revised Student Teacher Evaluation: Assessing the Teaching-Learning Process (RSTE) and An Assessment Device for Human Relations Skills.

Revised student teacher evaluation

The first step in the analysis was to transform the raw data to normal deviates. This was done because 1) statistically normal deviates more clearly distinguish differences between items, and 2) it was believed that the intervals between the response values were not equal, i.e., an evaluator
who clearly sees a teacher's performance above average or below average on an item does so with greater certainty than one who uses values around the middle of the scale. The original numerical responses of 1 to 99 were, therefore, non-linearly transformed by using these numbers as if they were cumulative proportions. These cumulative proportions were referred to a cumulative standard normal curve table, and the corresponding normal deviates replaced the original responses. Using this transformation, a response of 1 becomes \(-2.33\); a response of 50, 0; and a response of 99, 2.33. When X's were included in the analyses they were coded as 50 and thus transformed to 0. The analyses performed were based on these normal deviates rather than the original 99 scale responses.

Two analyses of variance (ANOVA) were calculated for the Revised Student Teacher Evaluation instrument. The first analysis of variance was computed for the portion of the instrument used in the three classroom observations for the 40 first year teachers. The second analysis of variance was computed for the cumulative over classes portion of the RSTE device.

The two analyses of variance were computed for the 47 items in the instrument. From these analyses, the reliability of the judges in rating items for discriminating
between teachers was calculated.

The model upon which the first analysis (classroom observation) was based is:

\[ Y_{ijk} = \mu + T_i + J_j + TJ_{ij} + K_{k(i)} + \varepsilon_{ijk} \]

where:

- \( Y_{ijk} \) = transformed score assigned \( i \)th teacher by \( j \)th judge for the \( k \)th class observed
- \( i \) = 1,2,3...40 teachers
- \( j \) = 1,2 judges
- \( k \) = 1,2,3 classes within teachers
- \( \mu \) = overall mean
- \( T \) = teacher effect
- \( J \) = judge effect
- \( TJ \) = teacher by judge interaction effect
- \( K \) = classes within teachers effect
- \( \varepsilon_{ijk} \) = random error

The expected mean squares for the ANOV are shown in Table 2. The level of significance selected for testing was the .05 level.

The intraclass correlation coefficient \( (r_I) \) between judges for each item on the classroom observation device was calculated using the estimated components of variance obtained from an analysis of variance on the mean response of
Table 2. Expected mean squares of the analysis of variance design

<table>
<thead>
<tr>
<th>Source of variation^a</th>
<th>df</th>
<th>Expected values of mean squares</th>
<th>F ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teachers (T)</td>
<td>39</td>
<td>$\sigma^2 + 2\sigma^2_{K(T)} + 6\sigma^2_T$</td>
<td>$T/K(T)$</td>
</tr>
<tr>
<td>Judge (J)</td>
<td>1</td>
<td>$\sigma^2 + 3\sigma^2_{TxJ} + 12\sigma^2_J$</td>
<td>$J/TxJ$</td>
</tr>
<tr>
<td>Teacher by Judge (TxJ)</td>
<td>39</td>
<td>$\sigma^2 + 3\sigma^2_{TxJ}$</td>
<td>$TxJ/\epsilon$</td>
</tr>
<tr>
<td>Class within Teacher [K(T)]</td>
<td>80</td>
<td>$\sigma^2 + 2\sigma^2_{K(T)}$</td>
<td>$K(T)/\epsilon$</td>
</tr>
<tr>
<td>Error (\epsilon)</td>
<td>80</td>
<td>$\sigma^2$</td>
<td></td>
</tr>
</tbody>
</table>

^a Teachers and classes random, judges fixed.
each judge over the 3 classes observed for each teacher. The model used in this situation was:

\[ Y_{ij} = \mu + T_i + J_j + \varepsilon_{ij} \]

where:

- \( Y_{ij} \) = mean response over the 3 classes observed for the ith teacher by the jth judge
- \( i = 1,2,3...40 \) teachers
- \( j = 1,2 \) judges
- \( \mu \) = overall mean
- \( T \) = teacher effect
- \( J \) = judge effect
- \( \varepsilon_{ij} \) = error

The expected mean squares for this ANOV are shown in Table 3.

Table 3. Expected mean squares of the analysis of variance design

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>df</th>
<th>Expected values of mean squares</th>
<th>F ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teachers (T)</td>
<td>39</td>
<td>( \sigma^2 + 2\sigma_T^2 )</td>
<td>( T/\varepsilon )</td>
</tr>
<tr>
<td>Judges (J)</td>
<td>1</td>
<td>( \sigma^2 + 40\sigma_J^2 )</td>
<td>( J/\varepsilon )</td>
</tr>
<tr>
<td>Error (( \varepsilon ))</td>
<td>39</td>
<td>( \sigma^2 )</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Teachers random, judges fixed.
Intraclass correlation coefficients were calculated for each item. These coefficients represent the relationship between teacher variance (true variance) over teacher plus error variance (error variance). The formula used was (Winer, 1971, p. 286):

\[ r_I = \frac{\delta^2_T}{\delta^2_T + \delta^2} \]

where:
\[ \delta^2_T = \text{estimated component of variance for teachers} \]
\[ \delta^2 = \text{estimated component of variance for error} \]

Reliability of the mean of 2 judges was computed by the average inter-item procedure (Nunnally, 1967, p. 193):

\[ r_{jj} = \frac{2r_I}{1+r_I} \]

The model upon which the second analysis (cumulative assessment over classes) was based is:

\[ Y_{ij} = \mu + T_i + J_j + \epsilon_{ij} \]

where:
\( Y_{ij} \) = transformed score assigned ith teacher by jth judge for an item
\( i = 1,2,3...40 \) teachers
\( j = 1,2 \) judges
\( u \) = overall mean
\( T \) = teacher effect
\( J \) = judge effect
\( \epsilon_{ij} \) = error

The analysis of variance format is shown in Table 4. The level of significance selected for testing was the .05 level.

Table 4. Expected mean squares of the analysis of variance design

<table>
<thead>
<tr>
<th>Source of variation( ^a )</th>
<th>df</th>
<th>Expected values of mean squares</th>
<th>F ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teachers (T)</td>
<td>39</td>
<td>( \sigma^2 + 2\sigma_T^2 )</td>
<td>( T/\epsilon )</td>
</tr>
<tr>
<td>Judges (J)</td>
<td>1</td>
<td>( \sigma^2 + 40\sigma_J^2 )</td>
<td>( J/\epsilon )</td>
</tr>
<tr>
<td>Error (( \epsilon ))</td>
<td>39</td>
<td>( \sigma^2 )</td>
<td></td>
</tr>
</tbody>
</table>

\( ^a \)Teachers random, judges fixed.
Intraclass correlation coefficients ($r_I$) were computed as:

$$r_I = \frac{\delta^2_T}{\delta^2_T + \delta^2}$$

where:

$\delta^2_T = \text{teacher variance}$

$\delta^2 = \text{error variance}$

Reliability of the mean of 2 judges was computed by the average inter-item procedure.

$$r_{jj} = \frac{2r_I}{1+r_I}$$

where:

$r_{jj} = \text{reliability of the item}$

$r_I = \text{intraclass correlation}$

Establishment of cluster composition

The final step in the analysis of the RSTE instrument was to run a 47x47 correlation matrix of the combined items of the classroom performance and cumulative over classes sections. The correlation matrix was calculated from the mean rating of the two judges to the items. The matrix was examined to identify items with correlations of at least .60 which
seemed to be forming clusters.

The items in each cluster were recorded and examined for similarity of content. Those items that were not included in any of the clusters were studied for resemblance to the content of the defined clusters; where similarities were found, the correlation matrix was examined to see if the item correlated with the other items above .40. If it did, the item was added to the cluster. In general, this method introduced items into a cluster which, while similar in content, tended to have correlations below .60 with the other items. Final choice of which items would remain in a cluster was determined by maximizing the reliability of the cluster.

The following formula was used to determine the reliability of each cluster:

$$r_{jj} = \frac{n \bar{r}}{1+(n-1)\bar{r}}$$

where:

- $r_{jj}$ = reliability of the cluster
- $n$ = number of items in a cluster
- $\bar{r}$ = mean of correlations for a cluster

A 3x3 cluster intercorrelation matrix was computed using the cluster scores. The matrix was inspected to determine the independence of the clusters.
An assessment device for human relations skills

A chi-square analysis was used for the human relations device. The technique was employed to identify if there was a pattern in how the judges assigned scores. This pattern can be a pattern of either agreement or disagreement. Therefore, it was necessary to inspect each item for which the test was significant to determine the cause of significance, e.g., agreement or disagreement.

A 6x6 contingency table, illustrated in Table 5, was set up and frequencies calculated as though judges were independent. When the table was examined, it was found there

Table 5. Frequency distribution of tallies from human relations skills device

```
<table>
<thead>
<tr>
<th>JUDGE 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUMBER OF TEACHERS</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>X</td>
</tr>
</tbody>
</table>
```
were not enough teachers to meet the requirement of an expected frequency of 5 in each cell. To avoid an invalid use of the chi-square technique, all the diagonal cells were added together and all the off diagonal cells added together to collapse the table into a 1x2 table as shown in Table 6.

Table 6. Reduced frequency distribution of tallies from human relations skills device

<table>
<thead>
<tr>
<th>NUMBER OF TEACHERS</th>
<th>AGREE</th>
<th>DISAGREE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIAGONAL</td>
<td>OFF DIAGONAL</td>
</tr>
<tr>
<td>f₁ᵃ</td>
<td>f₂</td>
<td>F₁ᵇ</td>
</tr>
</tbody>
</table>

\(^{af} = \text{number of teachers observed.}\)

\(^{bf} = \text{number of teachers expected.}\)

The formula used to obtain a chi-square with 1 df was (Snedecor and Cochran, 1967, p. 21):

\[ \chi^2 = \frac{(f₁⁻F₁)^2}{F₁} + \frac{(f₂⁻F₂)^2}{F₂} \]

where:
\[ f_1 = \text{frequency of teacher ratings on which judges agreed (sum over diagonal cells in contingency table)}. \]

\[ f_2 = \text{frequency of teacher ratings on which judges disagreed (sum over off diagonal cells in contingency table)}. \]

\[ F_1 = \text{expected frequency of teacher ratings on which judges agreed}. \]

\[ F_2 = \text{expected frequency of teacher ratings on which judges disagreed}. \]

A significance level of .05 was chosen for testing the independent assignment of scores by judges.
FINDINGS AND DISCUSSION

The first section of this chapter includes reports of results concerning the items of the Revised Student Teacher Evaluation (RSTE) instrument. The section is divided into three subsections; the 25 classroom performance items are discussed first, the 22 cumulative over classes items second, and the results of clustering the teaching performance items last.

The second section focuses on An Assessment Device for Human Relations Skills (HRS). The discussion reports the usefulness of the instrument as a teaching and/or evaluation device.

Assessment of RSTE Items

The most promising classroom performance items

The most promising classroom performance items of the RSTE were determined by inspection of the F ratios associated with the teacher (T) effect, judge (J) effect, and the teacher by judge interaction (TxJ) effect. If the teacher effect was significant, then the judges were observing differences among teachers. Effects which differentiated between teachers were desired as an indication that the items were not assessing competencies that were essentially mastered by all teachers.

If the judge effect was non-significant, the two judges
were agreeing on the rating of the item. Further, if the teacher by judge interaction effect was non-significant, the judges were agreeing on how to evaluate teachers, i.e., one judge was not rating one teacher high on a competency while the other judge was rating the same teacher at a lower level. Since not only agreement between the judges in observing differences among teachers but also agreement in the rating of the item was desired, items which possessed these characteristics were designated as the most promising items. The items in this group were divided into three subgroups and are reported in Table 7.

The classes within teacher K(T) effect is also reported in the classroom performance tables and discussed at the end of each subgroup. If the K(T) effect is significant this suggests that a teacher's performance varies over classes.

Subgroup I contains those items which possess a non-significant teacher by judge interaction effect. These items contain highly significant teacher differences and a non-significant judge effect. These results are interpreted as indicating the performance of the teacher differed on these items; the judges agreed on the rating of the item and ordered the items in the same way; and that judges were not rating teachers differently. Non-significant teacher by
judge interaction is illustrated graphically in Figure 1. Observe that judges essentially rated each teacher at the same rating level as evidenced by the fact that rating lines are essentially horizontal and that there are few intercepting ratings. Further interpretation of Figure 1 is found in the discussion of Subgroup III.

In the first group of items intraclass correlation coefficients (\( r_1 \)) range from .74 to .95. The \( r_1 \) represents the proportion of variance attributable to teachers. Inspection of the total Table 7 shows the coefficients range from .54 to .98 with 14 of the 17 coefficients above .83. These items also possess the highest reliabilities because the reliability estimates between the two judges were based on \( r_1 \).

These items are useful for discriminating among teachers by trained judges and are observable in a single class period. Discriminating among teachers is often desired in CBE programs and in teacher evaluation. However, if teachers are to be held accountable for a certain level of mastery performance, more work will be necessary to help all teachers meet the criterion.

The \( K(T) \) effect for items in this group is highly significant which provides evidence that even though judges are agreeing on what is happening in a class the teacher's performance varies over classes. The F ratio for this
Table 7. F ratios of analysis of variance components of most promising items in classroom performance section of RSTE

<table>
<thead>
<tr>
<th>Item</th>
<th>F ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T^a$</td>
</tr>
<tr>
<td>4. Learning opportunities suited to student's needs and/or interest.</td>
<td>3.46**</td>
</tr>
<tr>
<td>6. ...indicated the objectives of the lesson.</td>
<td>2.18**</td>
</tr>
<tr>
<td>8. ...selected appropriate teaching techniques and strategies for the situation.</td>
<td>3.49**</td>
</tr>
<tr>
<td>10. ...used example(s) or illustration(s) for conveying ideas during the lesson.</td>
<td>2.41**</td>
</tr>
<tr>
<td>13. ...effectively used teaching materials and/or instructional materials.</td>
<td>3.46**</td>
</tr>
<tr>
<td>17. ...created a positive, success-oriented learning environment.</td>
<td>5.25**</td>
</tr>
<tr>
<td>19. ...helped the students recognize the relationship of the lesson to previous learning or experience.</td>
<td>1.80**</td>
</tr>
<tr>
<td>20. ...assisted the students in synthesizing, summarizing and/or drawing conclusions.</td>
<td>4.24**</td>
</tr>
<tr>
<td>25. ...recognizes individual student's needs.</td>
<td>1.83**</td>
</tr>
<tr>
<td>Subgroup II</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>---</td>
</tr>
<tr>
<td>1. Lesson planned to concentrate primarily on</td>
<td>2.40**</td>
</tr>
<tr>
<td>one main idea.</td>
<td></td>
</tr>
<tr>
<td>2. Objectives written in behavioral terms.</td>
<td>9.21**</td>
</tr>
<tr>
<td>7. ...conducted the lesson in a logical</td>
<td>2.55**</td>
</tr>
<tr>
<td>sequence.</td>
<td></td>
</tr>
<tr>
<td>22. ...treats the students with respect.</td>
<td>3.72**</td>
</tr>
<tr>
<td>Subgroup III</td>
<td></td>
</tr>
<tr>
<td>3. Objectives based on student's needs and/or</td>
<td>2.55**</td>
</tr>
<tr>
<td>interest.</td>
<td></td>
</tr>
<tr>
<td>5. Lesson provided an opportunity for students</td>
<td>4.01**</td>
</tr>
<tr>
<td>to participate actively in several ways.</td>
<td></td>
</tr>
<tr>
<td>14. ...effectively used levels of questions to</td>
<td>3.02**</td>
</tr>
<tr>
<td>evoke thinking beyond the level of recall.</td>
<td></td>
</tr>
<tr>
<td>16. ...encouraged open-ended inquiry and/or</td>
<td>3.15**</td>
</tr>
<tr>
<td>discussion when consistent with instructional</td>
<td></td>
</tr>
<tr>
<td>goals.</td>
<td></td>
</tr>
</tbody>
</table>

\[a\] Degrees of freedom for F are 39, 80. Table values for F are 1.58 at 5 percent and 1.88 at 1 percent.

\[b\] Degrees of freedom for F are 1, 39. Table values for F are 4.09 at 5 percent and 7.35 at 1 percent.

\[c\] Degrees of freedom for F are 80, 80. Table values for F are 1.48 at 5 percent and 1.74 at 1 percent.

\[d\] 30 of the 120 observations were rated X by both judges and 8 observations rated X by one judge.

* Significant at \( P < 0.05 \).

** Significant at \( P < 0.01 \).
Figure 1. Illustration of non-significant teacher by judge interaction (TxJ) effect: Item 6
effect ranges from 3.13 to 15.34 indicating some behaviors vary more than others across classes. This finding has substantial implications for teacher education research because it indicates that there are very real differences between teacher performance in any given classroom. Prior to this time, the investigator is unaware of any study that had addressed itself to the differences in teacher performance between classes.

Subgroup II items in Table 7 include those with a significant F ratio for the teacher by judge interaction effect. These items also possess a highly significant teacher effect and non-significant judge effect. The significant teacher effect and non-significant judge effect are desirable, however, the significant teacher by judge interaction means the judges are not ordering the items in the same way.

The K(T) effect is highly significant for items in this group, with the F ratios ranging from 6.58 to 31.53. This finding is interpreted that the judges are observing differences in the teacher's performance across classes, and that these differences vary greatly depending on the item.

Subgroup III items in Table 7 include those with a highly significant teacher by judge interaction effect. These items possess a highly significant teacher effect and non-significant judge effect as did the items in subgroups
I and II in this table. Although judges were observing differences among teachers and agreeing on the rating of the item, they were not agreeing on how to evaluate the teachers on these items. These items are least fair to the teacher as her evaluation depends on the judge.

Item 3 in this group appears to be different from the other items. Analysis of the competency represented by the item suggests that this presents a difficult item to observe in one class period because of the difficulty of outside observers knowing what are the true needs of the learner prior to a classroom session. The intraclass correlation coefficient of .54 as well as smaller F for the teacher effect suggests that the item was not distinguishing well between teachers. Further, since the interaction effect is twice as large, it suggests judges were differing markedly on ratings given teachers.

The K(T) effect for items in Subgroup III was also highly significant, with three of the four items displaying large differences in teacher's performance over classes. Inspection of the content for items 3, 14, and 16 reveals these are rather common behaviors and yet teachers varied in these behaviors over classes.

Highly significant teacher by judge interaction is illustrated in Figure 2. Contrasting the graphic illustration
Figure 2. Illustration of highly significant teacher by judge interaction (TxJ) effect: Item 16
of teacher by judge interaction found in Figure 1, Item 6 and Figure 2, Item 16, it is apparent in Figure 2 that judges were using different rating scores for this item as evidenced by the large number of sloping lines connecting judge ratings versus the relatively large number of horizontal lines caused by similar ratings in Figure 1.

There are two reasons why the items differ in the interaction effect. The first reason is there are larger differences in how the two judges score the teachers on Item 16 than on Item 6. For example, on Item 6, non-significant TxJ interaction, the largest discrepancy between the judges scores occurs when judge 1 scores a teacher -.50 on the transformed scale and judge 2 scores that teacher a +.50 on the same scale, a difference of 1.00. However, on Item 16, significant TxJ interaction, the bottom line shows judge 1 scores a teacher -2.33, while judge 2 scores the same teacher -.67, a difference of 1.66. A second reason why the items differ in terms of significant teacher by judge interaction is that the error terms for testing the interaction term differ in size. The mean square error for Item 6 is .158, and for Item 16 is .091, thus making the denominator in the F value for testing significance of the teacher by judge interaction larger for Item 6 that for Item 16, therefore, the F value is less likely to be significant.
The least promising classroom performance items

The six least promising items were those that had significant F ratios for T, J and TxJ effects. While the highly significant T effect indicates the judges were observing differences among teachers and was designated desirable, the significant J effect and the significant TxJ effect suggests these items appeared to be the least desirable for observation of teaching behaviors. These effects denote the judges were not agreeing on the rating of the item and were ordering the teachers differently. The items are reported in Table 8.

The items were inspected to see if some rationale could be developed as to why the items did not function. Items 9, 23 and 24 contain phrases such as "interested in the subject," "open positive rapport," and "communicates effectively;" even with the aid of descriptors these are items on which it is difficult for observers to reach consensus. Further inspection of the means of the three observations per judge per item revealed more than half of the teachers were performing above average on Items 9, 11, 23 and 24. There appeared to be very little variation in the scores the teachers received.

Item 18 was the only item in the group to receive any X's by the judges. An X rating was used by the observers to
<table>
<thead>
<tr>
<th>Item</th>
<th>F ratios</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9. ...was interested in the subject being taught</td>
<td>2.99**</td>
<td>7.53**</td>
<td>1.76*</td>
<td>4.31**</td>
<td>.76</td>
<td>.86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. ...clarified concepts and concerns as needed during the lesson.</td>
<td>2.51**</td>
<td>9.39**</td>
<td>2.00**</td>
<td>8.10**</td>
<td>.82</td>
<td>.90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. ...pointed out reasons and relationships concerning facts.</td>
<td>2.65**</td>
<td>10.85**</td>
<td>2.77**</td>
<td>13.46**</td>
<td>.85</td>
<td>.92</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18. ...arranged physical environment conducive to teaching technique being used.</td>
<td>2.21**</td>
<td>26.17**</td>
<td>1.73*</td>
<td>4.81**</td>
<td>.71</td>
<td>.83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23. ...maintains an open positive rapport with the students.</td>
<td>6.65**</td>
<td>5.20*</td>
<td>1.48</td>
<td>4.10**</td>
<td>.89</td>
<td>.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24. ...communicates effectively at levels appropriate for the students.</td>
<td>3.97**</td>
<td>6.70*</td>
<td>2.19**</td>
<td>5.23**</td>
<td>.80</td>
<td>.89</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a Degrees of freedom for F are 39, 80. Table values for F are 1.58 at 5 percent and 1.88 at 1 percent.
*b Degrees of freedom for F are 1, 39. Table values for F are 4.09 at 5 percent and 7.35 at 1 percent.
*c Degrees of freedom for F are 80, 80. Table values for F are 1.48 at 5 percent and 1.74 at 1 percent.

*Significant at \( P < 0.05 \).
**Significant at \( P < 0.01 \).
indicate that there was no opportunity to exhibit the behavior. A complete listing of X's received for each item in the instrument can be found in Appendix J.

The K(T) effect for the least promising items are all highly significant, denoting the teacher's performance differed over classes. Item 12 is the most significant with an F ratio of 13.46.

The remaining classroom performance items

The two remaining items are reported in Table 9. These items have a non-significant T effect, J effect and TxJ effect. Hence while the analysis suggests the items do not discriminate among teachers, the non-significant J and TxJ effects indicate judges agree on how to rate and evaluate the teachers. An inspection of the content and rating of the items indicated Item 12 does not occur often enough to be observable during a single class period or over three class periods with 100 of the 120 observations receiving X by both judges. Item 21 mean ratings revealed over half of the ratings on the transformed scale suggested that teachers were performing above average or at an above average mastery level.

The K(T) effect for the remaining items is highly significant. These large differences could be attributed to the fact these behaviors would not be commonly exhibited
Table 9. F ratios of analysis of variance components of remaining items in classroom performance section of RSTE

<table>
<thead>
<tr>
<th>Item</th>
<th>F ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>12.</td>
<td>.81</td>
</tr>
<tr>
<td>21.</td>
<td>1.27</td>
</tr>
</tbody>
</table>

<sup>a</sup>Degrees of freedom for F are 39, 80. Table values for F are 1.58 at 5 percent and 1.88 at 1 percent.

<sup>b</sup>Degrees of freedom for F are 1, 39. Table values for F are 4.09 at 5 percent and 7.35 at 1 percent.

<sup>c</sup>Degrees of freedom for F are 80, 80. Table values for F are 1.48 at 5 percent and 1.74 at 1 percent.

<sup>d</sup>100 of the 120 observations were rated X by both judges and 4 observations were rated X by one judge.

**Significant at P < 0.01.
during every class period, but may occur only once during the three observations or over many observations, thus producing highly significant class within teacher differences. Perhaps these items would be more effectively observed in the same class period over time.

The most promising cumulative over classes items

The most promising cumulative over classes items of the RSTE were determined by inspection of the F ratios associated with the teacher (T) effect, and judge (J) effect (Table 10). This analysis lacks the effects of T×J and K(T) as each judge observed each teacher only once on these items and thus there were not enough degrees of freedom for the other F tests.

If the teacher effect was significant, then items were discriminating among teachers. If the judge effect was non-significant, the two judges agreed on the rating of the item. Since not only differences among teachers but also agreement in the rating of the item by judges was desired, items which possessed these characteristics were designated as the most promising items. These items are arranged in descending order by size of the F ratio for the T effect and reported in Table 10.

All of the most promising items possess a highly significant teacher effect and non-significant judge effect.
Table 10. F ratios of analysis of variance components of most promising items in cumulative over classes section of RSTE

<table>
<thead>
<tr>
<th>Item</th>
<th>F ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T^a$</td>
</tr>
<tr>
<td></td>
<td>$J^b$</td>
</tr>
<tr>
<td></td>
<td>$r_I$</td>
</tr>
<tr>
<td></td>
<td>$r_{jj}$</td>
</tr>
<tr>
<td>18. ...summarizes scores obtained from evaluation devices to determine range and mean of scores.</td>
<td>64.21**</td>
</tr>
<tr>
<td>14. ...has used at least 2 techniques for assessing psychomotor skills in formal evaluation.</td>
<td>56.30**</td>
</tr>
<tr>
<td>19. ...performs an item analysis on appropriate devices.</td>
<td>51.43**</td>
</tr>
<tr>
<td>22. C ...uses techniques of assigning marks or grades consistent with school policy.</td>
<td>37.74**</td>
</tr>
<tr>
<td>8. ...bases evaluation of pupil growth on the degree of accomplishment toward expected behaviors specified in the instructional objectives.</td>
<td>36.15**</td>
</tr>
<tr>
<td>1. ...encouraged the students to describe or show how the learning affects them personally.</td>
<td>18.60**</td>
</tr>
<tr>
<td>17. ...test directions are clearly stated and easy to follow.</td>
<td>17.91**</td>
</tr>
<tr>
<td>16. ...constructs well defined test items which reflect the principles of item writing.</td>
<td>16.05**</td>
</tr>
<tr>
<td>15. ...uses a table of specifications in planning tests.</td>
<td>15.77**</td>
</tr>
<tr>
<td>5. ...relates to each student.</td>
<td>15.67**</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>2.</td>
<td>...talks with students during or after class about their concerns.</td>
</tr>
<tr>
<td>20.</td>
<td>...uses results to assess the educative process.</td>
</tr>
<tr>
<td>3.</td>
<td>...handles discipline problems in a positive, consistent manner.</td>
</tr>
<tr>
<td>7.</td>
<td>...strives for improvement through professional growth activities.</td>
</tr>
<tr>
<td>21.</td>
<td>...returns scores and reviews and interprets the device with pupils to provide them with feedback.</td>
</tr>
<tr>
<td>4.</td>
<td>...uses criticism, either positive or negative, in a way that contributes to the student's growth.</td>
</tr>
<tr>
<td>11.</td>
<td>...in evaluation of performance activities, assesses process and product independently.</td>
</tr>
<tr>
<td>6.</td>
<td>...works cooperatively with other teachers and the administration.</td>
</tr>
</tbody>
</table>

^a Degrees of freedom for F are 39, 39. Table values for F are 1.70 at 5 percent and 2.13 at 1 percent.

^b Degrees of freedom for F are 1, 39. Table values for F are 4.09 at 5 percent and 7.35 at 1 percent.

^c 24 of the 80 observations were rated X by both judges.

** Significant at P < 0.01.
These results are interpreted to mean that judges were observing differences among teachers and agreeing on the rating of the item. A highly significant teacher effect and non-significant judge effect is illustrated in Figure 3. Item 14 is used in Figure 3 to graphically illustrate significant teacher effect and non-significant judge effect. The significant teacher effect can be seen by the wide range of scores along the diagonal line for the different teachers. The non-significant judge effect can be seen by the same range of scores assigned by the two judges, from -2.33 to 2.00 for both.

The intraclass correlation coefficients $r_I$ for these items range from .67 to .96 with 16 of the 18 coefficients above .83. These coefficients provide a measure of the precise agreement on scores. Items in this table also possess reliabilities above .80. These reliabilities are also high as the reliability estimates between the two judges were based on $r_I$.

Inspection of the content of the first four items with extremely large F ratios for the teacher effect indicate the competencies delineated by the items are all rather specific and therefore lend themselves to observation by raters. Eight of the items, 6, 7, 11, 15, 19, 20, 21, and 22 were assessed through conference questions due to the difficulty
Figure 3. Illustration of ratings which resulted in a highly significant teacher (T) effect and non-significant judge (J) effect: Item 14
of observing these behaviors even over several class periods. Thus many of these assessments were determined by what the teacher reported she did rather than actual observations. A listing of conference questions can be found in Appendix D.

The least promising cumulative over classes items

The three least promising items were those that had significant F ratios for both the T and J effects and are reported in Table 11. The highly significant T effect indicates the judges were observing differences among teachers. However, the significant J effect denotes the judges were not agreeing on the ratings of the item, thus these items appear to be the least desirable for observation of teaching behavior and the most unfair to the teacher as her evaluation depends upon the judge.

Ratings from Item 13 are graphically illustrated in Figure 4 showing a significant teacher effect and significant judge effect. The significant teacher effect can be seen by the wide range of scores along the diagonal line for the different teachers. The significant judge effect can be seen by the different range of scores assigned by the two judges, from -2.33 to 1.67 for judge 1 and from -2.33 to 2.00 for judge 2.

Inspection of the content of the items in this group suggest that even cumulatively they are difficult to observe
Table 11. F ratios of analysis of variance components of least promising items in cumulative over classes section of RSTE

<table>
<thead>
<tr>
<th>Item</th>
<th>F ratios</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T^a$</td>
<td>$J^b$</td>
<td>$r_i$</td>
<td>$r_{jj}$</td>
</tr>
<tr>
<td>9. ...encourage pupil's self-evaluation.</td>
<td>27.31**</td>
<td>9.19**</td>
<td>.92</td>
<td>.96</td>
</tr>
<tr>
<td>10. ...in evaluation of performance activities, clearly identifies to student whether process or product is being evaluated.</td>
<td>39.45**</td>
<td>4.24*</td>
<td>.95</td>
<td>.97</td>
</tr>
<tr>
<td>13. ...has used at least 3 techniques for assessing affective growth in formal evaluation.</td>
<td>172.08**</td>
<td>12.38**</td>
<td>.98</td>
<td>.99</td>
</tr>
</tbody>
</table>

$^a$Degrees of freedom for F are 39, 39. Table values for F are 1.70 at 5 percent and 2.13 at 1 percent.

$^b$Degrees of freedom for F are 1, 39. Table values for F are 4.09 at 5 percent and 7.35 at 1 percent.

*Significant at $P < 0.05$.

**Significant at $P < 0.01$. 
Figure 4. Illustration of ratings which resulted in a highly significant teacher (T) and judge (J) effect: Item 13
in the classroom. Items 9 and 10 were assessed through conference questions (Appendix D) and classroom observation when the behavior appeared. Assessing affective growth in formal evaluation, Item 13, was absent from some teachers' evaluation packets. An explanation for this could be the teachers were requested to include in the packet examples of devices used in the classes to be observed. In some cases these three types and/or levels of classes tended to accommodate cognitive and psychomotor devices rather than affective devices.

The remaining cumulative over classes item

The one remaining item is reported in Table 12. Item 12 has a non-significant teacher effect and significant judge effect. The non-significant teacher effect is interpreted that judges did not observe differences among teachers. The significant judge effect indicates that judges were not agreeing on the rating of the item.

An inspection of the data for Item 12, uses five techniques for assessing cognitive growth, revealed that almost every teacher was performing clearly above average. Thus, it appears teachers are functioning at a mastery level in using a variety of cognitive techniques. The significant judge effect for the item appears to be caused by one observation in which one judge did not observe five techniques, thus rating the teacher 0, while the other judge observed
Table 12. F ratios of analysis of variance components of remaining item in cumulative over classes section of RSTE

<table>
<thead>
<tr>
<th>Item</th>
<th>( F_{r} )</th>
<th>( F_{j} )</th>
<th>( r_{I} )</th>
<th>( r_{jj} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>12. ...has used at least 5 techniques for assessing cognitive growth in formal evaluation.</td>
<td>1.45</td>
<td>4.92*</td>
<td>.18</td>
<td>.31</td>
</tr>
</tbody>
</table>

*Significant at \( P < 0.05 \).

\(^{a}\) Degrees of freedom for \( F \) are 39, 39. Table values for \( F \) are 1.70 at 5 percent and 2.13 at 1 percent.

\(^{b}\) Degrees of freedom for \( F \) are 1, 39. Table values for \( F \) are 4.09 at 5 percent and 7.35 at 1 percent.
five techniques and rated the teacher clearly above average. Figure 5 uses Item 12 to illustrate graphically non-significant teacher effect and significant judge effect. The non-significant teacher effect can be seen by the short range of scores assigned the teachers along the diagonal. The significant judge effect is observed by the difference in range of scores assigned by the two judges, from -2.33 to 2.00 for judge 1 and from 1.00 to 2.00 for judge 2.

Inspection was made of the most promising items of the STE (Appendix H) with those in the RSTE (Table 7 and 10) to ascertain the effectiveness of the revision of the instrument. Five items (3, 5, 19, 23, and 47) of the STE discriminated between teachers and were reliably judged (Gilbert, 1974). This study found 17 items which could be observed during a single class period, discriminated between teachers and be reliably judged; 1, 2, 3, 4, 5, 6, 7, 8, 10, 13, 14, 16, 17, 19, 20, 22, and 25. Eighteen items could be observed over three class periods, discriminated between teachers and be reliably judged; 1, 2, 3, 4, 5, 6, 7, 8, 11, 14, 15, 16, 17, 18, 19, 20, 21, and 22. These findings evince the improvement of the STE by the revision.

To summarize, 17 of the items from the classroom performance section, Part A, and 18 from the cumulative over classes section, Part B, of the RSTE were found to be most promising in evaluating teaching performance. All 35 of the
Figure 5. Illustration of ratings which resulted in a non-significant teacher (T) effect and significant judge (J) effect: Item 12
items in both sections of the device discriminated between teachers and the two judges agreed on the item rating.

Ten of the 17 items in the classroom performance section, Part A, indicated that judges were also agreeing on how to evaluate teachers. The nine items identified as least promising of the combined sections appear to be difficult for judges to agree on the rating of the item as well as how to evaluate teachers. Since all of these items discriminate among teachers and thus do possess potential, the items and descriptors should be refined and further investigations made. These findings suggest the revision of the STE has improved the instrument.

Cluster Composition of RSTE Items

All items of the RSTE, classroom performance and cumulative over classes sections, were intercorrelated. The correlation matrix was calculated from the mean rating of the two judges on each item. Inspection of the correlations \( \geq .60 \) and item content produced three major clusters. Eight items which were isolates or possessed low correlations were not included in the clusters and are listed in Appendix I.

The three clusters varied in size from 11 to 14 items and were named according to the content of the items. The clusters were labeled as follows:
Cluster J, instructional modes

Cluster K, learning environment

Cluster L, evaluation methods

Cluster J, instructional modes:  The high end of this cluster describes a teacher who bases objectives on student's needs, uses a variety of teaching strategies and effectively uses instructional illustrations and materials. The reliability of the cluster is .94. The intercorrelation coefficients between the items are shown in Table 13.

The items numbered as they were in the original observational device include:

1. Lesson planned to concentrate primarily on one main idea.
3. Objectives based on student's needs and/or interest.
4. Learning opportunities suited to student's needs and/or interest.
5. Lesson provided an opportunity for students to participate actively in several ways.
7. . . . conducted the lesson in a logical sequence.
8. . . . selected appropriate teaching techniques and strategies for the situation.
9. . . . was interested in the subject being taught.
10. . . . used example(s) or illustration(s) for conveying ideas during the lesson.
11. . . . clarified concepts and concerns as needed during the lesson.
12. . . . followed through with her plans and yet remained flexible enough to adjust as needs became evident.
13. . . . effectively used teaching materials and/or instructional materials.
14. . . . effectively used levels of questions to evoke thinking beyond the level of recall.
15. . . . pointed out reasons and relationships concerning facts.
16. . . . encouraged open-ended inquiry and/or discussion when consistent with instructional goals.
Table 13. Intercorrelations\(^a\) among items in Cluster J, instructional modes

<table>
<thead>
<tr>
<th>Item</th>
<th>1(^b)</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>37</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>46</td>
<td>90</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>52</td>
<td>68</td>
<td>70</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>70</td>
<td>50</td>
<td>59</td>
<td>57</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>51</td>
<td>63</td>
<td>77</td>
<td>67</td>
<td>68</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>35</td>
<td>73</td>
<td>72</td>
<td>72</td>
<td>53</td>
<td>72</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>49</td>
<td>71</td>
<td>72</td>
<td>77</td>
<td>58</td>
<td>73</td>
<td>72</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>44</td>
<td>74</td>
<td>67</td>
<td>62</td>
<td>57</td>
<td>60</td>
<td>66</td>
<td>73</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>55</td>
<td>30</td>
<td>38</td>
<td>26</td>
<td>33</td>
<td>27</td>
<td>30</td>
<td>37</td>
<td>32</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>46</td>
<td>66</td>
<td>75</td>
<td>67</td>
<td>71</td>
<td>83</td>
<td>64</td>
<td>74</td>
<td>58</td>
<td>18</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>18</td>
<td>51</td>
<td>45</td>
<td>56</td>
<td>40</td>
<td>41</td>
<td>68</td>
<td>45</td>
<td>41</td>
<td>14</td>
<td>48</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>50</td>
<td>56</td>
<td>61</td>
<td>64</td>
<td>74</td>
<td>64</td>
<td>67</td>
<td>74</td>
<td>68</td>
<td>26</td>
<td>68</td>
<td>41</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>56</td>
<td>50</td>
<td>56</td>
<td>34</td>
<td>35</td>
<td>73</td>
<td>52</td>
<td>44</td>
<td>22</td>
<td>48</td>
<td>71</td>
<td>48</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) The decimal point has been omitted from the correlations on this table.

\(^b\) Items from classroom performance section of RSTE.
Cluster K, learning environment: The high end of this cluster describes a positive success oriented learning environment in which students are treated with respect and assisted in synthesizing, summarizing and drawing conclusions. The teacher in this environment communicates at an appropriate level for the students, provides opportunities for them to apply their learning in more than one way and uses criticism in a way that contributes to the student's growth. The reliability for Cluster K is .92. Table 14 shows the intercorrelation coefficients for Cluster K.

The items numbered as they were in the original observational device include:

17. ...created a positive, success-oriented learning environment.
18. ...arranged physical environment conductive to teaching technique being used.
19. ...helped the students recognize the relationship of the lesson to previous learning or experience.
20. ...assisted the students in synthesizing, summarizing and/or drawing conclusions.
21. ...provided an opportunity for students to apply their learnings in more than one way.
22. ...treats the students with respect.
23. ...maintains an open positive rapport with the students.
24. ...communicates effectively at levels appropriate for the students.
25. ...recognizes individual student's needs.
4. ...uses criticism, either positive or negative, in a way that contributes to the student's growth.
5. ...relates to each student.
Table 14. Intercorrelations$^a$ among items in Cluster K, learning environment

<table>
<thead>
<tr>
<th>Item</th>
<th>17$^b$</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>4$^c$</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>44</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>51</td>
<td>15</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>79</td>
<td>39</td>
<td>52</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>39</td>
<td>53</td>
<td>14</td>
<td>29</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>78</td>
<td>40</td>
<td>35</td>
<td>66</td>
<td>37</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>83</td>
<td>40</td>
<td>40</td>
<td>68</td>
<td>35</td>
<td>93</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>70</td>
<td>36</td>
<td>35</td>
<td>63</td>
<td>51</td>
<td>91</td>
<td>85</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>42</td>
<td>33</td>
<td>37</td>
<td>33</td>
<td>45</td>
<td>53</td>
<td>61</td>
<td>55</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>57</td>
<td>47</td>
<td>37</td>
<td>55</td>
<td>49</td>
<td>49</td>
<td>45</td>
<td>44</td>
<td>47</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>69</td>
<td>42</td>
<td>30</td>
<td>56</td>
<td>34</td>
<td>76</td>
<td>82</td>
<td>69</td>
<td>64</td>
<td>48</td>
<td>-</td>
</tr>
</tbody>
</table>

$^a$ The decimal point has been omitted from the correlations on this table.

$^b$ Items 17-25 are from classroom performance section of RSTE.

$^c$ Items 4-5 are from cumulative over classes section of RSTE.
Cluster L, evaluation methods: The high end of this cluster describes a teacher who incorporates a variety of evaluation techniques in the cognitive, psychomotor and affective domains and who utilizes sound analytical techniques to analyze the cognitive devices. The reliability of the cluster is .88 and the correlation matrix on which it is based is found in Table 15.

The items numbered as they were in the original observational device include:

8. ...bases evaluation of pupil growth on the degree of accomplishment toward expected behaviors specified in the instructional objectives.
9. ...encourages pupil's self-evaluation.
10. ...in evaluation of performance activities, clearly identifies to student whether process or product is being evaluated.
11. ...in evaluation of performance activities, assesses process and product independently.
12. ...has used at least 5 techniques for assessing cognitive growth in formal evaluation.
13. ...has used at least 3 techniques for assessing affective growth in formal evaluation.
14. ...has used at least 2 techniques for assessing psychomotor skills in formal evaluation.
15. ...uses a table of specifications in planning tests.
16. ...constructs well defined test items which reflect the principles of test writing.
17. ...test directions are clearly stated and easy to follow.
18. ...summarizes scores obtained from evaluation devices to determine range and mean of scores.
19. ...performs an item analysis on appropriate devices.
20. ...uses results to assess the educative process.
21. ...returns scores and reviews and interprets the device with pupils to provide them with feedback.
Table 15. Intercorrelations$^a$ among items in Cluster L, evaluation methods

<table>
<thead>
<tr>
<th>Item</th>
<th>8$^b$</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>28</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>49</td>
<td>50</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>44</td>
<td>34</td>
<td>64</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>11</td>
<td>19</td>
<td>16</td>
<td>19</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>23</td>
<td>2</td>
<td>30</td>
<td>26</td>
<td>48</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>38</td>
<td>22</td>
<td>31</td>
<td>53</td>
<td>34</td>
<td>31</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>31</td>
<td>32</td>
<td>35</td>
<td>39</td>
<td>-04</td>
<td>11</td>
<td>19</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>40</td>
<td>26</td>
<td>35</td>
<td>41</td>
<td>73</td>
<td>54</td>
<td>59</td>
<td>4</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>35</td>
<td>23</td>
<td>35</td>
<td>28</td>
<td>74</td>
<td>45</td>
<td>51</td>
<td>3</td>
<td>95</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>37</td>
<td>27</td>
<td>20</td>
<td>31</td>
<td>00</td>
<td>5</td>
<td>13</td>
<td>21</td>
<td>1</td>
<td>-04</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>49</td>
<td>15</td>
<td>36</td>
<td>43</td>
<td>30</td>
<td>17</td>
<td>50</td>
<td>24</td>
<td>42</td>
<td>40</td>
<td>20</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>80</td>
<td>25</td>
<td>62</td>
<td>54</td>
<td>17</td>
<td>22</td>
<td>48</td>
<td>32</td>
<td>45</td>
<td>45</td>
<td>34</td>
<td>76</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>64</td>
<td>44</td>
<td>42</td>
<td>52</td>
<td>27</td>
<td>15</td>
<td>57</td>
<td>36</td>
<td>46</td>
<td>39</td>
<td>56</td>
<td>53</td>
<td>63</td>
<td>-</td>
</tr>
</tbody>
</table>

$^a$The decimal point has been omitted from the correlations on this table.

$^b$Items from cumulative over classes section of RSTE.
The results of the cluster analysis look plausible for reducing the number of items needed for observation and for revision or adaption of the instrument for teacher evaluation. It should be noted however, that the analysis must be considered exploratory due to only 40 teachers in the study.

The relationship of the clusters as shown in Table 16 was studied by analyzing the amount of common variance ($r^2$)

<table>
<thead>
<tr>
<th>Cluster</th>
<th>J</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>J: Instructional modes</td>
<td>94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K: Learning environment</td>
<td>85</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>L: Evaluation method</td>
<td>45</td>
<td>36</td>
<td>88</td>
</tr>
</tbody>
</table>

\[ n = 40 \]

\[ \text{^aAll decimal points have been omitted.} \]

\[ \text{^bIntercorrelation coefficients reported in the off diagonals and reliability of the cluster are reported in the diagonal.} \]
between the clusters. It is apparent that the common variance between J and K is 72%; J and L, 20%, and K and L, 13%.

Interpretation of the degree of the common variance suggests that instructional modes and learning environment, Cluster J and K, are very dependent factors and one component contributes to the effectiveness of the other. Evaluation methods, Cluster L, has little common variance with either instructional modes, Cluster J, or learning environment, Cluster K. This finding suggests that evaluation methods is a relatively independent aspect to study in the analysis of teaching performance. Perhaps this assessment could occur at a different time or place from the direct observation of classroom performance.

While these clusters suggest the measuring of independent dimensions of teaching performance, it should not be inferred that any of the dimensions should be excluded without further analysis using a larger n.
Assessment of Human Relations Skills

The chi-square technique was used for analyzing An Assessment Device for Human Relations Skills (HRS). The decision to use the chi-square technique was made due to the use of the six point scale in observing teachers. The rating scale used for the HRS was X, 0, 1, 2, 3, 4. The chi-square for each competency of the 20 item device was inspected to determine significance. If a significantly large chi-square was observed it was interpreted that the number of teachers on which the judges agreed and/or disagreed differed from the number of teachers for which agreement and/or disagreement was expected due to chance, i.e., independent assignment of scores by judges to teachers. Further inspection of the frequency distribution of the chi-square table revealed whether the large chi-square occurred because the judges were agreeing more often or less often than expected by chance.

Items with significant chi-squares are divided into three subgroups according to the number of X's marked by the judges in rating the teachers. An X rating was used to indicate that there was no opportunity to exhibit the behavior. Since it was desired to look at the instrument as either an evaluation or teaching device, it was decided to report chi-squares computed without the X rating. Had the findings been
reported with X's included for the items, all 20 items would have been significant (see Appendix K).

Subgroup I contains significant chi-squares for 40 observations for each of two judges as shown in Table 17. Neither judge gave an X in rating any of the six skills. During the observation of three class periods, teachers demonstrated and raters were able to observe the skills. Inspection of the large chi-squares disclosed observed frequencies (f) in the diagonal cells were larger than their respective expected values (F). Half of the paired observations for each of the six skills in Subgroup I were reported as a one or two on the rating scale which indicated teachers are performing the skill with difficulty to some degree of competence. However, as with any skill, proficiency often comes with practice.

Subgroup II items are those in which 25 percent or less of the ratings were X. The paired observations range from 31 to 37. Eight human relations skills from the areas of active listening and giving positive and negative feedback are identified in the subgroup. The significant chi-squares range from 5.40 to 45.31. Since judges are able to observe and rate 75 percent of the observations in Subgroup II, it is believed that these items could be used in evaluating teachers and/or measuring their competence.
Table 17. Significant chi-squares for human relations skills behaviors

<table>
<thead>
<tr>
<th>Item</th>
<th>Pairs of observations</th>
<th>Percent X ratings</th>
<th>$f^a/p^b$</th>
<th>Chi-square$^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subgroup I - Items which did not receive X ratings</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Creating an Open, Positive Atmosphere:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOES THE PERSON...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. ...give messages of acceptance non-verbally; i.e., eye contact, posture, tone of voice, etc.?</td>
<td>40</td>
<td>0</td>
<td>26/15</td>
<td>12.77**</td>
</tr>
<tr>
<td>2. ...send simultaneous verbal and non-verbal messages that agree?</td>
<td>40</td>
<td>0</td>
<td>26/17</td>
<td>8.53**</td>
</tr>
<tr>
<td>Active Listening:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOES THE PERSON...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. ...relate and listen to every student?</td>
<td>40</td>
<td>0</td>
<td>27/18</td>
<td>8.76**</td>
</tr>
<tr>
<td>5. ...develop the student's ideas instead of translating them into her own answers and/or ways of expression?</td>
<td>40</td>
<td>0</td>
<td>25/17</td>
<td>7.45**</td>
</tr>
<tr>
<td>Giving Positive Feedback:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOES THE PERSON...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. ...vary the way in which she gives positive reinforcement.</td>
<td>40</td>
<td>0</td>
<td>25/18</td>
<td>4.69*</td>
</tr>
<tr>
<td>12. ...give feedback which refers to the individual's behavior, not the individual himself?</td>
<td>40</td>
<td>0</td>
<td>26/16</td>
<td>9.74**</td>
</tr>
<tr>
<td><strong>Subgroup II - Items which received less than 25% X ratings</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active Listening:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOES THE PERSON...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. ...reflect the student's thinking to him so that he can clarify his own ideas?</td>
<td>31</td>
<td>22.5</td>
<td>19/15</td>
<td>5.40*</td>
</tr>
<tr>
<td>8. ...leave the responsibility for a decision with the student rather than give the answer or solve a problem for the student?</td>
<td>37</td>
<td>7.5</td>
<td>30/13</td>
<td>34.27**</td>
</tr>
</tbody>
</table>
9. ...recognize and respond productively to a student's non-verbal cues which indicate he is reacting in a way that could affect his learning? 31

Giving Positive Feedback:

DOES THE PERSON...

11. ...follow her positive reinforcement with an explanation of why the praise was given; i.e., "that is a good answer because it..."? 37

Giving Negative Feedback:

DOES THE PERSON...

13. ...give negative feedback to another when it would contribute to his growth? 34

14. ...after pointing out errors, follow up with constructive help? 34

15. ...keep the reason for a negative message and the objective for giving it clear and open (as opposed to the use of insinuation, sarcasm or implication)? 32

16. ...in giving a negative message, communicate both by words and tone of voice a message of "information" rather than one of "should" or "ought"? A message of "what", not "why"? 31

\[ \chi^2 = \text{number of teachers observed on the diagonal.} \]

\[ \chi^2_f = \text{number of teachers expected on the diagonal.} \]

\[ \chi^2 = \text{Table value for chi-square with 1 degree of freedom is 3.84 at 5 percent and 6.63 at 1 percent.} \]

* Significant at \( P < 0.05 \).

** Significant at \( P < 0.01 \).
### Table 17 (Continued)

<table>
<thead>
<tr>
<th>Item</th>
<th>Pairs of observations</th>
<th>Percent X ratings</th>
<th>$\frac{a}{b}$</th>
<th>Chi-square $c$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subgroup III - Items which received more than 25% X ratings</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Giving Negative Feedback:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does the person...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17. ...use tentative expressions rather than absolutes when giving an opinion statement; (&quot;it seems&quot; vs. &quot;it is&quot;; &quot;some&quot;, &quot;more&quot;, &quot;perhaps&quot;)?</td>
<td>27</td>
<td>32.5</td>
<td>21/10</td>
<td>22.14**</td>
</tr>
<tr>
<td><strong>Receiving Negative Feedback:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does the person...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19. ...respond to negative feedback from others by asking for clarification or by checking for accuracy of meaning; (rather than responding with defensive remarks which, in effect, cancel out the information received in the negative message)?</td>
<td>12</td>
<td>70</td>
<td>12/6</td>
<td>13.35**</td>
</tr>
<tr>
<td>20. ...respond cognitively to the facts in a negative message from another instead of emotionally to an imagined motive; i.e., respond to the &quot;what&quot; rather than the &quot;why&quot;?</td>
<td>14</td>
<td>65</td>
<td>12/4</td>
<td>20.97**</td>
</tr>
</tbody>
</table>
Item 11 received the largest number of 0 ratings. A 0 rating indicated the behavior was appropriate but not attempted. Eleven of the 37 paired ratings were rated 0, which suggests teachers may need more help with the skill of following positive reinforcement with an explanation of why the praise was given.

Subgroup III contains three items, 17, 19, and 20, in which more than 25 percent of the paired ratings were X. These items also received paired 0 ratings of 2, 7 and 4 respectively. Although judges were agreeing on the observations, with such a large number of 0 and X ratings it appears these skills are difficult to observe during three class periods. Inspection of the items reveals the three are in the areas of giving and receiving negative feedback which are more difficult skills to develop and not likely to appear as often as some of the other human relations skills.

Non-significant chi-squares for items in the HRS device are reported in Table 18. Skills in the areas of creating an open, positive atmosphere, active listening and receiving negative feedback are represented in the three items. These skills were difficult for judges to observe and rate. There was not a pattern in how the judges assigned scores.

A closer inspection of the items reveals Item 3 did not
Table 18. Non-significant chi-squares for human relations skills behaviors

<table>
<thead>
<tr>
<th>Item</th>
<th>Pairs of observations</th>
<th>Percent of X ratings</th>
<th>(^a/F)</th>
<th>(^b/F)</th>
<th>Chi-square (^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creating An Open, Positive Atmosphere:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOES THE PERSON...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. ...make the negative messages she sends a response to the present moment?</td>
<td>33</td>
<td>17.5</td>
<td>19/16</td>
<td>1.11</td>
<td></td>
</tr>
<tr>
<td>Active Listening:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOES THE PERSON...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. ...maintain accurate understanding between speakers and listeners by checking for meaning?</td>
<td>30</td>
<td>25</td>
<td>14/10</td>
<td>3.30</td>
<td></td>
</tr>
<tr>
<td>Receiving Negative Feedback:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOES THE PERSON...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18. ...use reasoned agreement (consensus) to resolve a conflict among others or between another and herself, (vs. power or majority rule)?</td>
<td>8</td>
<td>80</td>
<td>6/5</td>
<td>3.44</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) F = number of teachers observed on the diagonal.

\(^b\) F = number of teachers expected on the diagonal.

\(^c\) Table value for chi-square with 1 degree of freedom is 3.84 at 5 percent and 6.63 at 1 percent.
receive any 0 ratings; however, Item 7 received four and Item 18 received 6 paired 0 ratings. Item 18 also received the largest percentage of X ratings which suggests the skill is demonstrated so seldom that evaluation is difficult. The skill itself is valuable and perhaps more opportunities should be provided in a teaching situation among peers to develop the skill.

In summary, 14 items of the HRS device can be observed in a relatively short observation period, 3 class periods, and judges can evaluate these skills. These items represent four of the five areas of the device. The only area not represented is receiving negative feedback. Six of the 14 items, which are 1, 2, 4, 5, 10, and 12 appear to occur frequently enough to perhaps be included in a single class period observation device.

The investigator cautions that the items should not be judged good or bad according to the X ratings, but rather suggests this information be used in working with teachers to help them feel more comfortable in using some of these skills in the classroom. The device does appear to have potential as an evaluative and/or teaching device, but, the instrument should be subjected to further statistical analysis.
SUMMARY

The reorientation of research on teaching from a study of teacher qualities to the observation of actual instruction in the classroom and the competency based education and certification movement have caused teacher educators to look closely at the effect of teacher preparation on teaching performance. One of the issues in competency-based teacher education programs is the problem of assessment, first in terms of defining student teacher competencies and, second, in terms of assessment measures.

Home economics teacher educators throughout the United States have combined and focused their efforts during the past decade toward defining competencies necessary for beginning home economics teachers. The Home Economics Education Department at Iowa State University has been involved in a study of program and curriculum revision. Components of the departmental study have been refining objectives of the home economics teacher education program; developing an evaluation matrix for the undergraduate teacher education program; developing an evaluation device for one of the five major competencies of the model, the teaching-learning process; and testing the device.

This study was part of the departmental study and was designed to revise and validate the teaching performance device.
used to assess competencies of home economics education student teachers related to the teaching-learning process.

Participants in the study were 40 first year home economics teachers from Iowa State University and South Dakota State University who had participated in Gilbert's 1974 study.

The instrument to be revised in this study contained items selected from instruments that had previously been found to discriminate between teachers, and items developed by a home economics education graduate student. Revision of the instrument involved taking into consideration results and recommendations from Gilbert's 1974 study. Items were reviewed for clarity, objectivity and observability and item descriptors established for some items. The instrument was reduced from 50 items to 47 items and divided into two parts. Part A, Classroom Performance, contains 25 items to be observed during a single class period. Part B, Cumulative Assessment Over Classes, includes 22 items to be observed over several class periods. A 99 point scale was used to record responses on the Revised Student Teacher Evaluation (RSTE) device.

The instrument was further expanded to include a human relations skills device, An Assessment Device for Human Relations Skills (HRS), developed by a department faculty member.
Data were collected during a school day visit to each of the 40 first year teachers. The scheduled visit included observation of three class periods selected by the observers, a 15 minute pre-conference, and a 30 minute post-conference. Five instruments were completed for each visit, they were: three of Part A and one of Part B of the RSTE instrument; and one of the HRS instrument.

The purposes of the analysis of data were to validate the instruments used in the study, to identify items that judges could reliably use to assess classroom teaching performance, and to identify items that differentiated between teachers. Analyses of variance, correlations and chi-squares were used in the data analyses.

The first step in the analysis of the RSTE was to transform the raw data to normal deviates. Using the latter, two analyses of variance (ANOV) were computed. The first analysis of variance was computed for Part A, Classroom Performance, and the second for Part B, Cumulative Assessment Over Classes. From the analyses, intraclass correlations were computed on each item. Reliability estimates of the judges in rating items that discriminated between teachers were based on the intraclass correlation coefficients and were projected for the two judges on each item using the average inter-item procedure.
A 47x47 correlation matrix was computed and examined for items which correlated and were similar in content. Clusters were formed and a 3x3 cluster intercorrelation matrix was computed and inspected to determine the independence of the clusters.

The analysis of variance results for Part A, Classroom Performance section were inspected. F ratios of the analysis of variance components for the teacher (T) effect, judge (J) effect, and teacher by judge (TxJ) effect were examined and revealed that of the 25 items, judges observed differences among teachers and agreed on the rating of 17 items. Additionally, on nine of the 17 items judges also agreed on how to evaluate the teachers, i.e., one judge was not rating one teacher high on a competency while the other judge was rating the same teacher at a lower level. Of the remaining items, six discriminated among teachers, however judges did not agree on the rating of the item nor on how to evaluate the teachers, and two did not discriminate among teachers, however, judges did agree on how to rate and evaluate the teachers. Of these latter two items, one did not occur often enough to be rated and the other revealed the majority of teachers rated slightly above average mastery level.

The significant class within teacher \([K(T)]\) effect for all
of the 25 items of Part A, suggests teacher performance does vary over classes. Therefore, assessment of classroom performance should be made over more than one class period and in more than one subject matter area. This finding has implications for teacher education research because it indicates that there are very real differences between teacher performance from one class to another.

The F ratios for Part B, Cumulative Over Classes section, were examined for the teacher (T) effect and judge (J) effect only, due to the single observation for each teacher on these items. Inspection revealed 18 of the 22 items were discriminating among teachers. The judges observed differences among teachers and agreed on the rating of the item. Of the remaining four items, two need more detailed descriptors, one item appears to be at the clearly above average mastery level and the other reflected a task one-third of the teachers were not performing.

The final analysis of the RSTE was to intercorrelate all items from section A and section B of the instrument. Inspection of the correlations and item content produced three major clusters. The clusters were labeled Cluster J, Instructional modes; Cluster K, Learning environment; Cluster L, Evaluation methods. The clusters were intercorrelated to analyze the common variance between the clusters. Results
suggest that Cluster J and K are dependent factors and one component contributes to the effectiveness of the other. Cluster L has little common variance with the other two clusters and appears to be a relatively independent aspect to study in the analysis of teaching, indicating that assessment of these tasks could occur at a time or place other than during the classroom performance observation.

In summary of the RSTE device, 35 of the 47 items discriminate between teachers and trained judges are able to observe the behavior and agree on the item rating. These items are the most promising in evaluating teaching performance. Although instructional modes, Cluster J, and learning environment, Cluster K have 72% common variance they are still relatively independent clusters. Therefore, it is reasonable to assess them as separate factors in a classroom setting and over classes. Evaluation methods, Cluster L, appears to be relatively independent suggesting these items could be assessed at the end of student teaching.

The chi-square technique was used for analyzing the 20 item Human Relations Skills device because of the 6 point rating scale employed. The instrument had been developed as a teaching and/or evaluation device and data were desired as to how many of these skills could be observed in
the classroom. The items were rated on a six point scale, which included an X rating that was used to indicate no opportunity to exhibit the behavior. Of the significant chi-square items, six received no X ratings, indicating over the three observed class periods, teachers demonstrated the skill and judges were able to observe and agree on the rating. Eight of the items were able to be observed 75% of the time, and three were observable approximately 25% of the time, however, when these skills were demonstrated the judges agreed on the rating. Only three items in the instrument did not receive significant chi-square ratings indicating judges were not able to agree on the rating of the skill. These three items also received X ratings, with one item receiving X ratings on 80% of the observations.

In summary of the HRS device, six of the items appear to be observable in a short observation period and perhaps could be included in a single class period observation device. It does appear that the majority of the skills are observable in the classroom. It is recommended however that the instrument be subjected to further statistical analysis.

Of the 47 items in the Revised Student Teacher Evaluation (RSTE), 35 items were identified that judges could reliably use to assess teaching performance and which discriminated among teachers. Analysis of An Assessment Device for Human Relations Skills (HRS), suggests that six items
are observable in the classroom setting. The device does appear to have potential as an evaluative and/or teaching device, but should be tested further to see at what levels teachers are able to perform these skills and if these skills vary over subject matter areas.

The revised device therefore appears highly useful for the evaluation of student teachers provided more training sessions are given cooperating teachers and college supervisors in the use of the device. Such a device is extremely useful as it provides data on a student's performance in a realistic rather than a simulated setting, a goal with which educational researchers have long been concerned.

The nine items identified as least promising of the combined sections of the RSTE appear to be difficult for judges to agree on the rating of the item as well as how to evaluate teachers. Since all of these items discriminate among teachers and thus do possess potential, the items and descriptors should be refined and further investigations made. These findings suggest that the revision of the STE has improved the instrument.


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


ACKNOWLEDGMENTS

As we journey through life, some of us are fortunate to interact with people who help us to grow. This author is fortunate and indebted to several people for helping her to grow professionally and personally. She is also aware of the numerous people who through support and assistance made this study possible. Among those to whom the author wishes to express appreciation are:

Dr. Ruth Hughes, advisor and major professor, for her confidence, support, and encouragement, and her example as a member of the profession;

Dr. Alyce Fanslow for her competent guidance, encouragement, and generous allotment of time, as well as serving as a committee member;

Dr. Helen Lebaron Hilton, Dr. Mary Heltsley, and Dr. Margaret Liston for their support and their willingness to serve as committee members;

Mrs. Julia Anderson and Dr. Harold Dilts for serving as committee members during the final oral examination;

Dr. Leroy Wolins and Dr. Holly Fuchs for their help with the statistical analysis involved in the study, and their patience in giving explanations of the process;

Bud Meador for the programming assistance he gave with the analysis of the data;
Leola Adams for her sharing and reacting to ideas during the planning of the research, as well as her cooperation and perseverance during the 2 1/2 winter months we traveled Iowa and South Dakota collecting data;

First year teachers in Iowa and South Dakota for their willingness to participate in the study and for their patience in having two observers spend an entire day;

Dean Ardyce Gilbert for her willingness to have South Dakota teachers participate in the study and to make the initial contact, as well as her warm hospitality while the researchers were in South Dakota and financial support for the mileage and airfare accrued while traveling in South Dakota;

College of Home Economics for the Home Economics Research Institute funds and the Graduate Dean's office for funds to conduct the research;

Weber State College for providing financial support through a sabbatical leave during the 1974-75 school year;

Dr. Helen LeBaron Hilton for her provision of space and comfort in which the study was written;

Friends and department faculty members for their constant support, interest and encouragement prior to and during graduate study;

Family, especially John, for assuming the roles of both
mother and father, and two very special children, Stephanie and Steven, for granting their mother a leave of absence for a year to complete this study.
APPENDIX A: HOME ECONOMICS TEACHER EDUCATION MODEL
HOME ECONOMICS TEACHER EDUCATION

I. PHILOSOPHY
II. PROGRAMS
III. LEARNERS
IV. ENVIRONMENT
V. TEACHING-LEARNING PROCESS

UNDERGRADUATE HOME ECONOMICS TEACHER EDUCATION PROGRAM

PROSPECTIVE TEACHER

BEGINNING TEACHER

CONTINUING EDUCATION
HOME ECONOMICS TEACHER EDUCATION

PART V. TEACHING-LEARNING PROCESS

IV. ENVIRONMENT

III. LEARNERS

II. PROGRAMS

I. PHILOSOPHY

TEACHER

PUPIL

BODY OF KNOWLEDGE

BEHAVIORS

A. PROGRAM PLANNING
B. OBJECTIVES
C. CONCEPTS GENERALIZATIONS
D. LEARNING EXPERIENCES
E. EVALUATION TECHNIQUES
F. RELATIONSHIP SKILLS
APPENDIX B: STUDENT TEACHER EVALUATION (STE)

AND EVALUATOR'S GUIDE
STUDENT TEACHER EVALUATION

Assessing the Teaching-Learning Process:  Part B, C, D, E, F

Area of H. E. __________ Name: __________________
Grade Level ________ Teaching Center: __________
Evaluator: __________ Time: 4 wk ___ 6 wk ___ 8 wk ___

Directions: This instrument will help assess the degree of achievement exhibited by the student teacher at 4 week, 6 week, and 8 week intervals. Please respond to each of the items in the following manner:

If you are certain the teacher was clearly below average in accomplishing the task, place a 1 in the space provided.

If you are certain the teacher was clearly above average in accomplishing the task place a 99 in the space provided.

If you are uncertain the teacher accomplished the task or if there was no opportunity to accomplish it, place a 50 in the space provided.

A score from 1-49 indicates the degree to which the teacher displayed to you below average performance in accomplishing the task.

A score from 51-99 indicates the degree to which the teacher displayed to you above average performance in accomplishing the task.

Feel free to use any number from 1-99 that best reflects your opinion. Please respond to every statement. The general scale is shown below.

clearly below average uncertain clearly above average

1 1 1 1 1 1 50 1 1 1 1 1 1 1 99

CLASSROOM PERFORMANCE

THE TEACHER:

Lesson Selection: (refer to lesson plans)

1....planned the lesson to concentrate primarily on one main idea.

2....planned objectives and activities that were suited to the student's needs and interests.
3. selected appropriate teaching techniques and strategies for the situation.
4. created a positive, success-oriented learning environment.
5. indicated the objectives of the lesson and their importance to the students.
6. helped the students recognize the relationship of the lesson to previous learning or experience.

Presentation of the Lesson:
7. conducted the lesson smoothly so that it was easy for the students to follow.
8. was sincerely interested in the subject being taught.
9. used meaningful examples or illustrations for conveying ideas during the lesson.
10. effectively used a variety of teaching materials and instructional materials.
11. effectively used a variety of levels of questions to evoke thinking beyond the level of recall.
12. emphasized reasons and relationships concerning the facts.
13. clarified concepts and concerns as needed during the lesson. (vocabulary, definitions, etc.)
14. encouraged open-ended inquiry and discussion when consistent with the instructional goals.
15. used questions to elicit thinking and student response consistent with the instructional goals.
16. followed through with her plans and yet remained flexible enough to adjust as needs became evident.
17. was creative and flexible in guiding the learning process.
18. assisted the students in synthesizing, summarizing and drawing conclusions.

Student Participation:
19. provided an opportunity for the students to participate actively and/or to apply their learnings in different ways. (verbal response, non-verbal response)
20. encouraged the students to describe or show how the learning affects them personally.
21. Encouraged the students to make generalizations during or at the end of the lesson.

**RELATIONSHIP SKILLS**

22. Treats the students with respect.

23. Maintains an open friendly rapport with the students.

24. Communicates effectively at levels appropriate to the preparedness of the students.

25. Supports and accepts each student as he is regardless of race, sex, nationality or learning potential.

26. Recognizes the individual's personal needs.

27. Can explain things so students are able to understand.

28. Is willing to talk with students during or after class about any problems which might be bothering them.

29. Tries to find things that students are "good at" instead of things they are "poor at".

30. Works well with other teachers and the administration.

31. Handles his/her own discipline problems; is firm but friendly, consistent in policy and self-confident in management of pupils.

32. Uses positive, encouraging and supportive criticism, rather than discouragement, blame or shame.

**EVALUATION SKILLS**

**Planning for Evaluation:**

33. Bases evaluation of pupil growth on the degree of accomplishment toward expected behaviors specified in the instructional objectives.

34. Encourages pupil's self-evaluation in both specific and informal ways.

35. In evaluation of performance activities, clearly identifies whether process or product is being evaluated, and keeps measures on these two aspects independent during scoring.

36. Has used at least five types of cognitive measures of student growth in formal and informal evaluation.

37. Has used at least three types of affective measures of student growth.

38. Has used at least two evaluation devices of psychomotor skills.
...in independent and/or group learning projects, evaluates objectives on the basis of each individual's ability to reach objectives stated before the project was begun.

Planning the device

...uses a table of specifications in planning tests.
...constructs well defined test items which reflect the principles of item writing.

Presentation

...takes precautions so that the pupils do not have an opportunity to receive improper aids during the evaluation process.

Use and interpretation

...summarizes scores obtained from evaluation devices to determine range and mean of scores and to estimate good and poor test items.
...uses results of evaluation to determine several aspects of the educative process, not only for assigning grades.
...returns scores and reviews and interprets the device with pupils to provide them with feedback as quickly as possible.
...uses techniques of assigning marks or grades consistent with philosophy of the school.

MANAGEMENT AND PROFESSIONALISM

...is well prepared for class.
...initiates responsibilities for physical environment conducive to teaching technique being used.
...searches for ideas, techniques and procedures in developing a teaching style.
...strives for improvement through positive participation in professional growth activities.
To maintain conciseness in form of this instrument, Assessing the Teaching-Learning Process, items have been described as simply as possible. Hence it was felt that this guide may be of assistance to evaluators in providing further explanation for items which may need elaboration.

The first section of the accompanying instrument "classroom performance, (items 1-22), is to be used to assess the competence of the student teacher in a specific teaching situation. The evaluator should have a copy of the lesson plan before the teacher begins teaching the class.

The items in the remaining sections may or may not be observed in the same classroom situation. Note: a score of 50 indicates "no opportunity to accomplish" or "uncertain the teacher accomplished the task." This is important for statistical reasons.

Clarification of specific items:

1. "One main idea" can be interpreted to include ideas closely related to the main idea."

Items 36, 37, 38

Items 36, 37, and 38 are cumulative items. It is necessary that the student teacher reach the minimum number stated in each item at some time during her student teaching. This may occur entirely within one evaluation period, or it may take two or three evaluation periods before she has had opportunities to reach the minimum stated.

To aid you in types of measures of student growth, lists for each item are provided here.

Item 36  Cognitive measures of student growth

<table>
<thead>
<tr>
<th>alternative answer</th>
<th>checklists</th>
</tr>
</thead>
<tbody>
<tr>
<td>alternative answer with corrections</td>
<td>games</td>
</tr>
<tr>
<td>completion</td>
<td>incomplete stories</td>
</tr>
<tr>
<td>essay</td>
<td>observation</td>
</tr>
<tr>
<td>matching</td>
<td>role playing</td>
</tr>
<tr>
<td>multiple choice</td>
<td></td>
</tr>
<tr>
<td>oral response</td>
<td></td>
</tr>
</tbody>
</table>
Item 37  Affective measures of student growth

anecdotal records  participation charts
checklists  rating scale
incomplete sentences  role playing
incomplete stories  self inventory
logs  sociometric diagram

Item 38  Psychomotor measures of student growth

Checklists
lab observation
performance test
rating scales
work sample test

Item 40

A table of specifications should include the content covered, and the level of behavior expected for the content. This would reflect the objectives of the unit. For each cell of the table, a relative weight should be indicated. These weights suggest the relative importance of each objective to the total test measurement.

Item 41

The following statements are included in the principles of item writing.

- items should be clear and concisely written at a reading level easily understood by pupils taking the tests

- items should avoid tricky and ambiguous wording, double negatives, and sly details which tend to confuse pupils

- terms such as "always", "never", "all", "none", "some", "few", "many", should be avoided where possible

- independent response items should be explicitly stated and qualified so that the intent of the question is evident

- multiple choice distractors should be equally appealing to pupils who do not know the correct answer

- true-false items should not be partially true or partially false

- matching tests should not have more than twelve alternatives

- choices in an item should begin in the same grammatical form

- the entire item should be visible on the page, and not split onto two pages
APPENDIX C: REVISED STUDENT TEACHER EVALUATION (RSTE) AND EVALUATOR'S GUIDE

Part A
Teacher   Observer   Class Observation Number

CLASS CHARACTERISTICS

<table>
<thead>
<tr>
<th>Type of Class</th>
<th>Class Enrollment</th>
<th>Grade Level of Class Members</th>
<th>Class Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>semester</td>
<td>girls only</td>
<td>7</td>
<td>1-10</td>
</tr>
<tr>
<td>comprehensive</td>
<td>boys only</td>
<td>8</td>
<td>11-20</td>
</tr>
<tr>
<td></td>
<td>coeducational</td>
<td>9</td>
<td>31-40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>over 40</td>
</tr>
</tbody>
</table>

Length of time class has met with teacher
       weeks

Through lesson plan and conference questions assess:

___ 1 Lesson planned to concentrate primarily on one main idea.
___ 2 Objectives written in behavioral terms.
___ 3 Objectives based on student's needs and/or interest.

Subject Matter Presented:

PRE CONFERENCE

___ 4 Learning opportunities suited to student's needs and/or interest.
___ 5 Lesson provided an opportunity for students to participate actively in several ways.

CLASSROOM PERFORMANCE

Presentation/Laboratory

___ 6 Indicated the objectives of the lesson.
___ 7 Conducted the lesson in a logical sequence.
___ 8 Selected appropriate teaching techniques and strategies for the situation.
___ 9 Was interested in the subject being taught.
___10 Used example(s) or illustration(s) for conveying ideas during the lesson.
___11 Clarified concepts and concerns as needed during the lesson.
___12 Followed through with her plans and yet remained flexible enough to adjust as needs became evident.
___13 Effectively used teaching materials and/or instructional materials.
___14 Effectively used levels of questions to evoke thinking beyond the level of recall.
___15 Pointed out reasons and relationships concerning facts.
___16 Encouraged open-ended inquiry and/or discussion when consistent with instructional goals.

Environment

___17 Created a positive, success-oriented learning environment.
___18 Arranged physical environment conducive to teaching technique being used.

Student Participation

___19 Helped the students recognize the relationship of the lesson to previous learning or experience.
___20 Assisted the students in synthesizing, summarizing and/or drawing conclusions.
___21 Provided an opportunity for students to apply their learnings in more than one way.

Relationship Skills

___22 Treats the students with respect.
___23 Maintains an open positive rapport with the students.
___24 Communicates effectively at levels appropriate for the students.
___25 Recognizes individual student's needs.
Part B
____ Teacher  ______ Observer

SCHOOL CHARACTERISTICS

Type of School

<table>
<thead>
<tr>
<th></th>
<th>K-12</th>
<th>9-12</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7-9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Enrollment of School

<table>
<thead>
<tr>
<th></th>
<th>less than 200</th>
<th>1001-1500</th>
</tr>
</thead>
<tbody>
<tr>
<td>201-500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>501-1000</td>
<td></td>
<td>over 2000</td>
</tr>
</tbody>
</table>

CUMULATIVE ASSESSMENT OVER CLASSES

Student Participation

1. Encouraged the students to describe or show how the learning affects them personally.

Relationship Skills

2. Talks with students during or after class about their concerns.

3. Handles discipline problems in positive, consistent manner.

4. Uses criticism, either positive or negative, in a way that contributes to the student's growth.

5. Relates to each student.

Personal Growth

6. Works cooperatively with other teachers and the administration.

7. Strives for improvement through professional growth activities.

EVALUATION

8. Bases evaluation of pupil growth on the degree of accomplishment toward expected behaviors specified in the instructional objectives.


10. In evaluation of performance activities, clearly identifies to student whether process or product is being evaluated.


12. Has used at least 5 techniques for assessing cognitive growth in formal evaluation.

13. Has used at least 3 techniques for assessing affective growth in formal evaluation.

14. Has used at least 2 techniques for assessing psychomotor skills in formal evaluation.

Planning the Device

15. Uses a table of specifications in planning tests.

16. Constructs well defined test items which reflect the principles of item writing.

17. Test directions are clearly stated and easy to follow.

Use and Interpretation

18. Summarizes scores obtained from evaluation devices to determine range and mean of scores.

19. Performs an item analysis on appropriate devices.

20. Uses results to assess the educative process.

21. Returns scores and reviews and interprets the device with pupils to provide them with feedback.

22. Uses techniques of assigning marks or grades consistent with school policy.
Evaluators Guide for
Revised Student Teacher Evaluation

Directions: This instrument is designed to assess professional competencies of the classroom teacher. In the space at the left of each item, place an appropriate rating for the behavior exhibited by the teacher.

If you are certain the teacher was clearly below average in accomplishing the task, place a 1 in the space provided.

If you are certain the teacher was clearly above average in accomplishing the task, place a 99 in the space provided.

If you are uncertain the teacher accomplished the task, place a 50 in the space provided.

If you are certain the teacher had no opportunity to accomplish the task, place an X in the space provided.

If you are certain the teacher did not accomplish the task when it was appropriate, place an 0 in the space provided.

A score from 1-49 indicates the degree to which the teacher displayed to you below average performance in accomplishing the task.

A score from 51-99 indicates the degree to which the teacher displayed to you above average performance in accomplishing the task.
APPENDIX D: RSTE ITEM DESCRIPTORS AND
CONFERENCE QUESTIONS
RSIE Item Descriptors

Part A:

Classroom Performance

Item 6. If students proceed without confusion with learning activities, it can be assumed that they understand the objectives.

Item 8. Appropriateness for the situation can be assessed by physical environment, subject matter being taught and student interest.

Item 10. The example tracked through the lesson being presented.

Item 13. Instructional materials were appropriate to learners, setting and content being presented.

Item 17. Positive, success-oriented environment can be assessed by organization, enthusiasm, interest and mannerism.

Item 21. Active participation by students can be assessed by their verbal responses, written work, and projects that occur during the lesson or as a result of the lesson.

Item 22. The teacher listens actively, acknowledges students' hands and accepts students' ideas through restatement and use.

Part B:

Cumulative Assessment

Item 5. Interactions initiated by students are recognized and are responded to with sensitivity and respect by the teacher.

Item 6. The teacher carries out routine duties promptly and accurately; the teacher is considerate of total school needs and shares professional competencies will colleagues.

Item 7. The teacher engages in studies and activities to improve professional competence.
RSTE Conference Questions

Questions asked to facilitate the rating of selected items.

Part A:

Classroom Performance

Item 4. What considerations had to be made in planning this lesson for these students?

Item 5. What did you do in this class the other days of the week.

Part B:

Cumulative Assessment

Item 6. Do you have administrative support for your program? What are faculty members' opinions towards this home economics program?

Item 7. Have you had the opportunity to attend any professional growth activities?

Item 8. Is the degree of accomplishment toward a stated objective considered in the overall evaluation of pupil growth?

Item 9. Do you encourage your students to evaluate themselves?

Item 10. When you evaluate performance activities do you separate process and product? Does the student know whether he is being evaluated for the product or the process of the product?

Item 15. In making up a test, how do you decide what percentage of questions to allocate to different topics? (What items to include and how many?)

Item 18. How do you report the results of your tests to students?

Item 19. Do you ever make a determination as to whether your tests are too easy or too difficult? How?
Item 20. How do you use the test results in making decisions about this program?

Item 22. What is the school policy for assigning grades? Are there restrictions?
Appendix E: Numerical Calculations for Preliminary Interrater Reliability Assessments

Numerically, the variance component for teachers for reliability estimate of the average correlation between two judges for a single item was calculated as follows:

\[
\sigma_T^2 = \sigma^2 + 2.20 \sigma_T^2 - 0.0780
\]

\[
\sigma_T^2 = \frac{40 \sigma_T^2}{\sigma_T^2} = 0.3361
\]

\[
\sigma_T^2 = 0.008
\]

The error variance for the reliability estimate of a single item was interpreted as \( \sigma_T^2 \) as in Table 1. The reliability coefficient calculation then became:

\[
r_{jj} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_T^2} = \frac{0.008}{0.008 + 0.078} = 0.093
\]

Numerically, the teacher variance component for the reliability estimate for the total score was calculated as follows:
\[ \sigma_T^2 = \sigma_\varepsilon^2 + 2\sigma_T^2 = .4141 \]

\[ -\sigma_\varepsilon^2 = \sigma_{T_{\text{XJ}}}^2 = .1143 \]

\[ 2\sigma_T^2 = .2998 \]

\[ \sigma_T^2 = .1499 = .150 \]

In this calculation the prefix due to items associated with the expected mean square for the teacher component of variance, i.e., 20, was omitted because for this estimate the numerical component due to item was not considered since the emphasis was on reliability across all items. The error variance was interpreted as \( \sigma_{T_{\text{XJ}}}^2 \) since it represents the error variance estimate across the total observational device, e.g., the sum of the 20 items in Part A of the device.

The reliability coefficient calculation then became:

\[ r_{jj} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_{T_{\text{XJ}}}^2} = \frac{.150}{.150 + .114} = \frac{.150}{.264} = .58 \]
APPENDIX F: AN ASSESSMENT DEVICE FOR HUMAN RELATIONS SKILLS (HRS)
AN ASSESSMENT DEVICE FOR HUMAN RELATIONS SKILLS

The instrument which follows is intended to be used as an aid in the identification and recording of a person's progress in the acquisition of the human relations skills listed. Over a period of time you may have several numerical ratings for each behavior.

**Instructions:** In the space at the right, place a numerical rating each time you observe the person performing that behavior. Use the following key to mark your response:

<table>
<thead>
<tr>
<th>Rating</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>not appropriate (no opportunity to exhibit behavior)</td>
</tr>
<tr>
<td>0</td>
<td>not attempted (appropriate but behavior not used)</td>
</tr>
<tr>
<td>1</td>
<td>attempted with difficulty</td>
</tr>
<tr>
<td>2</td>
<td>performed with some competence</td>
</tr>
<tr>
<td>3</td>
<td>performed with ease and proficiency</td>
</tr>
<tr>
<td>4</td>
<td>uncertain (rater uncertain behavior performed)</td>
</tr>
</tbody>
</table>

**BEHAVIOR BEING OBSERVED—Does the person...**

**CREATING AN OPEN, POSITIVE ATMOSPHERE:**

1. give messages of acceptance non-verbally; i.e., eye contact, posture, tone of voice, etc.?
2. send simultaneous verbal and non-verbal messages that agree?
3. make the negative messages she sends a response to the present moment?

**ACTIVE LISTENING:**

4. relate and listen to every student?
5. develop the student's ideas instead of translating them into her own answers and/or ways of expression?
6. reflect the student's thinking to him so that he can clarify his own ideas?
7. maintain accurate understanding between speakers and listeners by checking for meaning?
8. leave the responsibility for a decision with the student rather than give the answer or solve a problem for the student?
9. recognize and respond productively to a student's non-verbal cues which indicate he is reacting in a way that could affect his learning?

**GIVING POSITIVE FEEDBACK:**

10. vary the way in which she gives positive reinforcement?
11. follow her positive reinforcement with an explanation of why the praise was given; i.e., "that is a good answer because it..."?
12. give feedback which refers to the individual's behavior, not the individual himself?
13. give negative feedback to another when it would contribute to his growth?
14. after pointing out errors, follow up with constructive help?
15. keep the reason for a negative message and the objective for giving it clear and open (as opposed to the use of insinuation, sarcasm or implication)?
16. in giving a negative message, communicate both by words and tone of voice a message of "information" rather than one of "should" or "ought"? A message of "what", not "why"?
17. use tentative expressions rather than absolutes when giving an opinion statement; ("it seems" vs. "it is"; "some", "more", "perhaps")?
18. use reasoned agreement (consensus) to resolve a conflict among others or between another and herself, (vs. power or majority rule)?
19. respond to negative feedback from others by asking for clarification or by checking for accuracy of meaning; (rather than responding with defensive remarks which, in effect, cancel out the information received in the negative message)?
20. respond cognitively to the facts in a negative message from another instead of emotionally to an imagined motive; i.e., respond to the "what" rather than the "why"?

Copyright © 1974 by Elnora T. Huyck
APPENDIX G: CORRESPONDENCE
November 8, 1974

The faculty members of the Home Economics Education Department of Iowa State University are pleased that one of our graduates has joined your teaching staff. We hope that she is making a worthwhile contribution to your home economics program this year.

The interest that we have in our students and our concern for their continued professional growth does not end at the time of their graduation but continues as an inservice program. One part of this program provides that some of our faculty members will make visits to all of the home economics education graduates who are beginning teachers in Iowa. Through these visits we hope to assist and encourage these teachers.

We would also like to use these visits as an opportunity for the Home Economics Education Department at Iowa State University to collect some data related to our continuing effort to improve the teacher education program. The purpose of the research project in which these data will be used is to help us evaluate the effectiveness of the competency-based teacher education program. All information will be treated as confidential; the research project is being conducted by Dr. Ruth Hughes and Dr. Alyce Fanslow.

Participation in the study would involve having two faculty members spend the day with the home economics teacher. The faculty members would meet with the teacher in the morning prior to her first class for a 10 to 15 minute conference and after her last class or during her preparation period for a post conference. During the day we would like to observe some of her classes. The classes to be observed will be determined prior to the visit.
November 8, 1974
Page 2

Plans are being made at Iowa State for faculty members to visit from February through April, 1975. Enclosed is a copy of the letter written to your home economics teacher. She will be in contact with you to obtain your permission for her participation in the study.

We sincerely hope that your school will not only be willing for us to visit with the first year teacher but also that it will be possible for us to collect the designated data. If you will indicate your willingness on the card enclosed for your reply, we will begin to make our plans. If you have any questions, please call Dr. Alyce Fanslow collect at 515-294-5307.

Sincerely,

Ruth P. Hughes
Professor and Head
Home Economics Education

Alyce M. Fanslow
Associate Professor
Home Economics Education

Enclosures
(PLEASE CHECK)

___Yes  I am willing for you to visit our first-year teacher and for our school to participate in your research project evaluating the competency-based teacher education program.

___No   I would prefer that you did not visit.

Signed ________________________________
Principal

______________________________
School
Greetings to you from all of the Home Economics Education Faculty at Iowa State. We hope that you are having a challenging and rewarding experience in your first year of home economics teaching.

You recall from Miss Miller's previous letter that we would be contacting you concerning a visit. From February through April, 1975, Leola Adams and Colleen Caputo, two members of our faculty, will visit beginning teachers. Through this visit we hope to assist you in your home economics program, share some recent teaching aids and learn of the frustrations and pleasures in your teaching activities.

We would also like to use this visit as an opportunity to evaluate the competency-based home economics teacher education program by collecting data on our 1973-74 home economics education graduates. All data collected will be summarized for the group; data for each participating teacher will remain anonymous.

Our visit will involve a 10 to 15 minute pre-conference prior to your first class and a post-conference after your last class or during your preparation period. During the day we would like to observe some of your classes. The exact date of the visit and the classes to be observed will be established during January through further correspondence.

We hope that it will be possible for us to visit you and to collect the desired data. We have also contacted your principal requesting his permission for our visit. Please check with him concerning his willingness to have you participate. We encourage your participation in the proposed evaluation as we do need input from each of our teaching graduates in order to evaluate the teacher education program.
November 8, 1975

In order for us to make plans for our visit, would you please send us your schedule and an approximate idea of the subject matter areas you will be teaching from February through April, 1975. A form on which to provide this information along with a stamped, addressed envelope is enclosed to facilitate your reply. May we have your response by November 18?

We look forward to visiting you, sharing some ideas together, and learning about your experiences in your first teaching position. If you have any questions concerning the visit or the data to be collected, please call Dr. Alyce Fanslow collect at 515-294-5307.

Sincerely,

Ruth P. Hughes
Professor and Head
Home Economics Education

Alyce Fanslow
Associate Professor
Home Economics Education

Enclosure
Yes, I would like to have you visit and am willing to participate in the proposed evaluation of the home economics teacher education program. I have talked with my principal and determined that your proposed data collection is acceptable.

Would you give us some idea of the course title, type and grade level you will be teaching from February through April.

<table>
<thead>
<tr>
<th>Course Title</th>
<th>Semester/Comprehensive</th>
<th>Grade Level</th>
<th>*Tentative Content (if possible)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example:</td>
<td>Homemaking I</td>
<td>Comprehensive</td>
<td>9</td>
</tr>
</tbody>
</table>

Please indicate below your daily time schedule for the February through April period:

<table>
<thead>
<tr>
<th>TIME</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*If you do not have the content plans for second semester at this time, information will be obtained at a later date.
January 10, 1975

It is our hope that your holiday season was both stimulating and relaxing. We are looking forward to visiting with you concerning your experiences as a first year teacher. We would also like to express our appreciation for your willingness to participate in the research project designed to provide some data for helping us to evaluate our teacher education program.

In our first letter to you, we mentioned that two faculty members, Ms. Leola Adams and Mrs. Colleen Caputo, would be spending a day with you, having a 10 to 15 minute preconference prior to your first class and a post-conference of about 30 minutes during your preparation period or after school. If you have some questions related to your teaching, we would be pleased to discuss these with you also.

Upon inspecting your class schedule for February through April, the following classes have been selected for our research project:

The visit to your school has been scheduled for
Please return the enclosed card as verification of this date. If there are planned school activities for this date which will prohibit class meetings, will you please let us know on the enclosed card so that we can decide on another visitation date.

The classes we would like to observe should involve teacher-pupil discussions of approximately 30 minutes in length. The discussion should include an exchange of questions and ideas with the teacher leading the
discussion. Since we are not including the following in our definition of discussion, please do not include in the class session to be observed such activities as pupil buzz groups, pupil reports, extended role playing, or independent study. We would like to request that clothing construction and/or foods laboratory sessions be teacher-directed, i.e., you direct the learning activities such as illustrating, questioning, clarifying concepts, or laboratory demonstrations.

Sometime during the day spent with you, we would like to have a brief conference with your principal. Perhaps you can help to arrange an appointment around the classes we specifically would like to observe. If it is convenient with your schedule, you are welcome to attend this meeting.

We will be looking forward to visiting with you. Please return the enclosed card within three days so that we can make definite plans for our visit with you. Thank you.

Sincerely,

Alyce M. Fanslow, Ph.D.
Associate Professor

Ruth P. Hughes, Ph.D.
Professor and Head

Enclosure
(PLEASE CHECK)

Yes This date is acceptable. Classes are scheduled, will call collect if there are changes. 515-294-5307.

No This is not an acceptable date. Please call me at ____________ date

___________________ time

___________________ number to arrange for another visitation date.

Signed ________________________________  

____________________________ School
January 29, 1975

It is our hope that your teaching experiences have been both stimulating and rewarding. We are looking forward to visiting with you concerning these experiences as a first year teacher. We would also like to express our appreciation for your willingness to participate in the research project designed to provide some data for helping South Dakota State evaluate its home economics teacher education program.

In Dean Gilbert's letter to you, she mentioned that two graduate assistants, Ms. Leola Adams and Ms. Colleen Caputo, would be spending a day with you, having a 10 to 15 minute preconference prior to your first class and a post conference of about 30 minutes during your preparation period or after school. If you have some questions related to your teaching, we would be pleased to discuss these with you also.

Upon inspecting your class schedule for March through April, the following classes have been selected for our research project:

The visit to your school has been scheduled for February 10. We will be calling you on Monday morning February 10 to verify this date. If there are planned school activities for this date which will prohibit class meetings, will you please let us know at the time of our phone call so that we can decide on another visitation date. If you are unable to be reached for this call, please leave a message for us at the office.
The classes we would like to observe should involve teacher-pupil discussions of approximately 30 minutes in length. The discussion should include an exchange of questions and ideas with the teacher leading the discussion. Since we are not including the following in our definition of discussion, please do not include in the class session to be observed such activities as pupil buzz groups, pupil reports, extended role playing or independent study. We would like to request that clothing construction and/or foods laboratory sessions be teacher-directed, i.e., you direct the learning activities such as illustrating, questioning, clarifying concepts or laboratory demonstrations.

Sometime during the day spent with you, we would like to have a brief conference with your principal. Perhaps you can help to arrange an appointment around the classes we specifically would like to observe. If it is convenient with your schedule, you are welcome to attend this meeting.

We will be looking forward to visiting with you.

Sincerely,

Leola Adams
Graduate Assistant

Colleen Caputo
Graduate Assistant
The school year is half over and it will not be long until you will be looking back on your first year of teaching experience in home economics. We are anxious to hear about your experiences and look forward to visiting with you on

Our plan is to arrive at the school by and would like to have a pre-conference for 10-15 minutes about the classes we will be observing. Would you please have available a copy of your lesson plan for each class we plan to observe? A detailed lesson plan is not expected -- just a general indication of your lesson goals.

We would also like to have a set of any evaluation techniques you have used in the three classes. These evaluation techniques could include formal devices (tests, quizzes), rating scales, checklists, student assignments, projects and performance tests or informal devices which are verbal descriptions describing references made from observations, classroom questions or conversations with students. We would be most appreciative if there were duplicates of your evaluative materials that we could look over during the classes we are not observing.

We are still planning to observe the following classes:

Sincerely yours,

Leola Adams

Colleen Caputo
APPENDIX H: MOST PROMISING ITEMS IN STE
Table 19. Analysis of variance components of most promising items (Gilbert 1974, pp. 50-51)

<table>
<thead>
<tr>
<th>Item</th>
<th>F ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C&lt;sub&gt;a&lt;/sub&gt;</td>
</tr>
<tr>
<td><strong>Subgroup I</strong></td>
<td></td>
</tr>
<tr>
<td>3. The teacher selected appropriate teaching techniques and</td>
<td>1.13</td>
</tr>
<tr>
<td>strategies for the situation.</td>
<td>1.50</td>
</tr>
<tr>
<td>5. The teacher indicated the objectives of the lesson and</td>
<td>1.31</td>
</tr>
<tr>
<td>their importance to the students.</td>
<td>1.43</td>
</tr>
<tr>
<td>19. The teacher utilized available educational resources of</td>
<td>1.46</td>
</tr>
<tr>
<td>the community in classroom procedures.</td>
<td>1.07</td>
</tr>
<tr>
<td>23. The teacher treats the students with respect.</td>
<td>6.25**</td>
</tr>
<tr>
<td>47. The teacher is well prepared for class.</td>
<td>2.04*</td>
</tr>
<tr>
<td></td>
<td>1.37</td>
</tr>
<tr>
<td><strong>Subgroup II</strong></td>
<td></td>
</tr>
<tr>
<td>4. The teacher created a positive, success-oriented</td>
<td>2.19*</td>
</tr>
<tr>
<td>learning environment.</td>
<td>1.97*</td>
</tr>
<tr>
<td>6. The teacher helped the students recognize the relationship</td>
<td>.95</td>
</tr>
<tr>
<td>of the lesson to previous learning or experience.</td>
<td>2.32**</td>
</tr>
<tr>
<td>7. The teacher conducted the lesson smoothly so that it</td>
<td>1.52</td>
</tr>
<tr>
<td>was easy for the students to follow.</td>
<td>2.13*</td>
</tr>
<tr>
<td>8. The teacher used meaningful examples or illustrations</td>
<td>2.71**</td>
</tr>
<tr>
<td>for conveying ideas during the lesson.</td>
<td>2.30**</td>
</tr>
<tr>
<td>14. The teacher was creative and flexible in guiding</td>
<td>1.79*</td>
</tr>
<tr>
<td>the learning process.</td>
<td>1.48</td>
</tr>
<tr>
<td>15. The teacher used questions to elicit thinking and student</td>
<td>1.64</td>
</tr>
<tr>
<td>response consistent with the instructional goals.</td>
<td>1.72*</td>
</tr>
<tr>
<td>16. The teacher was sincerely interested in the subject</td>
<td>2.16*</td>
</tr>
<tr>
<td>being taught.</td>
<td>1.36</td>
</tr>
</tbody>
</table>
17. The teacher effectively used a variety of levels of questions to evoke thinking beyond the level of recall.

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgroup III</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. The teacher emphasized reasons and relationships concerning the facts.</td>
<td>2.43*</td>
<td>.83</td>
<td>1.52</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td>1.35</td>
<td>1.26</td>
<td>1.17</td>
<td>.10</td>
</tr>
<tr>
<td>11. The teacher followed through with her plans and yet remained flexible enough to adjust as needs became evident.</td>
<td>2.39**</td>
<td>.87</td>
<td>.91</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>1.71</td>
<td>.38</td>
<td>1.00</td>
<td>.07</td>
</tr>
<tr>
<td>13. The teacher effectively used a variety of teaching materials and instructional materials.</td>
<td>2.53**</td>
<td>.02</td>
<td>1.34</td>
<td>.11</td>
</tr>
<tr>
<td></td>
<td>1.46</td>
<td>.34</td>
<td>.89</td>
<td>.10</td>
</tr>
<tr>
<td>24. The teacher maintains an open friendly rapport with the students.</td>
<td>4.22**</td>
<td>.88</td>
<td>1.72</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>1.90</td>
<td>.00</td>
<td>.74</td>
<td>.06</td>
</tr>
<tr>
<td>27. The teacher recognizes the individual's personal needs.</td>
<td>3.32**</td>
<td>1.23</td>
<td>1.17</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>1.93*</td>
<td>3.30</td>
<td>.83</td>
<td>.00</td>
</tr>
</tbody>
</table>

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

**Degrees of freedom for F are 1, 33. Table values for F are 4.44 at 5 percent and 7.74 at 1 percent.

* Significant at P < 0.05.

** Significant at P < 0.01.
APPENDIX I: RSTE ITEMS NOT USED IN CLUSTERS

**Items from Classroom Performance Section**

2. Objectives written in behavioral terms.

6. ...indicated the objectives of the lesson.

**Items from Cumulative Over Classes Section**

1. ...encouraged the students to describe or show how the learning affects them personally.

2. ...talks with students during or after class about their concerns.

3. ...handles discipline problems in positive, consistent manner.

6. ...works cooperatively with other teachers and the administration.

7. ...strives for improvement through professional growth activities.

22. ...uses techniques of assigning marks or grades consistent with school policy.
APPENDIX J: ITEMS RECEIVING X RATINGS
Table 20. RSTE items receiving X ratings

<table>
<thead>
<tr>
<th>Item</th>
<th>Rated X by both judges</th>
<th>Rated X by one judge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classroom Performance Items&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>100</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>18</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>30</td>
<td>8</td>
</tr>
<tr>
<td>Cumulative Over Classes Items&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>24</td>
<td>0</td>
</tr>
</tbody>
</table>

<sup>a</sup> n = 120 observations.

<sup>b</sup> n = 40 observations.
Table 21. HRS items receiving X ratings

<table>
<thead>
<tr>
<th>Item</th>
<th>Rated X by both judges</th>
<th>Rated X by one judge</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>12&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>58</td>
<td>3</td>
</tr>
<tr>
<td>19</td>
<td>52</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
<td>1</td>
</tr>
</tbody>
</table>

<sup>a</sup>n = 80 observations.
APPENDIX K: CHI-SQUARE FOR HRS ITEMS WITH X RATINGS INCLUDED
Table 22. Significant chi-squares for human relations skills behaviors with X ratings included

<table>
<thead>
<tr>
<th>Item</th>
<th>Pairs of observations</th>
<th>f^a/(F^b)</th>
<th>Chi-square^c</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40</td>
<td>26/15</td>
<td>12.77**</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>26/17</td>
<td>8.53**</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>25/15</td>
<td>11.60**</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>27/18</td>
<td>8.76**</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>25/17</td>
<td>7.45**</td>
</tr>
<tr>
<td>6</td>
<td>40</td>
<td>22/12</td>
<td>10.70**</td>
</tr>
<tr>
<td>7</td>
<td>40</td>
<td>21/9</td>
<td>18.47**</td>
</tr>
<tr>
<td>8</td>
<td>40</td>
<td>33/12</td>
<td>50.66**</td>
</tr>
<tr>
<td>9</td>
<td>40</td>
<td>32/8</td>
<td>86.51**</td>
</tr>
<tr>
<td>10</td>
<td>40</td>
<td>25/18</td>
<td>4.69*</td>
</tr>
<tr>
<td>11</td>
<td>40</td>
<td>28/11</td>
<td>38.35**</td>
</tr>
<tr>
<td>12</td>
<td>40</td>
<td>26/16</td>
<td>9.74**</td>
</tr>
<tr>
<td>13</td>
<td>40</td>
<td>27/12</td>
<td>28.40**</td>
</tr>
<tr>
<td>14</td>
<td>40</td>
<td>27/13</td>
<td>23.27**</td>
</tr>
<tr>
<td>15</td>
<td>40</td>
<td>26/10</td>
<td>35.64**</td>
</tr>
<tr>
<td>16</td>
<td>40</td>
<td>28/11</td>
<td>35.92**</td>
</tr>
<tr>
<td>17</td>
<td>40</td>
<td>32/10</td>
<td>63.03**</td>
</tr>
<tr>
<td>18</td>
<td>40</td>
<td>35/24</td>
<td>12.02**</td>
</tr>
<tr>
<td>19</td>
<td>40</td>
<td>38/20</td>
<td>32.22**</td>
</tr>
<tr>
<td>20</td>
<td>40</td>
<td>37/18</td>
<td>37.53**</td>
</tr>
</tbody>
</table>

^a^ \(f\) = number of teachers observed on the diagonal.

^b^ \(F\) = number of teachers expected on the diagonal.

^c^ Table value for chi-square with 1 degree of freedom is 3.84 at 5 percent and 6.63 at 1 percent.

*Significant at \(P < 0.05\).

**Significant at \(P < 0.01\).