DOCUMENT RESUME

ED 333 009 TM 016 444

AUTHOR Grover, Barbara W.; And Others

TITLE Scoring a Semi-Structured Interview for Assessment of Beginning Secondary Mathematics Teachers.

INSTITUTION Pittsburgh Univ., Pa. Learning Research and Development Center.

SPONS AGENCY Connecticut State Dept. of Education, Hartford.; Office of Educational Research and Improvement (ED), Washington, DC.

PUB DATE Jun 90

NOTE 54p.; Synthesis and revision of papers presented at the Annual Meetings of the American Educational Research Association (New Orleans, LA, April 5-9, 1988) and (San Francisco, CA, March 27-31, 1989).

PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.

DESCRIPTORS *Beginning Teachers; Educational Assessment; Evaluators; *Interviews; *Licensing Examinations (Professions); *Mathematics Teachers; Pilot Projects; *Scoring; Secondary Education; *Secondary School Teachers; Teaching Skills

IDENTIFIERS Subject Content Knowledge; Teacher Candidates; *Teacher Competency Testing

ABSTRACT

The semi-structured interview was investigated as a content-based assessment designed to take into account the complexity of teaching. A semi-structured interview licensing assessment for secondary mathematics teachers was developed and tested by the Connecticut State Department of Education. The scoring system converted the open-ended verbal responses of candidates into a set of meaningful numerical scores. Assessment was made on four general dimensions of teaching: content knowledge; content pedagogy; knowledge of students; and basic communications. Four specific tasks were used: (1) organizing a unit; (2) organizing a lesson; (3) alternative mathematical approaches; and (4) evaluating student error patterns. The scoring system also evaluated the interview as a whole. The data analysis focused on the reliability of the ratings and the validity of the assessment. The results of a pilot study conducted with 13 interviewers, 10 candidates, and 20 raters suggest that the scoring approach was viable for new assessment instruments to measure the complexity of effective teaching. Seven tables and three figures illustrate the study. A 25-item list of references is included. (SLD)

Scoring a Semi-structured Interview for Assessment of Beginning Secondary Mathematics Teachers

Barbara W. Grover
Orit Zaslavsky*
Gaea Leinhardt

Learning Research and Development Center
University of Pittsburgh, Pittsburgh, PA 15260

June 1990

Scoring a Semi-structured Interview for Assessment of Beginning Secondary Mathematics Teachers

Barbara W. Grover
Orit Zaslavsky*
Gaea Leinhardt

Learning Research and Development Center
University of Pittsburgh, Pittsburgh, PA 15260

June 1990

*Department of Education in Technology and Science
Technion - Israel Institute of Technology
Haifa, 32000 Israel

This article is a synthesis and revision of a papers presented at the annual meetings of the American Educational Research Association in New Orleans, April, 1988, and in San Francisco, March, 1989.

The research reported herein was supported in part by funds from the Connecticut State Department of Education, Office of Research and Evaluation, and by a grant to LRDC's Center for the Study of Learning from the U.S. Office of Educational Research and Improvement (OERI). The opinions expressed do not necessarily reflect the positions or policies of the Connecticut State Department of Education or OERI and no official endorsement should be inferred.

Abstract

Teaching is a complex and intellectually demanding profession. The semi-structured interview is a technique being explored as one of the content based assessments designed to capture this complexity. This research focuses attention on an approach to the design and development of a scoring system for a semi-structured interview license assessment for secondary mathematics teachers. The scoring system converts the open-ended, verbal responses of candidates into a set of meaningful numerical scores. Assessment is made on four general dimensions of teaching, on four specific tasks, and on the interview as a whole.

A pilot study involving experienced high school mathematics teachers who were trained as interviewers and raters, and six volunteers was conducted. The results suggest that the approach used to design the scoring system is a potentially viable one for new assessment instruments that measure the complexity of effective teaching.

Scoring a Semi-structured Interview for Assessment of Beginning Secondary Mathematics Teachers

Barbara W. Grover
University of Fittsburgh

Orit Zaslavsky
Israel Institute of Technology

Gaea Leinhardt
University of Pittsburgh

The educational community has been engaged in a concerted effort to reform since the National Commission on Excellence Report (1983), "A Nation at Risk", declared that the United States' mediocre educational system represented a serious threat to the nation's well being. One of the important aspects of the educational reform movement is the recognition that teaching is a complex, demanding profession that requires a broad array of skills and knowledge and is as cognitively demanding as the practice of medicine, law, or architecture (Clark, 1988; Darling-Hammond & Hudson, 1986; Leinhardt & Greeno, 1986; Shulman, 1987). The teacher must be prepared to dea! with varied contexts taking into account individual students, subject matter, and setting when planning and implementing instruction (Leinhardt & Smith, 1985; Shulman, 1987). The recognition of the intricacy of teaching has resulted in the general call by the reform movement for realistic, content based assessments of teachers (Carnegie Forum, "A Nation Prepared," 1986; Holmes Group- "Tomorrow's Teachers," 1986). State agencies, school districts, prcfessional organizations, and teacher education institutions have increased their efforts to evaluate teachers' knowledge and performance in a valid, unbiased manner.

A consequence of this attention to evaluation has been the recognition that current forms of teacher assessment are insufficient and radically different forms must be considered. Two basic approaches are now in use, multiple choice tests and classroom observations. Although they can and do measure important aspects of teaching, an academic knowledge base and the teacher-in-action, they are not sufficient (Pecheone, 1988; Wise & Darling-Hammond, 1987). They do not capture the thinking, reasoning, or decision making skills of teachers. It is in these tasks that an integration of several knowledge bases come into play and substantiate the complexity of the professional activity. A dialogue with the teacher provides an opportunity to evaluate the application of these knowledge structures to instructional practice and the rationale for employing them.

An interview format has characteristics that seem to address some of the limitations of the multiple choice tests and classroom observation instruments. Teachers are given the opportunity to demonstrate their thinking and reasoning skills which, in turn, provides the opportunity to assess in greater depth the ways in which teachers create an atmosphere and environment that facilitates learning. Teachers with basically the same kind of knowledge create different plans. There is no one recipe for "good teaching". The interview questions can be open ended allowing for a variety of acceptable answers.

Multiple choice exams, on the other hand, are restricted to single correct answers and measure only a small portion of the knowledge, skills, and behaviors required for competent teaching (Cole, 1984; Darling-Hammond, 1986; Haertel, 1988; Pecheone, 1988; Rudner, 1987; Wise & Darling-Hammond, 1987). The complexity of teaching is ignored. The full range of the knowledge base is not represented and the reasoning process for applying that knowledge is not measured.

Classroom observations measure the teacher-in-action. The criteria for evaluation, however, are based on assumptions that all teachers perform basically the same activities and that characteristics of effective teaching have no relation to the subject matter or the grade level and ability levels of the students being taught (Haertel, 1988; McLarty, 1987). In addition, expert teachers build a learning atmosphere in their classrooms over time. A single lesson is a part of an integrated plan for an entire unit of instruction (Clark & Yinger, 1979; Leinhardt, Weidman, & Hammond,1987). To observe a single class in October and another in April cannot measure the complexity of the instructional planning nor the daily interactions that contribute to the learning atmosphere.

The semi-structured interview is one of the new assessment techniques being considered as an alternative assessment because it may be able to capture the complexity of teaching. A second reason for exploring more realistic content based assessment is that the nature of the assessment upon which licensure or certification is based will ultimately produce changes in the classroom by influencing the nature of teacher preparation programs.

There are two parts to developing such assessments. One part is the development of the tasks and questions to be asked during the interview and the second part is developing the scoring system to evaluate the responses to those questions. This paper deals with the second part, the development of a scoring system for a set of tasks and questions.

The advantages of the interview format, however, create considerable difficulties in the design of a measuring instrument. The multiple choice test and the classroom observation instrument immediately translate into a score for an individual which can then be judged. In contrast, the interview produces a dialogue with the candidate which does not immediately translate into a score. In order to accomplish the translation, a conceptualization of a score for the dialogue is needed. This conceptualization requires a theory about good teaching, a theory of how to design a scoring system, and how to develop scoring procedures based on the design. Evaluating the performance of the teachers in this setting raises issues not confronted by multiple choice tests or classroom observation instruments.

One challenge that is set forth for interviews is to develop a means of scoring the ensuing dialogue in a reliable, valid, and unbiased manner. In addition, the scoring mechanism should provide results that can be used not only for evaluation and selection but to provide information to enhance the professional growth of the individual teacher. The assessment must also be defensible in the eyes of the teaching profession, the public at large, and conform to applicable legal standards. The criteria for evaluating the performance on the tests should relate to what it means to be a "good" teacher of a particular discipline. The scoring system for that performance should be designed so that a teacher would receive a similar score if the task had been administered on a different day or time while allowing for a multiplicity of correct answers. The reliability of the scores assigned by those who evaluate the teacher's performance should be sufficiently strong to warrant confidence that the teacher would receive a similar score no matter which interviewers or raters completed the evaluation. The established standards should discriminate in a fair and unbiased manner among those teachers who successfully meet the standards, those who meet the standards only marginally, and those who lack the necessary qualifications. The critical feature of the scoring system is that it should produce scores that lead to reliable decisions about the classification of a candidate. Thus, the critical reliability is the reliability of the decision.

In summary, the semi-structured interview technique reflects the view that teaching is a complex and intellectually demanding profession. It allows the examination of a teacher's thinking and reasoning processes that classroom observations and multiple choice tests cannot assess. Its structure and design permit critical issues to be addressed and to be linked to specific subject matter. Consequently, the semi-structured interview is worthy of consideration as an evaluation instrument for the assessment of teacher performance. Scoring the performance of a teacher in that setting is the problem that has not been solved. This paper takes some first steps toward suggesting a solution to that problem.

The semi-structured interview

The Connecticut State Department of Education (CSDE), in 1986, developed a semi-structured interview for secondary mathematics teachers to be incorporated into their teacher licensure requirements. The decision to use this interview format made Connecticut a pioneer in the assessment area of the teacher reform movement.[1]

It is this semi-structured interview for which the scoring system was developed. The interview consists of four tasks: Organizing a Unit, Organizing a Lesson, Alternative Mathematical Approaches, and Evaluating Student Error Patterns (see Tomala, 1989, for a detailed discussion of the development of the tasks).

Task 1 Organizing a Unit: Candidates are given a set of cards listing 10 subtopics that would be taught within a given unit of study. Candidates are asked to discuss how they would order the subtopics in teaching the unit, why they would choose a particular sequence, and how they might change the overall structure depending on the ability levels of students.

Task 2 Organizing a Lesson: Candidates are given pages from a text book and are asked a series of questions about how they would teach a lesson on the topic covered in the text pages.

Task 3 Alternative Mathematical Approaches: Candidates are provided with descriptions of several different approaches to teaching a particular topic and are asked a series of questions concerning the advantages and disadvantages of these approaches.

[1] Currently Connecticut is sharing their experience with other states and the state of California, in particular, through the New Interstate Teacher Assessment and Support Consortium. The state of Minnesota and its Board of Teaching (Wise & Darling-Hammond, 1987) are capitalizing on the pioneering work of Connecticut in their efforts to revise entry standards into the teaching profession in their state.

Task 4 Evaluating Student Error Patterns: Candidates are asked to identify errors in the work of three students and discuss appropriate remediation strategies.

Imagine for a moment the scene of the interview. Two people, an interviewer and a beginning teacher, are seated at opposite sides of a table in a medium sized room. In an unobtrusive corner a technician operates a video camera, videotaping the interactions between the two people seated at the table. The beginning teacher is being interviewed on Task 1, Organizing a Unit. ,The teacher is given a set of topics, some time to review them, and is instructed to arrange them in the order in which she would teach the unit. Once the teacher has arranged the cards, the interviewer asks a series of questions about the arrangement. Why did you begin with this topic? Why does this topic follow that one? Imagine the possible responses. Imagine trying to evaluate those responses. This paper describes one approach to the development of a scoring system that allows this performance to be evaluated as competent or not competent in a way that would be reliable, valid, and fair.

Development of the system

An important feature of the dialogue of this semi-structured interview is that the tasks and questions are faithful to the practice of teaching. The dialogue is about teachers' reasoning, thinking, and decision making with regard to instructional practice. These are the acts involved in being a teacher. The approach we have taken to the evaluation of the performance is also faithful to the practice of teaching because it is grounded in theory, subject matter knowledge, and the wisdom of experienced mathematics teachers. This wisdom is the type of experience that Joseph Schwab and Lee Shulman refer to as the Wisdom of Practice and that Gaea Leinhardt refers to as Craft Knowledge. An important aspect of the scoring system is the fact that the criteria for evaluation are visible and public. The loosely structured dialogue is translated into a profile of scores which can be used to identify strengths and weaknesses for professional development as well as candidate selection.

In order to transform a lengthy free flowing discussion into a useable set of information, the interview had to be conceptualized in a particular way. The general structure of the scoring system for the semi-structured interview includes three major parts: tasks, dimensions, and components. The tasks correspond to the four parts of the interview described above: Organizing a Unit, Organizing a Lesson, Alternative Mathematical Approaches, and Evaluating Student Error Patterns. Dimensions refer to the general construct of teaching, incorporating different aspects of teacher knowledge. The dimensions assessed by the interviews are: 1) Content Knowledge which refers to the teacher's understanding of the mathematics involved in the task; 2) Content Pedagogy which refers to the teacher's knowledge of how to teach the mathematics; 3) Knowledge of Students which refers to the teacher's ability to take into consideration the interest, motivations, and abilities of the individual learner when organizing, planning, and implementing instruction; and 4) Basic Communication which refers to the teacher's ability to communicate in a cohesive and coherent manner. The choice of the dimensions was influenced by the work of Leinhardt (1987, 1990), Leinhardt and Greeno (1986), Shulman (1986), our work with the

Teacher Assessment Project at Stanford, and our own teaching experience.

The third major part of the scoring system is the Components. The components are the heart of the scoring system. A single component represents a particular competency of the candidate. Each component is associated with a particular task and is also associated with a particular dimension that cuts across tasks. In turn, each component has associated with it particular weights and anchors.

The concept of a component requires additional clarification. In Task 1, Organizing a Unit, candidates are asked to arrange and organize 10 subtopics of a unit of study in the order that they would teach them. The candidates are then asked to discuss and justify their arrangement as well as any groupings of topics within the overall arrangement they think are legitimate. One of the components for this task is:

1.1 Candidate takes into account the major math principles and concepts in planning and organizing the instruction. (CK).

0	*2*	*4*	*6*	*8*	*10*
no answer	*unacceptable order OR justification NOT based on math structure. or difficulty levels. OR no justification*	*acceptable order; justification given as general statements, does not include specifics*		*acceptable order; justifications based on math structure on levels AND or difficulty include specifics*	

This component assesses a particular competency of the candidate, the candidate's ability to use his/her understanding of major mathematical principles and the relationships between them to organize a unit of instruction. The component is associated with Task 1 and is associated with the dimension Content Knowledge, denoted by the letters CK in the parentheses. The component has a weighted rating scale from 0 to 10. Anchors are provided for low, middle, and high ratings.

A second example is:

1.4 Candidate justifies additions that can be made to improve the unit. (CP)

0	*1*	*2*	*3*	*4*	*5*
no answer	*general statements;no justification*	*1 specific addition AND justification based on completion of the unit OR adding variety OR general skills .*		*several additions AND justification based on at least 2 of completion of the unit, adding variety, or general skills.*	

As with the first example, this component is associated with Task 1, but this component is associated with the dimension Content Pedagogy denoted by the letters CP. The component has a weighted rating scale (0-5) and anchors for low, middle, and high ratings. Similar components are associated with each of the tasks.

An overview of these relationships is presented in Figure 1. The dimensions are listed in the left hand column; each of the tasks are column headings. Within each cell of the matrix are listed the components that are associated with each dimension within each task. For example, in the upper left hand cell is listed the abbreviation for component 1.1 Takes math into account[2] This abbreviation refers to the following component: *1.1 Candidate takes into account the major math principles and concepts in planning and*

[2] A decimal form of notation is used to designate the various components. The whole number part identifies the task and the decimal value identifies the position of the component in the sequential listing of all the components within the task. For example, Component 1.1 is the first component listed for Task 1, Component 2.4 is the fourth component listed for Task 2, and so on.

organizing the instruction. This component is associated with Task 1 and the dimension Content Knowledge. Figure 1 shows that each of the four dimensions are represented in each of the tasks.

Insert Fi; ure 1 about here

These three major parts of the scoring system are integrated to provide scores for the candidate. The relationship of the components to the tasks and the dimensions is shown in Figure 2. For each component, the decimal identification number and the maximum point value on its rating scale are listed in the appropriate cell. Working vertically, the ratings on all the components associated with one task are combined to form a Task Score. Each candidate receives four separate Task Scores. Working horizontally, the ratings on all the components associated with one dimension are combined to form a Dimension Score. Each candidate would receive four separate Dimension Scores. Finally, the scores combine to form an Interview Score. This system provides a mechanism which allows one to move from the specific, open-ended, ill structured, verbal responses of a candidate to a question and to the probes associated with that question to an assessment score on a general valued dimension of teaching. This design avoids the problem of localizing the score to a specific response to a specific query, an analytic path that would merely be a spoken version of the multiple choice tests. The ability to convert verbal responses to the interview questions to a score on a task and/or a dimension is the power of this scoring system developed for the semi-structured interview technique.

Insert Figure 2 about here

Special Features

Three special features of this assessment that distinguish it from the teacher tests currently in use should be noted. The first feature is that the dimensions of Content Knowledge, Content Pedagogy, and Knowledge of Students are assessed. Unlike the generic performance assessments currently used, this scoring system places a heavy emphasis on assessing the content knowledge and content pedagogy of the candidates. The teacher needs to understand the mathematics and needs to have a sense of how particular lessons or units fit into the overall mathematics curriculum (Content Knowledge). The teacher also needs to know how to teach the mathematics (Content Pedagogy). The teacher needs to take into consideration the interests and abilities of the learner when thinking about the instruction (Knowledge of Students). This involves making a series of decisions and the teachers should have valid reasons for the decisions they are making. The teachers are evaluated on their ability to demonstrate their knowledge of content, pedagogy, and students and defend their instructional decisions on the basis of the mathematics and the pedagogy.

A second feature is that the scoring system is generic at one level and topic specific at another. The interviews are written in such a way that the questions can be used with a variety of mathematics' topics. One of the characteristics needed to provide an objective, effective, and reasonably efficient scoring system is that the scoring system should be as connected to the specific subject matter as possible and yet be as general as possible. To be connected to the specific subject matter means that the mathematical content of a particular topic should play an important role in the evaluation process. To be general, means that the system could be used, with only minimal adjustments, for most significant topics in middle school through pre-calculus mathematics (e.g., linear equations, ratio/proportion/percent, linear functions). The components of each task need to be generic to mathematics, grounded in the topic of the task or the interview but not restricted to it. To resolve this dilemma of grounded specificity versus generality, the following approach was taken. The criteria on which each of the components are rated would be generic to mathematics and to the characteristics or skills being evaluated. Only in the specifics of the answers and the examples of a candidate's statements would the particular mathematical topic involved in the task play a part. At the same time the weight given to the specific details of the mathematics included in the answer was to be retained; otherwise candidates could get high scores without being able to appropriately use the necessary mathematics.

A major part of the evaluation of the candidates is to assess their content knowledge and how that knowledge influences other decisions. Thus, the content of the particular topic is a vital part of the answers and what constitutes quality answers. The independence of the other parts of the system (e.g., the components, the anchors on the rating scales) permits the topics to be altered for different interviews with minimal effort (i.e., only the answers and examples of candidate responses need to be created when a new topic is selected). The components and the criteria on which candidates are evaluated remain the same.

Table 1 summarizes the parallels between the interview questions and the scoring system with respect to their generic and specific elements. The basic tasks, the components associated with the tasks, the dimensions associated with the components and the anchors associated with each of the components are relatively topic free (e.g., ratio/proportion/percent, linear equations). However, the answers expected from the candidates are entirely dependent on the specific topic addressed in the interview. The statements required by the candidate to obtain high ratings are expected to contain specific information related to the topic.

Insert Table 1 about here

The arrows in Figure 3 show how the specifics of candidate responses and the specifics of acceptable answers for a particular topic were used to inform the development of the generic criteria for the anchors and the generic components.[3] The components and the anchors are related to a specific task but are not specific to a topic. The general dimensions also informed the creation of the generic components and each component was coded to one of the dimensions. The tasks are also generic. The questions asked during the interview

[3] The videotaped responses of 24 beginning and experienced mathematics teachers who volunteered to participate in these interviews in December of 1986 or in July of 1987 provided the data base for realistic specific responses.

related to the task are not directly related to the specific topic. Only the stimulus materials to which those questions refer are specific to the topic. It is this specific topic information that is included in the acceptable answers and expected in the candidate's responses. This structure of being both generic and topic specific is uniqıe and may prove to be exceptionally powerful.

Insert Figure 3 about here

The reader should note that the relationship between the interview questions and the scoring system is more complex than the usual one-to-one correspondence of other test situations (i.e., a question is asked and the answer to that question is scored). In answering a particular question, the interviewee may unintentionally answer a question that occurs later in the interview, may refer back to a prior question and add information, or may revise what has been said at any time. This fluidity means that the scoring system should not merely be a right/wrong decision on each question or a rating on each question but rather take into account all that is said during the interview and evaluate the response as a whole.

A third feature of the scoring system is that it is a multilevel assessment which serves a diagnostic as well as a selection purpose. The purpose of the system is to discriminate levels of competency and to aid in identifying strengths and weaknesses of candidates. A general classification based on overall performance can be made to identify those candidates who are incompetent, those who are competent but exhibit some serious weaknesses, and those who are highly competent with few weaknesses. The classifications are Not Pass, Pass, and High Pass, respectively. A profile can be generated for eight different aspects of teaching- the four tasks and the four dimensions. The state, the school district, and/or the teacher can use this profile to provide the support necessary for self-improvement.

Development of Documents Used for Rating and Classifying Candidates

The evaluation of a candidate's performance can be conducted in a variety of contexts. The procedure used in the evaluation of candidates influences the design and format of these documents. The procedure chosen was to have a rater evaluate a videotape of the interview. This procedure has the advantage of providing a permanent record of the interview if later review is required. In addition, it permits the rater the option of evaluating portions of the interview in "real time" and/or reviewing portions if desired.

The rater needs a framework for notetaking to help in the evaluation process because the components are not directly tied to a specific question and because the candidate may provide information related to any of the components at any time during the performance of the task. A guided notetaking form structures the organization of the rater's notes so that all information related to a particular component is recorded in the same place on the form. The rater's attention and notes are focused on the components to be evaluated not on the questions being asked. Table 2 shows a sample portion of the Guided Notetaking pages for Task 1. The listing of the topics to be ordered under the first section allows the rater to record the sequence in which the candidate orders the topics and space is provided to record justifications for that order. In section 2, the rater's attention is focused on classifying remarks about students into those dealing with familiarity of the topics to the students, motivating students, and different ability levels of the students. Again space is provided for justifications. In the third section, space is allotted for the groups identified by the candidate and the justifications provided for those groups. This information forms the basis of the criteria for rating the candidate on components 1.1, 1.2, and 1.3, respectively. Similar spaces associated with components 1.4, 1.5, and 1.6 are included on additional guided notetaking pages

Insert Table 2 about here

Once the raters have completed taking notes, they make a decision about what rating to give the candidate on each component ar.d circle a number or place a check mark between numbers on the rating scale. Table 3 shows a partial listing of the Component Ratings for Task 1. Note again how the anchors listed on the rating scale relate to the organization of the Guided Notetaking page in Table 2. Without this framework raters might take notes focusing on the answers to specific questions and would then have to reorganize those notes to fit the components. In addition, raters might miss some of the information related to the components if they were to work without this organizational structure.

Insert Table 3 about here

A documentation form is a duplication of the components but without the rating scales. The form includes a place to record any statements that reflect negatively on the candidate that were not evaluated in the components. These pages are used only for Pass or Not-pass candidates. Because the semi-structured interview is a part of the licensure process, documentation of weak or poor performance is essential. The licensing agent must be prepared to defend the evaluation of a candidate against legal action taken by the candidate or other parties contesting the evaluation. When a candidate receives a Pass or Not Pass classification on the first evaluation, the raters review the candidate's responses a second time and provide specific documentation to support low ratings. In addition, they provide important diagnostic information for the candidate to use as a guide for self-improvement during the coming year. The Task Score page allows the rater to aggregate the ratings on a task and classify the candidate. A sample of the Task Score page for Task 1 is shown in Table 4. A similar document was created for aggregating each of the dimension scores. In the blank spaces in the left hand column headed "Rating", the rater would write in the rating given to the candidate on each of the components. The sum of these ratings is placed in the blank next to the label Total Component Score. Appropriate values for negative and positive comments would be recorded in the blanks next to these labels and combined with the total component score to result in an Adjusted Score.

Insert Table 4 about here

The standards for determining classification of the candidate appear on the right side of the page. The rationale for the standards for both high pass and pass is as follows: In order to obtain a high pass classification, a candidate should have earned at least the highest pass points possible on two of the three major components (i.e., > or = 6). In addition, only a limited number of nonpass or low ratings are allowed. The adjusted point total is the sum of the minimum high pass point values on the components. For example, in Task 1, the minimum rating for High Pass is a 7 on components 1.1, 1.2, and 1.3. The minimum rating for High Pass is 3.5 on components 1.4, 1.5, and 1.6. Consequently, the minimum adjusted score to earn a High Pass rating is 31 ((3x7) + (3x3.5)).

In order to obtain a Pass classification, a candidate should have earned at least a Pass rating on two of the three major components (i.e., > or = 4). In addition, only a limited number of non pass or low ratings are allowed. The adjusted point total is the sum of the minimum pass point values on the components. For example in Task 1, the minimum rating for Pass is 4 on components 1.1, 1.2, and 1.3. The minimum rating for Pass is 2 on components 1.4, 1.5, and 1.6. Consequently to earn a Pass rating, the minimum adjusted score is 18 ((3x4) + (3x2)).

The three classifications for which standards are being established are High Pass (HP), Pass (P), and Not Pass (NP). High Pass means a candidate has demonstrated a high level of competence on most of the objectives measured by the task. The candidate is rated as competent with no weaknesses identified as needing remediation.

Pass means a candidate demonstrated a reasonable level of competence on most of the objectives measured by a task. The candidate is rated as competent with some weaknesses identified as needing remediation. Not Pass means a candidate demonstrated little or no competence on most of the objectives measured by a task. The candidate is rated as having a significant number of weaknesses identified as needing remediation.

In summary, the general sequence of procedures followed in developing the scoring system for the semi-structured interview was as follows:

1. Develop the general framework
2. Articulate the rationale and assumptions
3. Develop dimensions
4. Analyze the individual task
5. Develop the component/anchor descriptions and determine weights for each component
6. Develop documents used for rating and classifying candidates

In describing the development of the system and attempting to articulate the issues and concerns that were taken into account, the process loses its dynamic quality. In reality, several domains (i.e., components, anchors, specific acceptable answers) were considered simultaneously or interactively, moving back and forth from one domain to the other. The generic aspects of the components and the anchors developed from analyses of the specific data provided by the tapes of the interviewers and the research literature. However, often work on components and anchors in one task informed our thoughts about components within the same task or components and anchors for a different task. As we created the specifics that constituted acceptable answers or selected example statements that candidates might make to earn high ratings, our ideas and thoughts suggested revisions for the overall framework for the system or the documents that would be needed to convert the dialogue into scores. One area involved in the development of the scoring system (e.g., the components) often influenced our thinking on the other areas. In addition, the process suggested significant changes in the interview questions and prompts, which were incorporated in the fall of 1988. This dynamic process produced an integrated whole made up of interviews, scoring documents, and scoring procedures.[4]

Pilot Study

A pilot study was conducted to collect data on the various aspects of the semi-structured interview. The major objectives of the pilot study were: (a) to determine whether the administration and scoring of a semi-structured interview is a manageable task; (b) to determine whether raters could be trained to reliably score the interviews, and (c) to determine whether the scoring system would validly discriminate among non teachers, beginning teachers, and experienced teachers. In essence, the pilot study was conducted because preliminary data was sought to inform policy decisions by the Connecticut State Department of Education with respect to

[4] Technical reports detailing the development of the documents and the rater training manual are available from the authors. (see Grover, 1989; Grover & Zaslavsky, in preparation; Grover, Zaslavsky, & Leinhardt, 1989.)

continued support for the interview and the corresponding scoring system as a form of assessment in the teacher licensure program.

Participants.

An important factor in evaluating teachers in a fair and equitable manner is the competence of the people who will be administering the interview and the competence of the people who will be making ratings based on the scoring system. In order for the dialogue in the interview to be meaningful, the interviewer must share a similar professional knowledge base with the candidate (i.e., the interviewer must be knowledgeable, if not an expert, in the field). In addition, interviewers must be trained in interviewing techniques and the particulars of the specific interview in order to standardize the administration of the interview. It is important that the interviewers be aware of the extent to which they should and can probe for additional responses from candidates (see Winters, 1990, for details about interviewer training.).

Raters and candidates should also share a similar professional knowledge base. Two additional rater characteristics are essential to competent use of the system: expertise at teaching mathematics and a shared understanding of the criteria being evaluated. To produce this understanding a training manual was developed that describes, in detail, all the criteria for rating each of the components for each of the tasks.

Three groups of participants were involved in the pilot study, the interviewers, the raters, and the candidates to whom the semi-structured interview was administered. All three groups were volunteers recruited by mail by the State of Connecticut's Office of

Research and Evaluation (ORE) from the selected group of secondary mathematics teachers and mathematics educators available in Connecticut. The individuals to whom letters were sent were recommended by the State Curriculum Consultant for Mathematics. The basis for the selection of the raters and the interviewers as potential volunteers were their reputations as high quality secondary mathematics teachers, their contribution to the professional community, and the personal judgment of the Curriculum Consultant. Interpersonal skills were an additional factor considered in recommending interviewers and a high degree of knowledge of mathematics was an additional factor considered in recommending raters. Altogether, 13 interviewers, 10 candidates, and 20 raters participated.

The 10 candidates consisted of four beginning math teachers (i.e., less than three years of teaching experience), four experienced math teachers (i.e., more than five years of teaching experience), and two non math teachers. The non math teachers were people with high math knowledge who had not been enrolled in a teacher preparation program and had not receive any formal teacher training. The two non math teacher volunteers were recruited by a staff member of the Connecticut State Department of Education (CSDE).

Interviewer Training.

Thirteen secondary mathematics teachers underwent one day of interviewer training under the direction of CS staff. The interviewers were trained to administer an intervie for a specific task on a specific topic. Two interviewers were trained on Task 1 for

Ratio, Proportion, Percent, and two on Task 1 for Linear Equations, two on Task 2 for Ratio, and two on Task 2 for Linear, and so on. Interviewers were randomly assigned to the specific tasks.

Administration of interviews

In the fall of 1988, the interviews were administered in an assessment center setting at a high school in Connecticut. Each of the interviewers was assigned to a separate room in which s/he administered the same task interview throughout the day. A professional staff videotaped the interviews. Each candidate participated in two sets of interviews. The tasks in each set were Organiz'ng a Unit, Organizing a Lesson, Alternative Mathematical Approaches, and Evaluating Student Error Patterns. One set dealt with Ratio, Proportion, Percent and the other with Linear Equations. Within each designated group of beginning teacher, experienced teacher, and non-teacher, the candidates were randomly assigned to the topic on which they would be interviewed first.

The candidates completed a set of tasks on one topic in the morning session and then completed the remaining tasks on the second topic in the afternoon. To control for the order in which the tasks were administered to an individual candidate, the 10 candidates were randomly assigned to tasks within each scheduled time frame within each topic.

Rater Training

Twenty secondary mathematics teachers from the state of Connecticut underwent two and one-half days of training. These sessions included discussions of the purpose of each component, the meaning of each of the criteria, what constitutes justifications and details with respect to the specific topics, and examples of acceptable and unacceptable answers, especially the extreme cases: answers that would earn high ratings, and answers that would earn low ratings. Detailed discussions of the anchors of the components for each of the four tasks were held and practice in scoring eight videotapes, one tape of each task on each topic, was provided. The training tapes were selected from the data collected at the administration session described above.

Raters were trained to use a two stage process in evaluating performance on a given component. First, they were to make a general match between the evidence in their notes and one of the three anchors on the rating scale (HP, P, or NP). Then they were expected to engage their professional judgment as to the quality of the evidence within the range of that anchor. If the response was judged to be of average quality for that criteria for that anchor, the middle numerical value within the range of that anchor was to be used. If the response was judged to be of above average quality for the criteria for that anchor, the upper numerical value within the range of that anchor was to be used. Similarly, for responses of lesser quality, the lower numerical value within the range of the anchor was to be used.

During the training, the raters viewed videotapes of beginning teachers, scored the tapes, and then compared their notes and scores with standards that had been prepared in advance by the authors for each of the candidates whose videotapes were viewed. The standards included a sample set of completed working materials, with annotated notes on the guided notetaking pages that indicated why specific ratings were given for each of the components. Discussion followed the scoring of each videotape. These discussions and the practice of scoring videotapes similar to those that would be evaluated in actual scoring sessions promoted a shared view of the criteria for scoring.

Scoring

Of the 10 candidates' tapes, four were used as training tapes and six were used for the actual scoring in the pilot study. The videotaped interviews of those six candidates were scored during two scoring sessions. Each interview was scored by at least two raters. The individual interviews were arranged on the video tapes according to the interview administration schedule of a particular task in a particular room. Raters were assigned to score a particular task on a particular topic. Each rater viewed and scored his/her assigned set of videotapes independently in a separate room.

Each scoring session began with a brief refresher training session. All the raters scored three interviews for which a standard had been established by the authors. This review allowed raters to recallibrate their scoring to the shared view of the rating scales that was the focus of the training.

Data Analysis

Analysis of the data focused on the reliability of the ratings and the validity of the assessment. Reliability was evaluated at two levels for both the Tasks and the Components. The first level was the general classification into High Pass, Pass, or Not Pass categories. Percent of Agreement was used to measure reliability. The second level was the numerical value associated with the Task Score or the numerical value associated with the rating on the component. A Pearson Product Moment Correlation was used to assess reliability of these numerical scores.

Reliability

The scoring system is designed to provide nine separate scores for each candidate, four task scores, four dimension scores, and one overall interview score. Having a variety of scores supports the two major objectives of the assessment, to allow decisions to be made for selection into the teaching profession and to provide feedback to candidates on their strengths and weaknesses. These scores are dependent upon the rating a candidate receives on various subsets of the 22 components distributed throughout the four tasks. Consequently, if reliability among raters on the rating scale for each component is good, then reliability on the more global measures (i.e., task scores, dimension scores, interview score) will generally follow. The converse is not true. Good reliability indices for the more global measures do not necessarily mean reliability on the finer grained measures. However, the relationship between the reliability indices for more global measures and the reliability indices for the finer measures offers clues as to which parts of the scoring system are working and which need to be revised (e.g., the group of components for one task is working, but one particular component is not).

Reliability measures of classification are the most crucial for the purposes of this assessment. The goal is to classify individuals according to a set of absolute standards not according to their relation to each other. The objective of the assessment is to distinguish between those candidates who are qualified to teach and those who are not qualified. Percent of agreement measures the extent of agreement among raters on the classification of the candidate as to High Pass, Pass, and Not Pass. It is these classifications that are of significance in this evaluation. The Pearson Product Moment Correlation measures the extent of agreement among raters on the relative order of the candidates. These values are important in guiding revisions of the instrument but are not as relevant to the goal of the assessment as the percent of agreement. Licensure is based on a candidate meeting a particular standard not on being rated higher than other applicants.

Task Level Reliability. Agreement on classification on a given task and topic indicates that the system is working at a global level for selection purposes. Raters are able to reliably classify performance on a particular topic for a particular task as High Pass, Pass, or Not Pass. Lack of agreement on these classifications warrants attention to the extent of agreement on the total score for that task. Low agreement on classification may be caused by very discrepant total scores. These discrepant total scores would imply fairly large differences among raters on the numerical ratings each assigned to some or all of the components of that task which would be reflected in a low reliability coefficient. On the other hand, low agreement on classification may not mean serious discrepancies exist on the total scores but rather that slight discrepancies occurred at the critical cut points. This outcome would be reflected in a fairly high correlation coefficient between total scores and would warrant a look at the standards set for classification. It could also be the case that total scores are quite similar but the distribution of ratings among the individual components is discrepant (i.e., the overall evaluation coincided but the particular evidence related to a specific component was categorized or interpreted differently). This situation would be indicated by a low reliability coefficient between component scores and suggests a look at the finer grained decision making at the component level.

Table 5 shows a summary of the reliability of the raters. There were 17 pairs of raters across all the tasks and topics. The first row shows that 76% of the pairs of raters agreed on classifications of task performance for at least 4 out of the 6 candidates (> or = 67%), and 60% agreed on at least 5 of the 6 candidates (> or = 83%). In terms of the reliability coefficients, the reliability coefficients were above .70 for 82% of the pairs of raters and above .80 for 47% of the pairs

Insert Table 5 about here

The second row in the table reflects the reliability of the raters at the component level. Fifty-seven percent of the pairs of raters agreed on the classification of performance on particular components of 67% or more of the candidates and 31% agreed on the classification of 83% or more of the candidates. The reliability coefficients were above .70 for 55% of the pairs of raters and above .80 for 34% of the pairs

These two reliability measures provide two different kinds of information. The Pearson Product Moment Correlation indicates the extent to which raters agreed on the relative order of the candidates.

The percent of agreement indicates the extent to which they agreed on the classification of the candidate into one of the three categories. The classification is obviously the most critical evaluation and so the percent of agreement is the more important of the two reliability measures for our purposes. The combination of the two measures, however, informed the revision process. These results are encouraging and suggest that raters can be trained to reliably classify and score candidates at the task and component levels.

Validity

The second major issue in evaluating the scoring system is whether the ratings discriminate among the experienced teachers, the beginning teachers, and the non-teachers. Comparisons were made in three general areas, the tasks, the dimensions, and the interview performance as a whole.

Tasks. On the tasks, .ie expectations were that experienced teachers would do better than beginning teachers, who in turn would do better than non teachers.

Insert Table 6 about here

Table 6 shows the results of both the Kruskal-Wallis and ANOVA analyses of the Tasks. An ANOVA takes into account the within group variance as well as the between group variance. When within group variance is small relative to the between group variance significant differences result. However, the Kruskal-Wallis takes into account only the between group variance because it uses the ranks of the scores not the actual scores. In addition, it should be noted that because of the small sample, there is only one pattern of ranks that will result in significant differences with the Kruskal-Wallis test. If the rankings of the non teachers are 1 and 2, the beginning teachers are ranked 3 and 4, and the experienced teachers are ranked 5 and 6, then significant effects in the predicted direction will be found. Any other arrangement leads to non significant findings.

The hypothesis with respect to task scores received mixed support on the topic of Ratio in that differences were significant at the .05 level on Tasks 1 and 3 but not on tasks 2 and 4 for both the Kruskal-Wallis and ANOVA. There were no significant differences in performance on any task for the topic of Linear Equations.

Dimensions. Similar analyses were conducted on the dimensions. Table 7 shows the results of those analyses. The predictions in this case were that for the Content Pedagogy and Knowledge of Students dimensions the experienced teachers would perform better than the beginning teachers and the beginning teachers would perform better than the non teachers. The Kruskal-Wallis analysis of the data supports that prediction, but the ANOVA analysis supports that prediction only for the topic of ratio on the dimension Content Pedagogy. These differences occured because the within group variance on Dimensions scores was fairly large relative to the between group variance in the other situations and non-significant results occured. The rankings used in the Kruskal-Wallis obscure these differences in the data. The rankings are the same in all Dimensions for all topics (except Basic Communication on the topic of Linear Equations) despite differences in the actual data values.

Insert Table 7 about here

The prediction for the dimensions of Content Knowledge and Basic Communications was that the groups would perform equally well. The data do not support that contention in that significant differences were found between all groups in the Kruskal Wallis analysis for these two dimensions and for the topic of ratio in the ANOVA analysis. It is not clear why the differences occured in the Content Knowledge. One hypothesis is that the experienced teachers simply included their knowledge of mathematics in their descriptions more often. The other subjects did not provide this data for evaluation. Differences in Basic Communications may be due to rater bias. Many raters informally indicated that they could not separate the ability to use the English language from the ability to coherently discuss the mathematics. The raters tended to lower scores for those who could not articulate their mathematical ideas even though their general ability to communicate was adequate.

For the overall interview differences were expected in the same direction as for the tasks. The data supported this hypothesis for the topic of Ratio but not for Linear Equations. The results of the two analyses are shown in Table 7. In general, this preliminary evaluation suggests a potential for being able to validly discriminate among groups. Studies employing larger samples, more diverse populations, and additional topics are needed.

Conclusions

We have described a model for the process of developing a scoring system for one of the new content-based assessment techniques, namely the semi-structured interview. It appears the complexity of teaching can be captured in an interview and that the dialogue of the interview can be translated into a score. Four characteristics of this approach to the design and development of the scoring system are particularly noteworthy for new conceptualizations of assessment.

First, the system remains faithful to the practice of teaching; it reflects the acts involved in teaching; it recognizes that teaching is complex, dynamic, and cognitively demanding. By adding this kind of assessment to the licensure process, the licensing agency says that it recognizes that teaching is a complex, cognitive activity which involves the integration of a variety of types of knowledge and that aspect of teaching is worthy of assessment. Furthermore, it says that the agency supports the concept of employing only those individuals who display the qualifications which enable them to perform this complex cognitive activity.

Second, the scoring system is grounded in theory, subject matter knowledge, and the wisdom of practice, integrating general principles of good teaching with specific content knowledge in a matrix framework. The theoretical framework operationalizes the concept of good teaching. This framework helps to provide a clear target not only for the teacher applicants who will be taking the licensure exam, but for the interviewers and raters associated with the process, and all educators associated with teacher preservice and inservice preparation.

Third, the criteria for evaluation are visible and public and allow for a broad range of style and philosophy to be accommodated. Candidates will know ahead of time what the criteria are on which they are to be evaluated. The criteria are clear at the component level, at the task level, at the dimension level, and at the interview level. The translation of the interview performance into a score is transparent and can be traced from the raters notes on the Guided Notetaking pages, through the evaluation of specific components, to the standards for classification on tasks, dimensions, and the interview as a whole. There is no mystery.

Finally, the scoring system converts the open-ended dialogue of a conversation between professionals into a profile of numerical scores that can be employed for both selection decisions and diagnostic purposes. A general theme of the reform movement is an interest in professionalizing teaching. The emphasis is on helping beginning teachers become better teachers, not removing them from the profession. The components can be aggregated into different configurations to create a profile of strengths and weaknesses that a candidate can use as a guide to self improvement.

This structure of the interview and the scoring system takes a first step toward having members of the profession explicate what is expected from those who wish to enter the profession and accepting the responsibility of being the gatekeepers for the profession. Mathematics teachers must be interviewed by mathematics teachers. To interview properly, the interviewer must understand the responses of the candidate and be able to probe appropriately. An art teacher or an English teacher would have difficulty conducting an effective interview with a mathematics teacher and visa versa. Even though the scoring system has an analytic foundation, professional judgment about the adequacy and quality of the response is required in rating a candidate. Mathematics teachers must serve as the raters for mathematics teachers. This excludes individuals who have only a knowledge of mathematics from the process of evaluation. It is the intertwining of the subject matter knowledge and the knowledge of pedagogy that is critical. A knowledge of anatomy does not qualify a person to be a physician. A knowledge of mathematics is a necessary but not sufficient condition to qualify a person as a teacher of mathematics.

The recurring theme of this paper has been that teaching is a complex and demanding profession. The challenge was perceived as being the need to design and develop an assessment instrument that takes into account that complexity. In the process we have accepted the obligation to ground our instrument in theory and validate it empirically. The instrument described in this paper is well grounded in theory. The process of adequate empirical validation is just beginning.

Future issues

One of the major development issues is the conflict between What Is and What Should Be. In deciding what is important and what is to be measured, we must try to achieve some balance between evaluating the skills we believe ought to characterize the performance of beginning teachers, our vision of the future, and evaluating the skills that are currently being taught in teacher training programs or by their mentor teachers. This issue is of major importance to schools of education which must share in the vision if reform of teacher preparation programs is to become a reality.

There are also many policy issues to be addressed. How much information should be provided to beginning teachers prior to the administration of these exams? Should we be concerned if a "Stanley Kaplan" type preparation course could be organized to improve an individual's chances of passing the exam? What are the implications of the misclassification errors of false positives and false negatives? Is the cost in time, human resources, and money worth the additional information gained from this form of assessment? How does this form of assessment impact on minority candidates? Continued exploration of these issues is essential if we are to revitalize the mathematics classroom and that is the ultimate challenge addressed by the reform movement.

References

Carnegie Forum on Education and the Economy. (1986). A nation prepared: Teachers for the 21st century. New York: Author.

Clark, C.M. (1988). Asking the right questions about teacher preparation: Contributions of research on teacher thinking. Educational Researcher, 17 (2), 7-12.

Clark, C.M., & Yinger, R.J. (1979). Teachers' thinking. In P.L. Peterson and H. Walberg (Eds.). Research on Teaching: Concepts, findings, and implications (pp. 231-263). Berkeley: McCutchan.

Cole, N. (1984). Testing and the crisis in education. Educational Measurement: Issues and practice, 3 (3), 4-8.

Darling-Hammond, L. (1986). A proposal for evaluation in the teaching profession. Elementary School Journal, 86(4), 531-551.

Darling-Hammond, L., & Hudson, J. (1986). Indicators of teacher and teaching quality. Center for Statistics, WD-3064-ED.

Grover, B.W. (1989). Develoment and preliminary evaluation of a scoring system for a semi-structured interview for teacher assessment. Unpublished doctoral dissertation, University of Pittsburgh, Pittsburgh.

Grover, B.W., & Zaslavsky, O. (in preparation) Training manual for scoring system for semi-structured interview for secondary mathematics teachers. Pittsburgh, PA: Learning Research and Development Center.

Grover, B.W., Zaslavsky, O., & Leinhardt, G. (1989). An approach to the design and development of a scoring system for a new teacher assessment: The semi-structured interview. (Report No. CLIP-89-02). Pittsburgh, PA: Learning Research and Development Center.

Haertel, E. (1988). Assessing the teaching function. Applied Measurement in Education, 1(1), 99-107.

Holmes Group. (1986). Tomorrow's teachers. East Lansing: Michigan State University.

Leinhardt, G., Weidman, C., & Hammond, K.M. (1987) Introduction and integration of classroom routines by expert teachers. Curriculum Inquiry, 17(2), 135-176.

Leinhardt, G. (1987). Development of an expert explanation: An analysis of a sequence of subtraction lessons. Cognition and Instruction, 4 (4), 225-282.

Leinhardt, G. (1990). Capturing craft knowledge in teaching. Educational Researcher, 19 (2), 18-25.

Leinhardt, G., & Greeno, J. G. (1986). The cognitive skill of teaching. Journal of Educational Psychology, 78(2), 75-95.

Leinhardt, G., & Smith, D.A. (1985). Expertise in mathematics instruction: Subject matter knowledge. Journal of Educational Psychology, 77(3), 247-271.

McLarty, J.R. (1987). A single minded appraisal of a multiple criterion teacher appraisal system. Paper presented at the annual meeting of the American Educational Research Association. Washington, D.C.

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. Washington, D.C.: U.S. Department of Education.

Pecheone, R.L. (1988, April). The catalytic role of new teacher assessment strategies: Designing assessments to measure subject matter pedagogical understandings. Paper presented at the 72nd annual meeting of the American Educational Research Association, New Orleans.

Rudner, L.M. (1987). Content and difficulty of a teacher certification examination. In L.M. Rudner (Project Director), What's happening in Teacher Testing.(pp. 33-38) Washington, D.C. : U.S. Government Printing Office.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. Educational Researcher, 15(2), 4-14.

Shulman, L.S. (1987). Knowledge and teaching: Foundations of the new reform. Harvard Educational Review, 57(1), 1-22.

Tomala, G. (1989, March). Designing semi-structured interviews for statewide assessment. Paper presented at the annual meeting of the American Educational Research Association. San Francisco.

Winters, P. (1990). A study of interviewer behavior and teacher proficiency ratings in a state assessment program for beginning teachers. Unpublished doctoral dissertation, University of Connecticut, Hartford.

Wise, A.E. & Darling-Hammond, L. (1987). Licensing teachers: Design for a teaching profession. Santa Monica, CA: The Rand Corporation (R-3576-CSTP)

List of Figures

Table 1

Generic and Specific Characteristics of the Semi-structured Interview and its Scoring System

	Interview	Scoring system
Generic characteristics	Tasks Questions	Dimensions Components Anchors
Specific characteristics	Stimulus materials 1. Sets of subtopics 2. Textbook pages 3. Teaching approaches 4. Examples of student work	Acceptable answers Justifications Support for justifications

Table 2

Sample Guided Notetaking Page for Task 1

GUIDED NOTETAKING

1. Ordering on the basis of mathematics. (Use the list below to note the order in which the candidate places the topics.)

Order Justification

___ A. using ratios

___ B. proportions

___ C. solv. prob. w/propor.

___ D. meaning of %

___ E. fractions as %

___ F. decimals as %

___ G. finding % of number

___ H. sales tax and interest

___ I. discounts and mark-ups

___ J. percents > 100%

2. Taking students into account. Justification

Familiar to the student

Motivate

Different abilities

3. Grouping topics
Group(s) Justification

Table 3

Sample of Components for Task 1

1.1 Candidate takes into account the major math principles and concepts in planning and organizing the instruction. (CK)

0	2	4	6	8	10
no answer	unacceptable order OR justification NOT based on math structure. or difficulty levels, OR no justification	acceptable order; justification given as general statements, does not include specifics		acceptable order; justifications based on math structure or difficulty levels and includes specifics	

1.2 Candidate takes the students into account in planning and organizing the instruction. (KS)

0	2	4	6	8	10
no answer	general statements; no specifics about how influence order	determines order on basis of only 1 of: familiar to students, motivate students, or abilities AND includes specifics		deter. order on basis of at least 2 of: familiar to stud., motivate students, or abilities AND includes specifics	

1.3 Candidate demonstrates an ability to group topics within the unit and support the groupings with a valid rationale. (CP)

0	2	4	6	8	10
no groups	unacceptable group(s) OR only 1 acceptable group with no justification.	more than 1 acceptable group AND justification vague or not given for all groups.		at least 3 acceptable groups AND justification for each group	

1.4 Candidate justifies additions that can be made to improve the unit. (CP)

0	1	2	3	4	5
no answer	general statements; no justification	1 specific addition AND justification based on either completion of the unit OR adding variety, OR general skills		specific additions AND justification based on at least 2 of completion of ~~the~~ unit, adding variety, or general skills	

Table 4

Sample Task Scores Page for Task 1

TASK SCORES

Candidate____________ Date ________________

Rater _______________ Classification ___________

Classification Criteria

Component	Rating
1.1	_____
1.2	_____
1.3	_____
!.4	_____
1.5	_____
1:.6	_____

HIGH PASS

_____1.1,1.2, &1.3: two > or = 6

_____ No more than one rating of 1

_____ No more than one rating of 2

_____ Adjusted score > or = 31

PASS

_____1.1,1.2, &1.3;two> or = 4

_____ No more than two ratings of 1

_____ Adjusted score > or = 18

Total Component Score	______
Negative Comments Score	______
Positive Comments Score	______
ADJUSTED SCORE	______

Table 5

Reliability of raters for task scores and components as measured by percent of agreement and the Pearson Product Moment Correlation

	Classification		Numerical Score	
	% of Agreement		PPMC	
	> or = 67%	> or = 83%	> or = .70	> or = .80
Tasks	76%	60%	82%	47%
Components	57%	31%	55%	34%

Table 6

Kruskal-Wallis and ANOVA Analyses of Task Scores

		K-W	ANOVA
Task	Topic	T	F
1	Ratio	4.706*	9.906*
	Linear	3.714	3.082
2	Ratio	2.059	.542
	Linear	3.714	3.207
3	Ratio	4.571*	12.789*
	Linear	2.721	2.295
4	Ratio	3.714	2.561
	Linear	3.400	1.626

* $p < .05$

Table 7

Kruskal-Wallis and ANOVA Analyses of Total Dimension Scores

		K-W	ANOVA
Dimension	Topic	T	F
Content	Ratio	4.571*	10.284*
Knowledge	Linear	4.571*	8.521
Content	Ratio	4.571*	102.707*
Pedagogy	Linear	4.571*	3.306
Knowledge of	Ratio	4.571*	5.171
Students	Linear	4.571*	3.291
Basic	Ratio	4.571*	13.938*
Communication	Linear	13.824	4.058

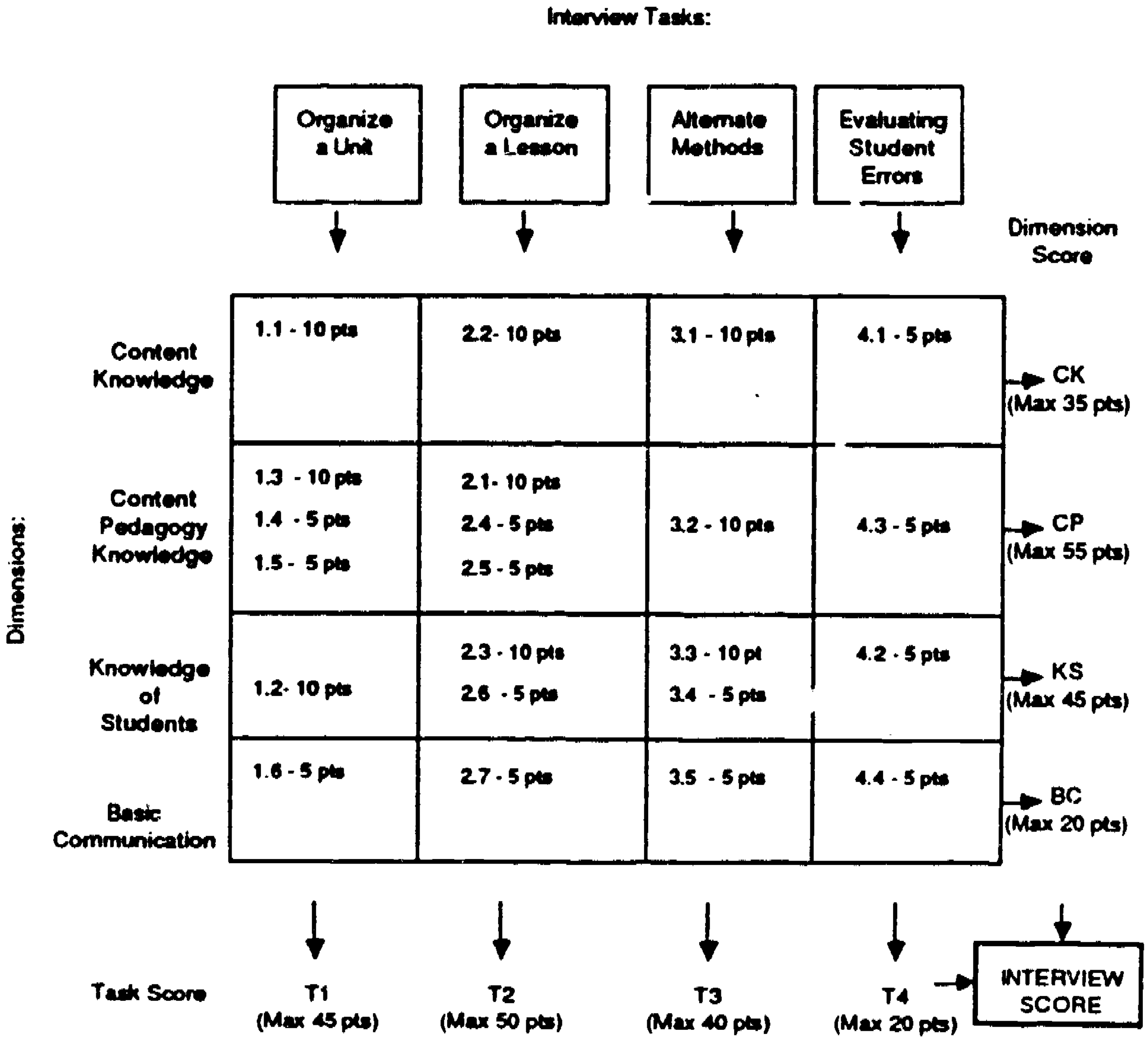
* p< or = .05

		K-W	ANOVA
Interview	Topic	T	F
	Ratio	4.571*	23.938*
	Linear	4.571*	4.953

* p< or = .05

Dimensions	Task 1 (T1) Organize a Unit	Task 2 (T2) Organize a Lesson	Task 3 (T3) Alternate Methods	Task 4 (T4) Student Errors
Content Knowledge	1.1 - Takes math into account	2.2 - Takes math into account	3.1 - Identifies math in approaches	4.1 - Identifies math errors
Content Pedagogy	1.3 - Groups subtopics within unit. 1.4 - Adds to Unit of study 1.5 - Relates subtopics to curriculum	2.1 - Presents overall plan 2.4 - Identifies student math difficulties 2.5 - Checks for comprehension	3.2 - Relates approach to pedagogy	4.2 - Describes remediation strategies
Knowledge of Students	1.2 - Takes students into account	2.3 - Takes students into account 2.6 - Adjusts lesson for different abilities	3.3 - Relates approach to student needs 3.4 - Evaluates approaches for different ability levels	4.3 - States reasons for errors
Basic Communication	1.6 - Is coherent and articulate	2.7 - Is coherent and articulate	3.5 - Is coherent and articulate	4.4 - Is coherent and articulate

Figure 1. Relationship among the tasks, the dimensions, and the components of the scoring system.



Interview Tasks:

Dimensions:	Organize a Unit	Organize a Lesson	Alternate Methods	Evaluating Student Errors	Dimension Score
Content Knowledge	1.1 - 10 pts	2.2- 10 pts	3.1 - 10 pts	4.1 - 5 pts	CK (Max 35 pts)
Content Pedagogy Knowledge	1.3 - 10 pts 1.4 - 5 pts 1.5 - 5 pts	2.1- 10 pts 2.4 - 5 pts 2.5 - 5 pts	3.2 - 10 pts	4.3 - 5 pts	CP (Max 55 pts)
Knowledge of Students	1.2- 10 pts	2.3 - 10 pts 2.6 - 5 pts	3.3 - 10 pt 3.4 - 5 pts	4.2 - 5 pts	KS (Max 45 pts)
Basic Communication	1.6 - 5 pts	2.7 - 5 pts	3.5 - 5 pts	4.4 - 5 pts	BC (Max 20 pts)
Task Score	T1 (Max 45 pts)	T2 (Max 50 pts)	T3 (Max 40 pts)	T4 (Max 20 pts)	INTERVIEW SCORE

Figure 2. Relationship between the framework of the scoring system and the scores produced

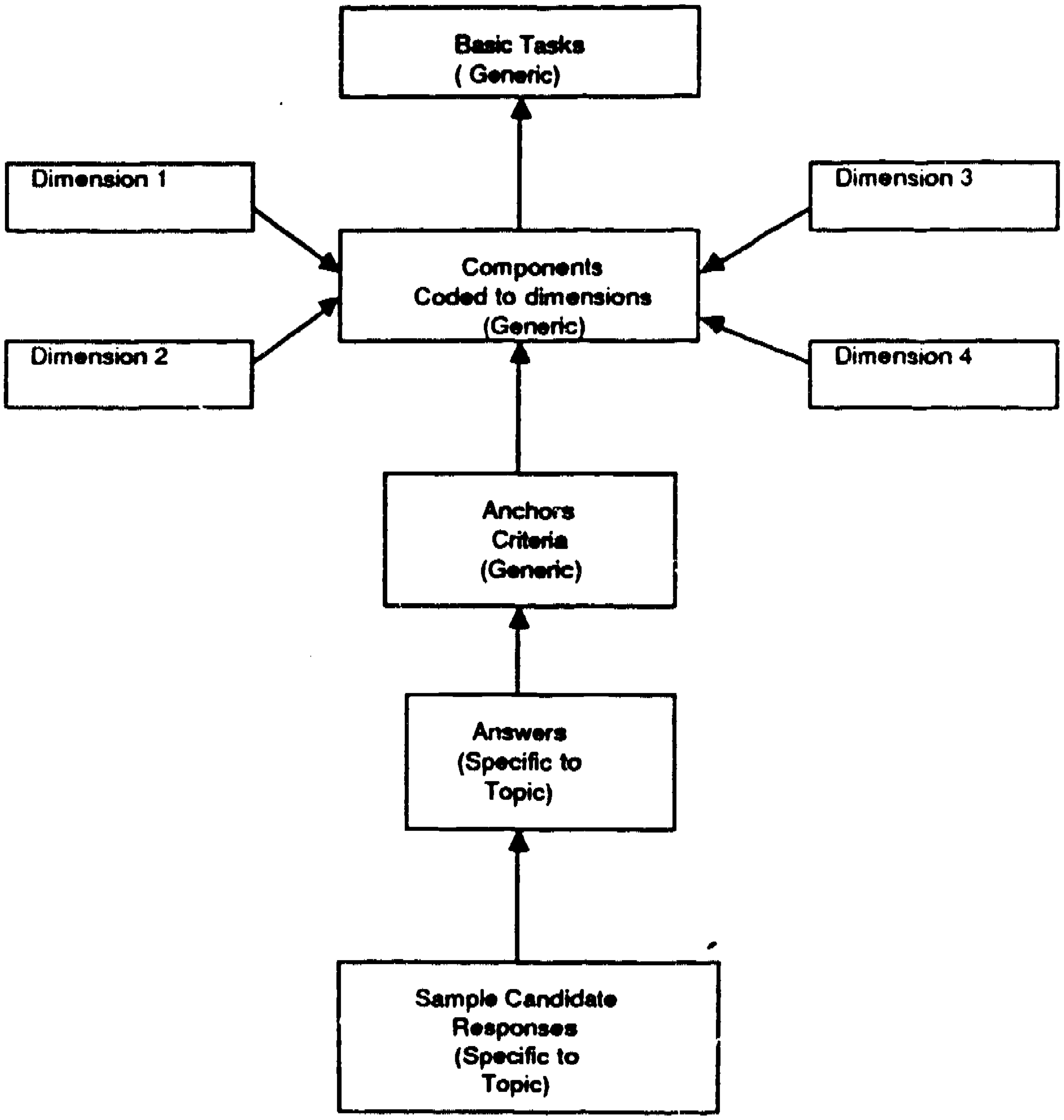


Figure 3. Relationship between grounded specificity and generic tasks.