DOCUMENT RESUME

ED 319 748 . TM 014 832

AUTHOR	Stevenson, Zollie, Jr.; And Others
TITLE	The Reliability of Using a Focused-Holistic Scoring Approach To Measure Student Performance on a Geometry Proof.
PUB DATE	Apr 90
NOTE	82p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Academic Achievement; Cost Effectiveness; Educational Assessment; *Geometry; High Schools; *High School Students; Holistic Evaluation; Mathematics Tests; Multiple Choice Tests; *Performance; *Proof (Mathematics); *Reliability; *Scoring; State Programs; Testing Programs
IDENTIFIERS	End of Course Testing Program; Final Examinations; North Carolina

ABSTRACT

During the Spring of 1989, the North Carolina Department of Public Instruction field-tested a geometry proof performance assessment as a component of the High School End-of-Course Testing Program (EoC). The existing geometry EoC test consisted only of multiple-choice items. The performance assessment was added to the multiple-choice component to form a more "authentic" assessment of student performance. Focus was on determining the reliability of the scoring process and the cost of adding a performance asse .ment to the existing geometry test. The sample for the proofs and the comprehensive geometry multiple-choice test was over 43,000 students, who comprised the entire statewide enrollment in North Carolina in geometry classes. The separate 32-item proofs field test was administered to °75 students. A high degree of consistency was found between the ratings assigned by two readers when perfect and adjacent a'r.ement findings w.re analyzed. The estimated cost of conducti'~ the performance assessment was 3.00 dollars per student (2.44 dollars for school year 1990). The educational significance of using this process for developing authentic assessment strategies is discussed. Eight tables contain study data. Appendices provide North Carolina geometry goals/objectives; rated samples of geometry proofs and the score scale; percentage of students receiving each score on the 1988-89 Geometry Proof Field Test; statistics for proof scores and grades and geometry proof focused-holistic score scale distribution; a summary of teacher assigned proof grades and geometry proof scores; sample school system disaggregate report; and teacher survey data. (SLD)

THE RELIABILITY OF USING A FOCUSED-HOLISTIC SCORING APPROACH TO MEASURE STUDENT PERFORMANCE ON A GEOMETRY PROOF

Prepared by:

Zollie Stevenson, Jr., Ph.D.
District of Columbia Public Schools

Chris P. Averett, Ph.D.
North Carolina Department of
Public Instruction

Daisy Vickers, M.A.
North Carolina Department of
Public Instruction

A Paper Presented at the
American Educational Research Association
Annual Meeting

Boston, Massachusetts

April 1990

The authors:

Zollie Stevenson, Jr., Ph.D.
Director (Acting)
Research and Evaluation Branch
Office of Educational Accountability and Planning
District of Columbia Public Schools*

Chris P. Averett, Ph.D.
Consultant for High School Testing Programs
Testing Section
Research and Development
North Carolina Department of Public Instruction

Daisy Vickers, M.A.
Consultant for Writing and Performance Assessment
Testing Section
Research and Development
North Carolina Department of Public Instruction

* - Dr. Stevenson formerly served as Region 5 Coordinator of Research and Testing with the North Carolina Department of Public Instruction in Greensboro, N.C.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

During the Spring of 1989, the North Carolina Department of Public Instruction field-tested a geometry proof performance assessment as a component of the High School End-of-Course Testing Program (E-o-C). The existing geometry E-o-C test consisted only of multiple-choice items. The performance assessment was added to the multiple-choice component to form a more "authentic" assessment of student performance. Of primary concern in this study were the reliability of the scoring process and the cost of adding a performance assessment to the existing geometry test.

The findings indicated a high degree of consistency between the ratings assigned by two readers when perfect and adjacent agreement ratings are analyzed. The cost of conducting the performance assessment field-test has been estimated at $3.00 per student ($2.44 for SY 1990).

The educational significance of using this process for developing authentic assessment strategies is discussed.

7

The Reliability of Using a Focused-Holistic Scoring Approach to Measure Student Performance on a Geometry Proof

Introduction

While Webb and Romberg (1988) were presenting the new standards for mathematics assessment adopted by the National Council of Teachers of Mathematics (NCMT) at the New Orleans meeting of the AERA, state and local testing departments were already putting together mathematics performance tasks to assess student knowledge beyond the realm of the multiple-choice test. The NCMT proposed that assessment be appropriate and meaningful in facilitating mathematical communication among students. Webb and Romberg (1988) provided examples of age/experience appropriate innovative assessment tasks that emphasized critical thinking and problem-solving.

As with the NCMT, Wiggins (1989) argues for "authentic" assessment to enable educators to have better knowledge of student ability in areas not amenable to multiple-choice assessment techniques. Wiggins (1989) states:

> Do we judge our students to be deficient in writing, speaking, listening, artistic creation, finding and citing evidence and problem solving? Then let the tests ask them to write, speak, listen, create, do original research and solve problems. Only then need we worry about scoring the performance, training the judges and adapting the school calendar to assure through analysis and useful feedback to students about results.

The North Carolina Department of Public Instruction has endeavored to translate the educational reform mandate legislated by state representatives into more meaningful assessment activities. The North Carolina End-of-Course (EOC) Testing Program at the secondary level is an outgrowth of the desire of state legislators to standardize the statewide course of study and the basic educational program offerings to the one hundred thirty-four (134) public school systems in North Carolina. A set of common, or core, items on the EOC tests are used to compare student performance across school systems. School systems are encouraged to use the EOC test scores on the core items as a factor in assigning final course grades for students. Additional items are assessed through those tests for use in evaluating the extent to which school systems are implementing the state-mandated curriculum goals and objectives in each subject area. For example, each Algebra I student takes a 100-item test, of which 60 items are common to all test forms and 40 items represent one of five forms. Therefore, 260 items are measured in each classroom. Some fourteen EOC tests will be put into place by School Year (SY) 1992. Presently, multiple-choice EOC tests have been implemented in Algebra I, biology, Algebra II, U.S. History, geometry, chemistry, physics and English I. When implemented in SY 1992, the English II assessment will have students write essays, some of which will be literature-based.

During 1988, NCDPI field-tested approximately 1,200 multiple-choice geometry test items which measured the mandated standard course of study curriculum goals and objectives (see Appendix A). Eight of the fourteen geometry curriculum goal areas include instruction in developing complete proofs. Traditionally, instruction in proofs has been considered an important objective in the high school curriculum for its focus on the development of logical and precise thinking skills. The Mathematic and Testing Sections of NCDPI determined that the best way to measure student ability to develop proofs is to have the students formulate actual proofs during EOC testing and to have the proofs scored on a common scale. Teachers and curriculum specialists advised that the geometry EOC test would have greater face and content validity if it also contained proofs. The item field test administered in 1988 therefore contained 20 proofs, two each on the ten test forms. A Geometry Advisory Group (GAG) composed of geometry teachers, school system mathematics supervisors, college mathematics teachers, and NCDPI staff was formed to provide guidance and feedback to the Testing and Mathematics Sections on the development of a proofs assessment.

After developing a scoring process for the proofs it was decided to determine the feasibility and reliability of a proofs assessment during a statewide field test during SY 1989. The statewide field test not only afforded an opportunity to assess the administration and scoring of geometry proofs, but also an opportunity for statewide staff development and awareness of the proposed measurement process. Although most EOC tests are administered at the end of the school year, the proofs were administered during the spring, and each student developed two proofs, one common proof and one of four variable proofs, so that five proofs were administered in each classroom.

Objectives

The goals of the statewide field-test were to determine the feasibility of adding a geometry proof to the geometry EOC examination and to determine the reliability of scoring geometry proof performance exercises using the focused-holistic scoring approach. Specifically, this study sought to answer the following questions:

1. What is the agreement rate for two independent readings of geometry proofs? Does the rate vary by type or difficulty of proof? Does the rate vary by scoring location?

2. What is the reliability of proof ratings and proof scores?

3. What are the relationships (predictive validity) between proof scores and geometry grades, geometry proof grades, a multiple-choice proofs test, and a multiple-choice geometry test?

4. What is the cost of a geometry proofs performance assessment?

Literature Review

One of the measurement issues arising during the educational reform movement of the late 70's/early 80's has been the most appropriate method for assessing student ability on authentic performance tasks. Most research on this topic has been done in the area of writing. Most educators have preferred to collect and analyze writing samples as authentic measures of writing ability, but the methodology for reliably assessing the resulting writing samples has raised many questions yet to be resolved.

Rating procedures for assessing direct writing samples are plagued by many sources of error including less than desirable scorer reliability. Some educators have sought to circumvent problems associated with direct assessment by using an indirect approach to measure writing ability: the language expression, total language, or reading vocabulary sections of multiple-choice examinations. While multiple-choice tests generally possess high levels of reliability and validity based, in part, on their conformity to traditional measurement techniques, others believe that direct measures of writing are more concrete and valid indicators of writing ability (in the particular writing domain chosen). The next section will identify some of the issues related to scoring direct performance measures through the research on writing assessment.

Spandel (1981) has summarized many of the issues dealing with methods of rating writing samples and the number of readings required for each. The four most commonly used approaches for rating writing samples are holistic, focused-holistic, analytic and primary-trait procedures. The holistic rating procedure involves assessing a piece of writing to get an "overall" impression of its merits based on a set of predetermined criteria. A range of factors are considered in defining the criteria or overall quality of the writing sample. The focused-holistic approach uses a specific, selected number of predetermined writing characteristics that are selected for "focus" to provide an overall assessment of the quality of the sample based on the selected domain. The key difference in the holistic versus the focused-holistic approaches is the number of characteristics and specificity of criteria included under the "overall" umbrella. The analytic approach, according to Spandel (1981) involves isolating one or more of the pred-defined characteristics of writing and rating each independently. Analytic procedures enable the rater to assess the students' ability to perform the specific skills of writing (e.g., grammar, punctuation, organization and/or style). The primary-trait procedure is similar to analytic in that attention is directed to specific characteristics of a writing product; however, this procedure endeavors to quantify the amount of the characteristic present in determining the appropriate rating to be assigned.

The process of rating a writing sample varies with the scoring approach used (Spandel, 1981). With holistic or focused-holistic approaches, a single score is assigned to a writing sample based on the overall impression of the rater using predetermined criteria and performance levels after <u>one</u> reading. Thirty to forty papers can be read in one hour using this approach. In analytic and primary-trait procedures, each factor is evaluated independently of other factors. If four factors were being considered, every essay would need to be read four times so that the strength of each factor could be judged independently.

Spandel (1981) concluded that the benefits of the rating method used depends on the purpose for evaluating student writing. If the goal is to identify overall quality of writing, then a holistic measure is the strongest indicator of ability in the writing domain chosen. Ratings from analytic and primary-trait approaches are more useful in directing instruction.

The most comprehensive review of research on standardized systematic assessment of writing ability was conducted by Cooper (1984) for the Educational Testing Service (ETS). Cooper's thorough review considered the nature and limitations of essay and multiple-choice tests of writing ability, the statistical relationships of those types of tests, disaggregate performance indicators and the comparative cost effectiveness of various types of writing assessment.

Cooper (1984) indicated that direct measures of writing are subject to lower reliability than are indirect measures because of the subjective nature of the ratings and the procedures used to assess writing samples. Conventional statistical approaches may fail to disclose differences based on changes in factors such as the rating standards used, flatness of the writing sample and time/order of rating. Essay ratings independently assigned by two to four readers (raters) were considered to be more reliable than those assigned by a single rater. Any rating assigned is only as good as the training provided to the raters and the strength of the guides used to anchor score points.

Cooper (1984) reported that the reliability of ratings assigned by experienced, rested (fresh) raters is subject to variation from one rating to the next depending on the quality of writing being scored. Errors are more likely to be counted in poor rather than more skillfully produced essays. Interesting essays are more likely to receive higher ratings than minimally adequate but uninteresting essays. Further, the quality of preceding essays is likely to impact on the ratings assigned to subsequent essays. Fatigue also impacts on scores assigned. A tired rater can become more lenient, stricter or erratic in the scores assigned for a given essay depending on the level of fatigue experienced. The length of the reading period across a day or number of days can result in lower scores. Ratings done on the first day of a multiple period of readings tend to be higher than ratings assigned towards the end of a multiple day reading period (Cooper, 1984).

A study of interrater agreement by Myers, McConville and Coffman (1966) across five days of ratings indicated that while the average daily rating correlations for readers across all papers over five days was .406, the average correlation on the fifth day was only .264. Among the conclusions of Myers et al. (1966) is that lack of vigilance exists as the end of any arduous task approaches no matter how little time has been allocated to the scoring process. More recently, North Carolina and other states have found that reader agreement increases over time, testifying to the effectiveness of the monitoring of reader reliability that is common in statewide assessments.

Conlan (1980) indicated that for any reading of essays, some effort must be made to control variables such as the number and length of rest breaks; rules for off-topic papers; and a system for handling unique or emotionally evocative papers.

Cooper (1984) and Breland et al. (1987) reported that reliability estimates of essay scores increase with the number of topics and/or readings per topic. "If each essay receives two independent readings, the rater reliabilities are .65 for three topics, .55 for two topics and .38 for one topic (the latter being the most common)." Score reliabilities for single topics read one time ranged from .361 to .411. Breland et al. found that adding more topics contributes more to estimated reliability and predictive validity than adding additional readings.

Research on the raters of performance tasks have been completed by Blok (1985), Cooper (1984), Breland and Jones (1982) and others. Blok (1985) studied multiple ratings obtained by having different raters and the same raters repeat the judging of essays. Testing the theory of rater equivalence, that the "true" scores of one rater will correlate perfectly with the true scores of another rater, was the goal of the Blok (1985) study. Sixteen elementary school teachers rated one hundred five (105) essays on a scale from 1 (very poor) to 10 (excellent) using a holistic approach without providing training to the scorers. A second rating was made on the 105 essays by the same 16 teachers three months later. The four rater equivalence theories were testing using tests of linearity and the method of linear structural equation model. In terms of the equivalence ratings of the essays used in the Blok study (1985), the ratings of different raters were essentially different measures (with rater correlations ranging from .415 - .910).

The rater variable is considered by Cooper to be a major and "unique" source of measurement error in direct assessment. Cooper (1984) indicated that scoring inconsistencies become more pronounced when dealing with a group of readers. Some of the inconsistences are the result of random error while other inconsistencies are systematic based on differences within the groups assigning ratings. Breland and Jones (1982) indicated that inexperienced readers assigned higher ratings than did experienced readers. Further, even when "experienced" English teachers agree on scoring criteria and standards, they do not agree on the extent to which any one of the criteria ought to be applied in any given circumstance even with training. Coffman (1977) added that inexperienced raters are reluctant to assign rating scores that are too high or too low so their scores tend to cluster around the middle. As a result, the matter of which rater assigns a score to a writing sample can make a difference in the scores of poor and good essays.

Cooper reported (1984) that " over and over it has been shown that there can be wide variance between the grades given to the same essay by two different readers, or even by the same reader at different times". Coffman (1971) indicated that ratings by a single pair of raters may result in excessive overestimations of reliability. This finding differed from Cooper and Odell (1977) who estimated 90% reliability between pairs of raters under controlled conditions. They cited rigid control of the training protocols as the single most important factor in producing consistency (Cooper and Odell, 1977).

Hughes and Keeling (1984) studied the effect of offering model essays as training devices to eliminate context effects for persons preparing to score writing samples. Essays were collected from thirty-eight (38) students on a topic determined by the researcher. The essays were typed, retaining the original errors, and formed into twenty-five (25) booklets (randomly ordered). Each set of essays was scored on a 0-25 point scale by a cadre of experienced English teachers using predetermined criteria. The same five good, five poor, one criterion and four filler essays were selected for placement in good or poor context booklets. In the good context booklet, the five good essays were randomly placed first, next the criterion essay and last the filler essays randomly arranged with the poor essays. In the poor context booklet, the poor essays were randomly arranged in the first section of the booklet and the good essays were randomly arranged at the end of the booklet. Three annotated model essays (one good, one average and one poor) were used as models for the trained group. A set of five essays were scored by each group and served as the covariate group for the study. The findings were that context effects existed even after efforts were made to eliminate them. Hughes and Keeling (1984) concluded that contextual factors were likely to persist in performance assessments where writing samples or non-factual responses were being assessed. However, they held out hope that contextual factors would be less evident in performance measures that dealt with factual answers.

The literature presented leads to the following conclusions. Ratings assigned to a writing sample by scorers untrained in the interpretation of the criteria are less reliable than ratings assigned by trained scorers. Adequate training, frequent rest breaks and pre-established rules for rating on the off-topic responses improves the consistency of the ratings assigned by scorers. Rules for recalibrating raters and monitoring the consistency of ratings assigned also improves the reliability. Some issues, such as the number of raters that should read each paper, the best approach for scoring writing samples and the appropriate statistical procedures to use in determining reliability/validity, have not been resolved to the satisfaction of many "experts". However, it does appear that rule setting and training are key factors required to improve the consistency of the ratings assigned to writing samples no matter what scoring approach is used.

Methodology

Development of Scoring Process

During the summer of 1988 the Geometry Advisory Group (GAG) reviewed student responses to sample geometry proofs developed by by selected geometry teachers from across the state. The proofs had been field-tested during the late spring of 1988 in selected schools across North Carolina. Borrowing from successful performance task measurement applications in writing, the scoring approaches considered were analytic and focused-holistic.

Initially, the GAG favored the analytic approach since specific errors could be marked. After a period of study and some calculations, the GAG found that the analytic approach could result in as many as twenty-seven (27) different scoring guides for one proof! This was partly due to the fa that the GAG strongly felt that any proof administered statewide should be vable from multiple approaches. The GAG recommended use of the focused-holistic approach to assess student performance on geometry proofs for four reasons: 1. the ability to develop a single scoring guide containing various strategies for solving the proofs amenable to the focused-holistic scoring approach; 2. the belief that training of scorers could better be accommodated using the focused-holistic approach; 3. the efficiency and speed of focused-holistic scoring; and 4. the previous success of the focused-holistic approach with writing assessment in North Carolina. The group determined the scoring criteria and score point descriptions to be used in the statewide field test. The score scale contained five (5) points with a four (4) reflecting a nearly perfect proof and a zero (0) reflecting a blank or completely erroneous proof (see Appendix B).

The focused-holistic scoring process has been widely used in the assessment of writing (e.g., Cooper, 1984; Stevenson, 1988). This approach considers the overall sense of completeness of the writing sample based on a pre-determined set of criteria. To appropriately use the focused-holistic approach, agreement has to be made on the criteria to be used and characteristics of the criteria at each score point level. Each individual data sample is read and evaluated (using the established criteria and score points) by one or more raters. Cooper (1984) has reported scorer reliability in writing assessment-related studies for two or more readers ranging from .41 to .89 varying with the background and length of training on the scoring criteria. The NCDPI has reported perfect scorer agreement rates of more than 70%, adjacent agreement rates (no more than one point difference between the two ratings) approaching 30% and less than .7% differing by more than one point on a four-point focused-holistic score scale.

Staff Development and Awareness Training

NCDPI Mathematics and Testing Coordinators based in the eight regional centers were trained to use the scoring guides during November 1988. These Coordinators then conducted regional Geometry Proof Awareness Session(s) attended by at least one (1) geometry teacher from each school in the region that provided geometry instruction during December of 1988 and January of 1989.

Participants were trained to score the sample proof and informed of the logistics of the spring 1989 statewide field-test. After developing skill in scoring practice sets of geometry proofs, participants received scoring guides, training practice sets and scoring keys for use in training other geometry teachers and exposing students to the test format, scoring process and assessment criteria prior to the actual field-test administration. Many of those same geometry teachers returned during late April to score the geometry proofs of students from their regions.

Data Collection

Proofs were administered to more than 43,000 geometry students during the period March 20 - April 7, 1989 on scannable 11x17 folded answer documents. Student identification information was printed on page 1, directions on page 2, the common proof exercise on page 3 and one of four variable proofs on page 4. The four different forms, identified by the four variable proofs, were printed in different ink colors for ease in identification during scoring. The four forms were spiraled throughout each classroom.

The central office Test Coordinator for each of the school systems collected completed geometry proof field-tests and forwarded them to the Regional Research and Testing Coordinator. Proofs were shipped from regional centers to an outside contractor so that identifiable data linking a proof with a specific student, school or school system were removed to eliminate those factors as potential sources of scorer bias and to allow scoring packets to be developed. The outside contractor separated the proofs into four form/color groups, printed packet identification sheets, and stapled these sheets to the proof sets. Four scannable monitor sheets for recording independent scores were produced for each packet. The first two listed proofs in the order they were in face up in the packets. The other two listed the proofs in reverse order for use in scoring the variable proofs on the reverse side. The four sheets were inserted into each packet along with the proofs.

During scoring reader one removed monitor sheet number one from the packet, recorded a reader identification number, and verified that the proofs in the packet matched the proof identification numbers on the monitor sheet. After scoring the proofs, reader one returned the proofs to the packet envelope, which still contained the other three monitor sheets, and placed the monitor sheet used on top of the packet. NCDPI staff retrieved completed packets, reviewed the monitor sheet, and randomly re-circulated the packet to a second reader. The completed monitor sheets were then scanned on NCS Sentry 3000 tabletop scanners connected to an IBM personal computer. Data were stored on floppy diskettes using a software program developed by NCDPI. In addition, reader reliability reports were generated to monitor reader agreement and progress in scoring. Highly discrepant readers were retrained and proof scores requiring resolution (discrepant by more than one score point) were identified and resolved on the spot by specially trained scorers.

Data on the diskettes produced in the eight regions were merged at the NCDPI state testing office with student background information provided by the outside contractor on data tape. Rosters of student scores were returned to geometry teachers prior to the end of the school year so that final scores could be coded on answer sheets for the EOC multiple-choice geometry test. All EOC tests were scored at a high school or central office site in each school system and rosters of scores were produced for use in assigning grades to students. (Geometry grade rosters included core multiple-choice and common proof scores.) Data diskettes with all EOC scores were forwarded to the Regional Testing Coordinator for final (in-region) editing and shipment to the Testing Section where the statewide report was prepared (see Appendix F). Summary reports of proof scores were generated in this fashion for 43,926 secondary school students.

Five multiple-choice proofs were also field-tested in selected sites during May 1989. The multiple-choice proofs paralleled proofs administered during March-April 1989.

Reader Training and the Scoring Process

Each school system provided a minimum of one (1) geometry teacher from each school where geometry was taught to participate in the regional scoring process. In the three largest regions, two full day scoring sessions were held. All of the regional sessions were held on school days. School systems paid substitute teacher expenses for geometry participants from their school systems. Participating teachers received certificate renewal credit for their participation. The number of teachers involved in the scoring process ranged from twenty-nine (29) in the smallest region to ninety-one (91) in one of the larger regions.

On the first day of scoring, geometry teachers selected by the school systems were trained by one of three scoring directors to use the common proof scoring guide. Next, teachers scored three (3) training sets of 10 to 15 proofs (35 total) each that provided exposure to the scoring characteristics and distinctions between each score point and the variability within score points. Finally, teachers took a qualifying exam with 70% accuracy (perfect agreement with the designated score point) required in order to be eligible to score common proofs. Teachers who fell below the criterion were re-trained and administered additional qualifying exams. The entire training and qualifying process took approximately three hours. Statewide, only 1% of the 423 readers failed to qualify to score.

Common proofs were read twice. The first reader assigned a rating (0 - 4) to the common proof based on the criteria (see Appendix B for the score scale). Ratings for two readers that differed by one point were averaged (e.g., if rater one assigned a rating of "1" and rater two assigned a rating of "2", a score of 1.5 was assigned as the proof score). When ratings assigned by a reader differed by more than one (1) point, a staff person from the NCDPI Mathematics Section or a participating GAG member read the proofs with discrepant scores and assigned the final score.

After scoring of the common proof was complete, section leaders trained participants to score the variable proofs. Section leaders were either members of the GAG or teachers from the winter awareness sessions that demonstrated skill and understanding of the rating process as observed through the winter practice exercises. Each section leader provided training on specific characteristics of their variable proof. Each variable proof was initially read once with second reading occurring only if time permitted. Scores from the variable proofs were not considered by teachers in evaluating student course performance.

Sample

The sample for the proofs and the comprehensive geometry multiple-choice test was the entire statewide enrollment in geometry classes in North Carolina, more than 43,000 students. The separate, 32-item multiple-choice proofs field test was administered to a convenience sample of 875 students from schools in each of the eight educational regions.

Measures

Each student completed two proofs, one common and one of four variable proofs (see Appendix B for the proof exercises). Each common proof was scored twice. Most variable proofs were scored once, but a substantial portion were scored twice. This resulted in four reader scores for each student, and as many as ten reader scores across the sample of five geometry proofs. For the common proof, and when possible for the variable proofs, scores were combined to produce composite scores. All students also took a comprehensive multiple-choice geometry test at the end of the year. In addition, a sample of students took a multiple-choice test focusing on the same five proofs. Six to seven items were specific to each of the five proofs. Teachers recorded the final course grade they expected to give each student at the end of the year, the course grade as of the proofs assessment in the spring, and a grade assessing the student's proofing skill at the time of the proofs assessment. The following list gives all the variables used in this study. The N counts in parentheses are the counts for all analyses using these variables. When two or more variables are related with differing N counts, the analysis is based on the lower of the two N counts.

Focused-holistic scores of proofs on a scale of 0 (low) to 4 (high):

1. Common Proof, Reading 1 (N=43,926)
2. Common Proof, Reading 2 (N=43,926)
3. Variable Proof A, Reading 1 (N=11,177)
4. Variable Proof A, Reading 2 (N=2,773)
5. Variable Proof B, Reading 1 (N=11,017)
6. Variable Proof B, Reading 2 (N=5,612)
7. Variable Proof C, Reading 1 (N=10,925)
8. Variable Proof C, Reading 2 (N=4,951)
9. Variable Proof D, Reading 1 (N=10,807)
10. Variable Proof D, Reading 2 (N=4,304)

Focused-holistic score composites

11. Common Proof: (1+2)/2 (N=43,926)
12. Variable Proof A: (3+4)/2 (N=2,773)
13. Variable Proof B: (5+6)/2 (N=5,612)
14. Variable Proof C: (7+8)/2 (N=4,951)
15. Variable Proof D: (9+10)/2 (N=4,304)

Multiple-choice test scores

16. NC Test of Geometry: score on 60-item core test (N=43,325)
17. Multiple-choice proofs test: score on 32-item test of same 5 proofs (N=875)
18. Multiple-choice common proof test: score on 6-item subtest (N=875)
19. Multiple-choice variable proof A test: score on 7-item subtest (N=217)
20. Multiple-choice variable proof B test: score on 6-item subtest (N=221)
21. Multiple-choice variable proof C test: score on 6-item subtest (N=218)
22. Multiple-choice variable proof D test: score on 7-item subtest (N=219)

Instructor's ratings

23. Course grade in geometry at end of the year (N=43,067)
24. Grade in geometry at time of proofs assessment (N=43,400)
25. Grade in proofing skill at time of proofs assessment (N=43,103)

Results

Agreement Rates

Tables 1 and 2 summarize the reader agreement rates for the geometry proof field test. On the common proof, approximately 66% of the proofs received the same score on two different readings. Adjacent agreement, or the percentage of proofs receiving scores within one point of each other, was 30.7%, and 3.4% of the common proofs received scores differing by more than one point and were "third read" by a specially-trained scorer.

Agreement rates on the other proofs varied somewhat by type of proof, from a low of 65.6% to a high of 80.6% perfect agreement. The highest agreement rates occurred for the three-dimensional proof (variable proof B) and the parallel line proof (variable proof C), both of which were difficult, with almost 60% of the scores either a 0 or 1. In addition, variable proof C appeared to be one where students "either knew it or they didn't". Although there were a large number of 0 and 1 scores, students also received a relatively large percentage of 4 scores. Lower agreement rates were evidenced on proofs with the largest percentage of scores in the 1 to 3 point range, i.e. in the middle of the score scale, where it is usually more difficult to score accurately.

Table 1

Percentage Agreement Between Two Readings of Geometry Proofs

Proof	Perfect Agreement	Adjacent Agreement*	Difference Requiring Resolution
Common	65.9%	30.7%	3.4%
A	68.1%	30.3%	1.6%
B	73.8%	25.7%	0.4%
C	80.6%	18.4%	1.0%
D	65.6%	32.2%	2.1%

*1 point difference

Table 2 gives the agreement rates for the various scoring sites. The perfect agreement rates ranged between a low of 62.8% to a high of 68.3%, with 5 of the 8 sites having rates of approximately 65 to 66%. Different readers were involved at each site, and a total of three different scoring directors/trainers were used across the eight sites.

Table 2

Percentage Agreement Between Two Readings of Common Proof for Each Scoring Site

Scoring Site	Perfect Agreement	Adjacent Agreement*	Difference Requiring Resolution
1	65.4%	31.6%	3.1%
2	65.9%	30.6%	3.5%
3	67.3%	29.7%	3.0%
4	62.8%	32.9%	4.3%
5	65.0%	31.8%	3.3%
6	68.3%	28.7%	3.0%
7	65.1%	31.6%	3.3%
8	66.0%	30.3%	3.7%

*1 point difference

Correlational Estimates of Reliability

A common method of assessing essay scoring reliability is to correlate the scores assigned by different readers to the same essay. As noted by Breland et al. (1987) these estimates are inflated because they reflect only one source of error. Table 3 gives the correlations between two independent scorings of the same proof for the common proof and the four variable proofs. These estimates of the reliability of one reading of each proof range between .822 and .948. Breland et al. (1987) also point out that these estimates can be "stepped up" using the Spearman-Brown formula to obtain estimates of the reader reliability with two readings of each proof.

Table 3

Correlational Estimates of Reader Reliabilities

Proof	r	Reader Reliability Estimate for 2 Readings
Common	.871	.931
A	.869	.930
B	.822	.902
C	.948	.973
D	.854	.921

Since students responded to two proofs each, reliability estimates can also be calculated by correlating the scores on each proof. Table 4 gives these correlations for one reading, two readings of the common proof and one reading of the variable proof, and two readings of each proof. The three correlations per proof combination give the reliabilities of giving one proof under the three different scoring conditions. The estimates for one proof read once range from .522 (Common vs. B) to .627 (Common vs. A), with an average of .590 across proof types. As would be expected, as the number of readings increases, the reliability estimates increase slightly.

The last column gives reliability estimates for two proofs obtained using the Spearman-Brown formula. These estimates demonstrate that the reliability of a proof assessment can be increased dramatically with the addition of another proof.

Table 4

Correlational Estimates of Reliabilities of Proofs Receiving One Reading or Two Readings

Proofs	Number of Readings	r	Reliability Estimate for Two Proofs
Common vs. A	1	.627	.771
	2 common	.649	.787
	2 each	.664	.798
Common vs. B	1	.522	.686
	2 common	.538	.700
	2 each	.564	.721
Common vs. C	1	.619	.765
	2 common	.642	.782
	2 each	.649	.787
Common vs. D	1	.590	.742
	2 common	.610	.758
	2 each	.634	.776
Average	1	.590	.741
	2 common	.610	.757
	2 each	.628	.771

Predictive Validity of the Proof Assessment

The analysis below gives the predictive validities of both the proofs performance assessment and two multiple-choice tests. A total of five outcomes can be analyzed for the proofs themselves:

1. course grade--grade instructor expects to give the student, reflecting overall geometry performance, not just proofing ability;

2. proofs grade--instructor judgement of proofing skill at time of assessment ;

3. MC proofs test--32-item multiple-choice test of same five proofs;

4. MC proofs subtest--6 to 7-item subtest of MC proofs test of the same proof as solved by the student;

5. MC geometry--60-item test covering entire geometry course content.

Table 5 gives the correlations of the various proofs with the five outcome variables for one and two readings of each proof. As would be expected, the validity estimates are slightly higher for the scores based on two readings, reflecting their higher reliability. When two proof scores are combined (one reading) between .05 and .10 is added to the predictive validity related to proofs grades. Reading the proofs twice adds negligibly to the correlations. Also as expected, the correlations are somewhat higher when related to grades in proofing skill rather than overall geometry performance.

Correlations with the multiple-choice proofs test are generally higher than those with the proofs grade, reflecting the higher reliability of the 32-item test than of teacher judgements about proofing skill. The subtest scores are for the multiple-choice items that relate to the same common or variable proof.

Table 5

Predictive Validities of Proof Assessments

Proof	Number of Readings	Correlations with: Course Grade*	Proofs Grade	MC Proofs Test	MC Proofs Subtest	MC Geometry
Common	1	.511	.585			
	2	.528	.603	.655	.482	.625
A	1	.519	.614	.644	.561	.614
	2	.534	.629			
Common+A	1	.570	.675			
	2	.580	.683			
B	1	.480	.548	.557	.457	.573
	2	.507	.574			
Common+B	1	.579	.658			
	2	.596	.674			
C	1	.542	.632	.650	.651	.686
	2	.554	.644			
Common+C	1	.589	.685			
	2	.598	.693			
D	1	.521	.622	.577	.530	.640
	2	.543	.647			
Common+D	1	.576	.682			
	2	.593	.698			

*Obtained at time of geometry proof assessment.

Note: Missing cells are due to the fact that the datasets containing the multiple-choice tests include only the final scores on the geometry proofs, which were the combined readings for the common proof and one reading for each of the variable proofs.

Table 6 displays the distribution of scores for students who participated in the multiple-choice proofs field test. More than half of the students who could not complete a proof at all on the performance assessment (scores of 0 through 1.0) got 4 to 5 of 6 items correct on the same proof in a multiple-choice (completion) format, and almost one-quarter received perfect scores in this format that requires only recognition, rather than recall and production. These results are somewhat confounded by the fact that the students had already responded to this proof at the statewide administration several months earlier. However, most of the teachers in the multiple-choice proofs assessment reported that they had not reviewed the proofs after the spring administration.

Table 6

Frequency Distribution of Focused-Holistic Scores on the Common Proof and Multiple-Choice Scores on the Same Proof Topic

Multiple-Choice Proof Scores	Focused-Holistic Proof Scores				
	0-1.0	1.5-2.0	2.5-3.0	3.5-4.0	Totals
0-1	2.8%	0.0%	0.0%	0.0%	1.0%
2-3	17.5%	10.0%	0.6%	1.0%	8.7%
4-5	56.3%	57.1%	39.2%	22.6%	45.0%
6	23.4%	32.9%	60.2%	76.5%	45.3%
Totals	36.5%	19.4%	20.7%	23.2%	100.0%

N=875

Table 7 gives the correlations between the two multiple-choice tests and the two instructor ratings. The 32-item multiple-choice proofs test correlated .527 with proofs grades, while the correlations between the performance-based proofs and proofs grades ranged between .574 and .698, depending on the type of proof, the number of readings, and the number of proofs (see Table 5). If differences in reliability of the tests were taken into account, the difference in predictive validity, using proofs grade, would be even greater.

Table 7

Predictive Validities of Non-Essay Assessments

Multiple-Choice Test	Course Grade	Proofs Grade
Geometry	.639	NA
Proofs	.406	.527

Finally, Table 8 gives the multiple correlations (R) when the proofs performance test is combined with the multiple-choice tests to predict grades. Note that the proofs assessment adds between .02 and .05 to the predictive validities.

Table 8

Predictive Validities Combining Common Proof Score and Multiple-Choice Components

Multiple-Choice Test*	Course Grade	Proofs Grade
Geometry	.659	NA
Proofs	.434	.577

*Test combined with focused-holistic proof score.

Feasibility of Statewide Proofs Performance Assessments

Analyses of the Testing Section of the NCDPI indicate that the specific cost to the state of North Carolina for conducting the geometry proof field test was approximately $3.00 per student. This cost includes a curriculum consultant, materials, development, training, scoring, and report generation. Excluded are costs for travel, some facilities expenses, and the salaries of staff of the NCDPI. For SY 1990 the cost is estimated at approximately $2.44 per student.

The statewide field test demonstrated that student responses to proof problems can be scored and reported in a reliable and relatively cost-efficient manner. The logistics developed for this aspect of the EOC testing program are quite feasible, and could be generalizable to other statewide performance efforts.

Discussion

The results of this study indicate that scorer reliability on proof ratings was high, as demonstrated both by the agreement rates and the correlational reliability estimates for the common proof and the variable proofs which were scored twice. This finding is of particular interest since the scoring involved over 400 raters reading proof papers distributed among eight different scoring sites. Five of the scoring sites (1, 2, 4, 7, and 8) are largely rural yet the reader agreement rates for the rural sites were similar to those in the large urban centers. The consistency of scoring across sites and different groups of readers is testimony to the clarity of the scoring guide, the consistency of the scoring criteria, and the willingness of teachers selected as scorers to accept the scoring process. Actual time devoted to training was less than three hours during which time the scoring guide was reviewed, three sets of proofs were scored and discussed, and qualifying rounds were held. The context effects were minimized, resulting in more reliable scoring (Keeling and Baker, 1985).

Further analyses of the relationship between scores on the two proofs each student took indicates that the addition of one extra item contributed dramatically to the overall reliability of the proofs test, beyond that of scorer reliability alone. Except for the addition of training time on the second proof, the cost of scoring one proof twice is similar to the cost of scoring two proofs once, and may be more cost-effective due to the increase in overall reliability. In this assessment students could not receive scores based on two proofs because the students took one of four variable proofs which differed in overall difficulty. The primary purpose of the variable proofs was to provide broader curriculum coverage by assessing five proofs in every schools.

The relationship between the proof scores and other measures of proof performance and geometry performance indicate that the performance-based proofs measures are valid indicators of proofing ability. Furthermore, the performance-based proof scores were more highly related to grades in proofing skill than were the scores on a multiple-choice completion test format. This finding lends support to the subjective impression of many educators that performing actual tasks are more valid the multiple-choice tests related to those tasks. Not only do these tasks have face validity, but a degree of predictive validity as well.

In a time of performance or outcome based accountability systems and measurement-driven instruction, the measurement of skills in a more authentic, performance-based, manner takes on additional meaning. Suhor (1985) reported that both a poll of 350 language arts supervisors and research data indicate that writing instruction decreases with objective tests, and increases where direct writing assessments are implemented. This has certainly been the experience in North Carolina. One purpose, therefore, of performance-based assessments like the proofs test is that they encourage certain types of instruction. This field-test demonstrated that alternative strategies for assessing student ability can be implemented in an objective and reliable fashion. The findings were that teachers can be trained to score proofs with a high degree of consistency. The approach used by North Carolina to measure student performance on a geometry proof could be adapted for use in other subject and skill areas, and by any school system.

References

Blok, H. Estimating the reliability, validity and invalidity of essay ratings. Journal of Educational Measurement, 1985, 22,1, 41-52.

Breland, H.M., Camp, R., Jones, R.J., Morris, M.M., and Rock. D. A. Assessing Writing Skill. New York: College Entrance Examination Board, 1987.

Breland, H.M. and Jones, R.J. Perceptions of Writing Skill. New York: College Entrance Examination Board, 1982.

Coffman, W.E. Essay examinations. In R.L. Thorndike (Ed.), Educational Measurement, 2nd ed., Washington , D.C.: American Council on Education, 1971.

Conlon, G. Comparison of analytic and holistic scoring techniques. Princeton, N.J.: Educational Testing Service, 1980.

Cooper, C.R. and Odell, L. (Eds.) Evaluating Writing: Describing, measuring and judging. Urbana, IL.: National Council of the Teachers of English, 1977.

Cooper, P.L. The assessment of writing ability: A review of research. GRE Board Research Report GREB No. 82-15R, ETS Research Report 84-12, May 1984.

Hughes, D.C. and Keeling, B. The use of model essays to reduce context effects in essay scoring. Journal of Educational Measurement, 1984, 21,3, 277-82.

Myers, A., McConville, C. and Coffman, W.E. Simplex structure in the grading of essay tests. Educational and Psychological Measurement, 1966, 26, 41-54.

North Carolina Department of Public Instruction. Report of Student Performance-1988-89: Writing, Grades 6 & 8. Division of Research and Development, Testing Section, January 1989.

Noyes, E.S. Essays and objective tests in English. College Board Review, 1963, 49, 7-10.

Suhor, C. Objective tests and writing samples: how do they affect instruction in composition. Phi Delta Kappan, 1985, May, 35-39.

Spandel, V. Classroom applications of writing assessment. Clearinghouse for Applied Performance Testing, Northwest Regional Educational Laboratory, Portland, OR., 1981.

Stevenson, Z.J. The relationship of training to the ratings assigned by teachers to student writing samples. District of Columbia Public Schools, Division of Quality Assurance and Management Planning, Washington, D.C., 1988.

Webb, N. and Romberg, T. Implications (of) NCTM Standards for Mathematics Assessment. Paper presented at the American Educational Research Association Annual Meeting, New Orleans, 1988.

Wiggins, G. A true test: Toward more authentic and equitable assessment. Phi Delta Kappan, 1989, May, 703-13.

Appendix A

(North Carolina Geometry Goals and Objectives)

NORTH CAROLINA E[illegible]
GEOMETRY GOA[illegible]

Goal 1: The learner will state the characteristics of sets of points.

1. Identify and name sets of points, such as line, ray, segment and plane.
2. Draw representations of points, lines, and planes.
3. Identify and name unions and intersections of sets of points.
4. Find the coordinate of a point on a line.
5. Find the length of a segment.
6. Identify congruent segments.
7. Identify the midpoint of a given segment.
8. Use a protractor to find the measure of an angle.
9. Determine when two angles are congruent.
10. Identify interiors and exteriors of geometric figures.
11. Identify the bisector of an angle.

Goal 2: The learner will use the structural properties of the real number.

1. State and use the properties of equality.
2. State and use the properties of inequality.

Goal 3: The learner will develop geometric proofs.

1. Translate a geometric statement into an "If-Then Statement".
2. State the converse of a conditional statement.
3. State the hypothesis and conclusion for a conditional statement.
4. Use the process of deductive reasoning in mathematical and non-mathematical situations.
5. Write a proof using the two-column format.
6. Write an indirect proof.

Goal 4: The learner will use some of the properties of angles and lines to develop proof and solve exercises.

1. Use three letters, a number, or a single letter to name an angle.
2. Classify an angle.
3. Identify adjacent and vertical angles.
4. Determine the complement and supplement of a given angle.
5. Apply the Angle Addition Postulate.
6. Apply the Segment Addition Postulate. (Definition of Betweenness)
7. Recognize congruent angles.

Goal 5: The learner will recognize perpendicular lines and planes and use this information to complete proofs and exercises.

1. Apply definitions of perpendicular lines and planes.

ND-OF-COURSE TESTING
S AND OBJECTIVES

Goal 6: The learner will recognize parallel lines and planes and use this knowledge to complete proofs and exercises.

1. Identify parallel lines and planes, and skew lines.
2. Identify corresponding angles and alternate interior angles which are formed when two parallel lines are cut by a transversal.
3. State conditions under which lines are parallel.
4. State which angles are congruent when two parallel lines are cut by a transversal.
5. Identify which angles are supplementary when lines are cut by a transversal.

Goal 7: The learner will identify polygons and complete proofs and exercises related to them.

1. Classify a triangle according to its sides.
2. Classify a triangle according to its angles.
3. Classify a polygon according to the number of its sides or angles.
4. Classify a convex polygon according to the measure of its angles.
5. Apply the fact that the sum of the measures of the angles of a triangle is 180.
6. Find the measures of the exterior angles of a triangle.
7. Find the measures of the interior and exterior angles of a convex polygon.
8. Apply the characteristics of various quadrilaterals.

Goal 8: The learner will identify congruent triangles and complete proofs and exercises related to them.

1. List the corresponding parts of two congruent triangles.
2. Use various postulates and theorems to prove two triangles are congruent and their corresponding parts are congruent.
3. Identify the altitudes and medians of triangles.
4. Apply the theorem about the segment joining the midpoints of two sides of a triangle.
5. Apply the theorem about the intersection of the medians of a triangle.

Goal 9: The learner will demonstrate when two polygons are similar and develop proofs and solve exercises related to them.

1. Identify regular polygons and determine the measures of the angles.
2. Solve a proportion.
3. Use proportions to solve geometric problems.
4. Find the geometric mean of two numbers.
5. Determine whether or not two polygons are similar.
6. Prove two triangles are similar.
7. Apply properties of similar triangles to find corresponding proportional sides.
8. Apply theorems which involve dividing segments proportionally.

Goal 10: The learner will state some of the characteristics of a right triangle and solve exercises related to them.

1. State two relationships that exist in a right triangle.
2. Use the Pythagorean Theorem and its converse to find the lengths of the sides of a right triangle or a quadrilateral.
3. Use the relationships that exist in special right triangles to solve problems.
4. Using a table and/or calculator, apply the definitions of sine, cosine, and tangent to solve right triangles.

Goal 11: The learner will list some characteristics of a circle and develop proofs and solve exercises related to them.

1. Use the definitions of a circle and the lines and segments related to it.
2. Recognize polygons inscribed in or circumscribed about a circle.
3. Apply the properties involving arcs and angles of circles.
4. Apply the theorems about the chords of a circle.
5. Apply the theorems that relate to the tangents, secants, and radii of a circle.

Goal 12: The learner will find the perimeter, area, and volume of geometric figures.

1. Find the perimeter of a geometric figure.
2. Compute the area of a triangle, parallelogram, trapezoid, and rectangle.
3. Find the ratio of both the areas and the perimeters of similar triangles.
4. Compute the apothem, radius, and area of special regular polygons.
5. Compute the circumference and area of a circle.
6. Compute arc lengths and the areas of sectors of a circle.
7. Identify and describe space figures.
8. Compute the lateral area, total area, and volume of a right prism or pyramid.
9. Compute the lateral area, and volume of a right circular cylinder or cone.

Goal 13: The learner will complete a geometric construction and describe the locus of a point or points.

1. Construct a segment congruent to a given segment.
2. Construct an angle congruent to a given angle.
3. Construct the bisector of an angle.
4. Construct a line perpendicular to a line through a point on the line.
5. *Construct a line perpendicular to a line through a point not on the line.
6. Construct the perpendicular bisector of a segment.
7. Construct a line parallel to a line through a given point.
8. Construct the tangents to a circle from a point outside the circle.
9. Circumscribe a circle about a triangle.
10. Inscribe a circle inside a triangle.
11. Divide a segment into a given number of congruent segments.
12. Given three segments, construct a fourth segment such that the lengths of the four segments are proportional.
13. Construct a segment whose length is the geometric mean between the lengths of two given segments.
14. Construct quadrilaterals which meet certain criteria.
15. Construct a circle through three non-collinear points.

---

*These objectives would be included in an enriched course but not in a basic course.

Goal 14: The learner will investigate some of the properties of coordinate geometry.

1. Write the coordinates for a point in the coordinate plane.
2. Write equations for vertical and horizontal lines in the coordinate plane.
3. Use the distance formula to solve problems.
4. Use the midpoint formula to find the coordinates of the midpoint or endpoint of a segment.

*5. Find the slope of the line given two points on the line.

*6. Find the slope and y-intercept of a line.

*7. Write an equation for a line which is parallel or perpendicular to a given line.

*8. Write the equation and draw the graph of a line when given either two points on the line, one point and the slope of the line, or the slope and y-intercept of the line.

*9. Use coordinate geometry to prove some of the properties of polygons.

*10. Write an equation of a circle given its center and radius length.

*11. Find the center and radius length of a circle given an equation.

Appendix B

(Rated Samples of Geometry Proofs
and the Score Scale)

The Score Scale

Annotated examples of the score points are given on the following pages. Note that although proofs may differ in difficulty and complexity, the criteria for each score point should remain the same. Differences in difficulty will then be evident in the proportions of students receiving each score point. In addition, the score scale is not meant to be interval in nature; the difference between a 1 and a 2 will not be the same as the difference between a 2 and a 3, etc. Just as there are many varieties of 'B' students, there will be relatively wide variations in quality within score points and these will occur in some score points more than others.

4 = The response demonstrates a clear understanding of the proof. The proof is complete. All logical steps are given and wording is accurate. All statements are logically sequenced and all reasons are correctly aligned with these statements. Mathematically equivalent variations of the answers given in this guide are given a score of 4. Complete and geometrically correct proofs arrived at by different methods than those presented here are also given a score of 4, as long as the logic is sound. Unconventional wording and abbreviations, minor misspellings, correct, but irrelevant, statements that do not seriously detract from the solution as a whole are allowed as long as the statements and reasons are mathematically correct. Proofs scored a 4 do not contain any incorrect statements or reasons, even if the incorrect information is irrelevant to the proof.

3 = The response exhibits a reasonable command of geometric logic in developing the proof. The proof indicates considerable thought and sound logic in the sequence of statements and reasons, but may be lacking in precise notation, wording of theorems, postulates, etc. The proof is generally coherent and complete overall, with major steps always present, although minor weaknesses are present, i.e. a part of a step such as a reason may be missing or stated incorrectly if the corresponding statement is present and correct or incorrect irrelevant statements may be present.

2 = The response demonstrates a weakness in geometric logic in developing the proof. A proof is attempted but is not complete in logic or sequence of statements and reasons. In some proofs, although the student demonstrates a fair understanding of the problem, he or she has omitted or incorrectly stated a major step(s) (including the given) required of the proof. Statements and reasons following an incorrect step may be logical and geometrically sound but they follow from a false conclusion. In other responses the sequence of logical steps is not maintained to the extent that it detracts from the solution.

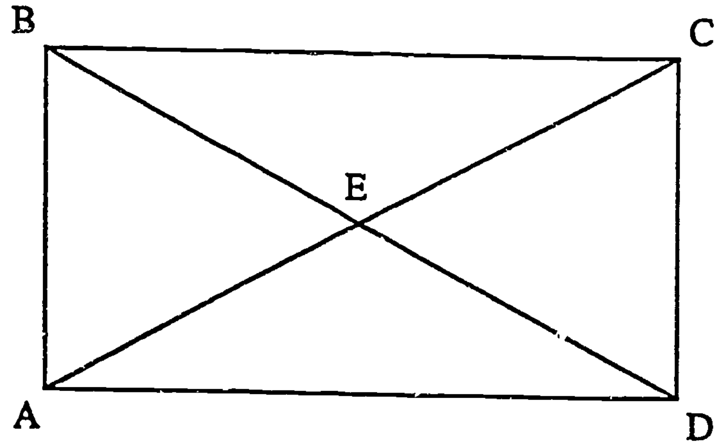
1 = The response exhibits a lack of command of geometry in developing the proof. There is evidence that the student has seen the problem and has attempted the proof, but the proof is off-base. The student demonstrates a vague knowledge of the steps in the proof, but there is very little substance to the proof. The first and last steps may be present, however the majority of the intervening statements and reasons are incorrect or irrelevant. The proof must contain some bit of relevant and correct information other than the given and the prove.

0 = Either the proof is not attempted, the paper is blank, or only the given and/or prove steps are present, or all other steps are totally (statement and reason) incorrect or irrelevant. Nothing is correct except the given and/or prove steps.

Use the figure to prove the following exercise.

1. Given: Figure ABCD is a rectangle with diagonals $\overline{BD}$ and $\overline{AC}$.

   Prove: $\triangle AED \cong \triangle CEB$



Statements	Reasons
1. Figure ABCD is a rectangle with diagonals $\overline{BD}$ and $\overline{AC}$	1. Given
2. $\angle A, \angle B, \angle C, \angle D$ are rt. $\angle$'s	2. If it is a rectangle then it has 4 rt. $\angle$'s
3. $\overline{BD} \perp \overline{AC}$, $\overline{BD} \cong \overline{AC}$	3. Diagonals of a rectangle are $\perp$ and $\cong$
4. $\overline{AE} \cong \overline{ED}$, $\overline{BE} \cong \overline{EC}$	4. CPCTC
5. $\therefore \triangle AED \cong \triangle CEB$	5. SAS

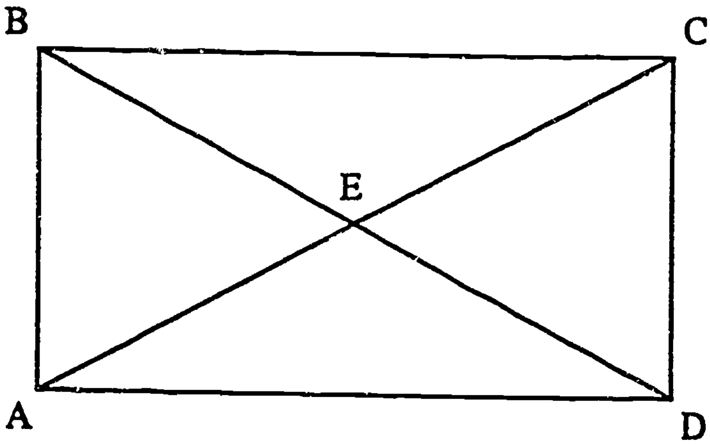
Score Point 0.

All steps other than the given and prove are totally incorrect or irrelevant.

Use the figure to prove the following exercise.

1. Given: Figure ABCD is a rectangle with diagonals $\overline{BD}$ and $\overline{AC}$.

Prove: $\triangle AED \cong \triangle CEB$



Statements	Reasons
1 Figure ABCD is a rectangle with diagonals $\overline{BD}$ and $\overline{AC}$	Given
2 $\overline{BC} \parallel \overline{AD}$, $\overline{BA} \parallel CD$	Definetion of prepdralell lines
3. E is the midpoint of $\overline{BD}$ And $\overline{AC}$	Definetion of midpoint
4 ∠ BAD and ∠ DCD are right angles	All right angles are congruent
5 △ AED ≅ △ CEB	SSS
6 $\overline{AE}$ ≅ $\overline{CE}$ and $\overline{BE}$ ≅ $\overline{DE}$	segments are conguent

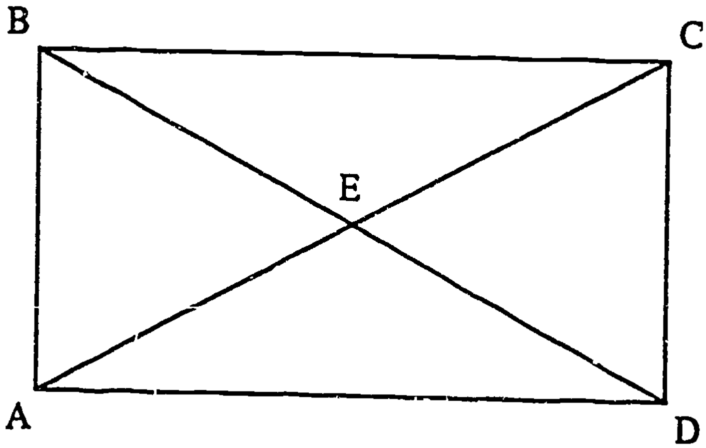
Score Point 1.

The response exhibits a lack of command of geometry in developing the proof. The student demonstrates a vague knowledge of how to prove that two triangles are congruent, but there is very little substance to the proof. The only bits of relevant and correct information in this proof, other than the given and prove, are the first part of Statement 2, Statement 3, and Statement 6 (too late though!).

Use the figure to prove the following exercise.

1. Given: Figure ABCD is a rectangle with diagonals $\overline{BD}$ and $\overline{AC}$.

Prove: $\triangle AED \cong \triangle CEB$



Statements	Reasons
1) Rectangle ABCD with diagonals $\overline{BD}$ and $\overline{AC}$	1) Given
2) $m\angle A = 90°$	2) All angles of a rect. are rt. angles. , Def. rt. angle
3) $\angle A \cong \angle B \cong \angle C \cong \angle D$	3) If one angle is a rt. ∠, then all ∠'s of a rect. are rt. ∠'s.
4) $m\angle BCE = 45°$	4) Diagonals of a rect. bisect ∠'s.
5) $\angle BCE \cong \angle DAE$	5) Transitive
6) $\overline{AC} \cong \overline{BD}$	6) Diagonals of a rect. are ≅.
7) $\overline{AE} \cong \overline{EC}$	7) Diagonals bisect each other.
8) $\angle AED \cong \angle CEB$	8) Vertical ∠'s are ≅.
9) ∴ $\triangle AED \cong \triangle CEB$	9) ASA

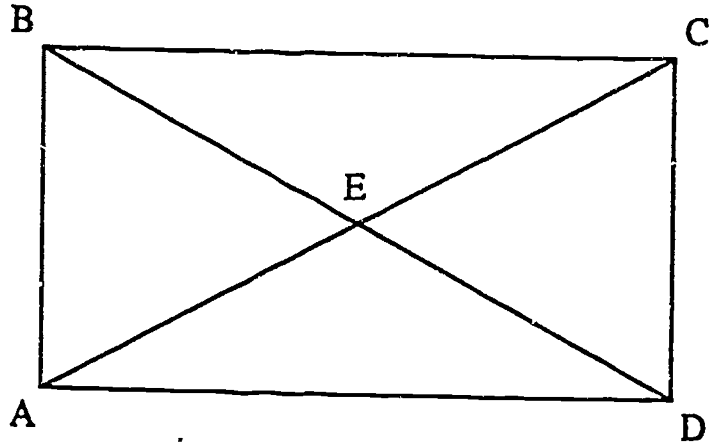
Score Point 2.

The response demonstrates a weakness in geometric logic in developing the proof. Step 4 and Reason 5 are incorrect. Steps 2, 3, and 6 are irrelevant. Also, Statement 3 and Reason 3 do not "agree"; and "$\overline{BC} \parallel \overline{AD}$" must precede Statement 5. The sequence of logical steps is not maintained to the extent that it detracts from the solution.

Use the figure to prove the following exercise.

1. Given: Figure ABCD is a rectangle with diagonals $\overline{BD}$ and $\overline{AC}$.

Prove: $\triangle AED \cong \triangle CEB$



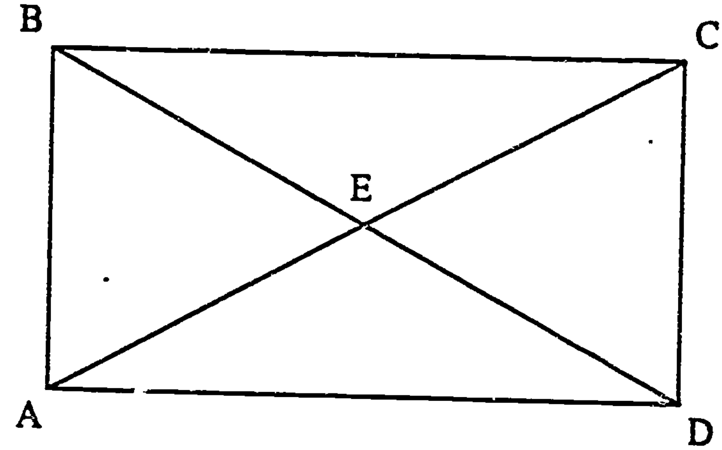
Statements	Reasons
① Figure ABCD is a rectangle with diagonals $\overline{BD}$ and $\overline{AC}$.	① Given
② $\overline{AD} \cong \overline{CB}$	② If a figure is a rectangle, then opposite sides are ≅.
③ $\overline{AE} \cong \overline{CE}$ and $\overline{DE} \cong \overline{BE}$.	③ If diagonals bisect, then the divide segments into 2 ≅ parts.
④ $\triangle AED \cong \triangle CEB$	④ If SSS ≅ exists, then △'s are ≅.

Score Point 3.

The student demonstrates a reasonable understanding of how to do the proof using SSS method, although a minor weakness is present. The student failed to state that the diagonals of a rectangle bisect each other. This is needed prior to Step 3. This omission is a minor weakness since Step 3 is present. Although a minor weakness is present, the proof indicates considerable thought and sound logic in the sequence of statements and reasons.

1. Given: Figure ABCD is a rectangle with diagonals $\overline{BD}$ and $\overline{AC}$.

Prove: $\Delta AED \cong \Delta CEB$



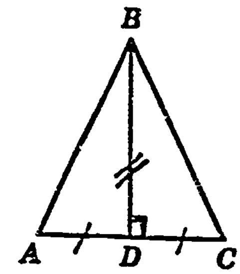
Statements	Reasons
1) Figure ABCD is a rectangle with diagnols $\overline{BD}$ and $\overline{AC}$.	1) Given
2) point E is the midpoint of both $\overline{BD}$ and $\overline{AC}$	2) if it is a rectangle, then the diagnols bisect each other.
3) $\overline{BC} \cong \overline{AD}$	3) if it is a rectangle, then oppisite sides are $\cong$.
4) $\overline{BE} \cong \overline{ED}$ and $\overline{AE} \cong \overline{EC}$	4) if it is the midpoint of a segment, then it will divide into 2 $\cong$ segments.
5) $\Delta AED \cong \Delta CEB$	5) if. SSS, then Δ's are $\cong$.

Score Point 4.

The student demonstrates a clear understanding of how to prove that two triangles are congruent using SSS method by including all steps in the proof. All statements are logically sequenced and all reasons are correctly aligned with these statements. The abbreviations used are acceptable. Although there are minor misspellings in Statement 1, Reason 2, and Reason 3, they do not seriously detract from the solution as a whole. The proof is accurate and complete.

C2. Given: In $\triangle ABC$, $\overline{BD}$ is the perpendicular bisector of $\overline{AC}$.

Prove: $\overline{AB} \cong \overline{BC}$



Proof:

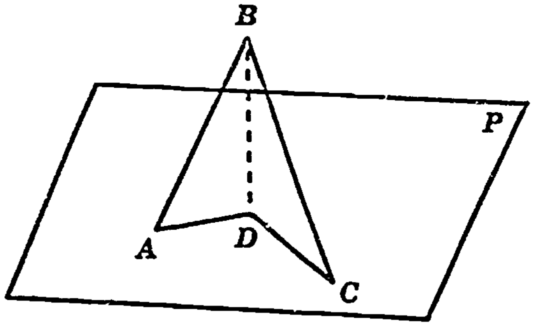
1. In $\triangle ABC$, $\overline{BD}$ is the $\perp$ bisector of $\overline{AC}$.	1. Given
2. $\angle BDA$ and $\angle BDC$ are rt. $\angle$s; $\overline{AD} \cong \overline{DC}$	2. def. of $\perp$ bisecto
3 $\overline{BD} \cong \overline{BD}$	3. reflexive
4. $\angle BDA \cong \angle BDC$	4. all right $\angle$s are $\cong$
5. $\triangle ABD \cong \triangle CBD$	5 SAS
6. $\overline{AB} \cong \overline{BC}$	6 CPCTC

Score Point 4

The student demonstrates a clear understanding of how to prove that two segments are congruent by including all steps in the proof. All statements are logically sequenced, and all reasons are correctly aligned with these statements. The abbreviations used are acceptable. The proof is accurate and complete.

D2. Given: $\overline{BD} \perp$ plane P at D and $\overline{AB} \cong \overline{BC}$

Prove: $\overline{AD} \cong \overline{DC}$



Statements	Reasons
1. $\overline{BD} \perp$ plane P at D and $\overline{AB} \cong \overline{BC}$	1. Given
2. $\overline{BD} \perp \overline{AD}$, $\overline{BD} \perp \overline{DC}$	2. definition of a line perpendicular to a plane
3. $\angle BDA$ & $\angle BDC$ are Rt. $\angle$'s	3. definition of perp. lines
4. $\triangle BDA$ & $\triangle BDC$ are Rt. $\triangle$'s	4. definition of Rt. $\triangle$
5. $\overline{BD} \cong \overline{BD}$	5. Reflexive prop. of $\cong$
6. $\triangle BDA \cong \triangle BDC$	6. Hypotenuse Leg
7. $\overline{AD} \cong \overline{DC}$	7. CPCTC

Score Point 4

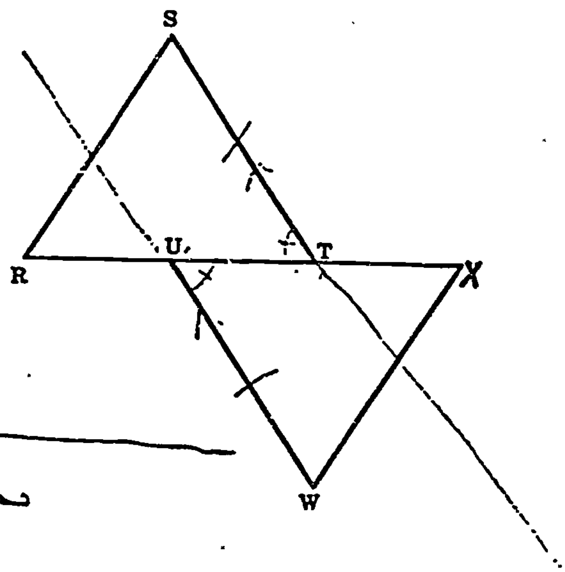
The student demonstrates a clear understanding of how to prove that two segments are congruent by including all steps in the proof. All statements are logically sequenced, and all reasons are correctly aligned with these statements. The abbreviations used are acceptable. The proof is accurate and complete.

G2. Given: $\overline{ST} \cong \overline{WU}$

$\overline{RT} \cong \overline{XU}$

$\overline{ST} \parallel \overline{WU}$

Prove: $\overline{RS} \parallel \overline{XW}$



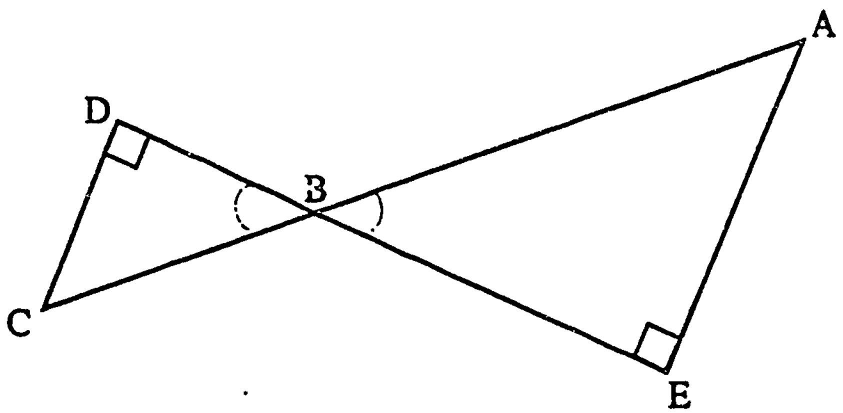
Statements	Reasons
1. $\overline{ST} \cong \overline{WU}$; $\overline{RT} \cong \overline{XU}$; $\overline{ST} \parallel \overline{WU}$	1. given
2. $\angle STR \cong \angle XUW$	2. If 2 lines are ∥ & cut by a transversal, then alternate interior ∠s are ≅.
3. $\triangle RTS \cong \triangle XUW$	3. SAS
4. $\angle R \cong \angle X$	4. CPCTC
5. $\overline{RS} \parallel \overline{XW}$	5. If alternate interior ∠s are ≅, then the 2 lines cut by the transversal are ∥. QED

Score Point 4

The student demonstrates a clear understanding of how to develop the proof by including all steps. All statements are correct and logically sequenced, and all reasons justify these statements. The abbreviations used are acceptable. The proof is accurate and complete.

2. Given: $\overline{CD} \perp \overline{DE}$
$\overline{AE} \perp \overline{DE}$

Prove: $\frac{BC}{BA} = \frac{DC}{EA}$



Statements	Reasons
① $\overline{CD} \perp \overline{DE}$; $AE \perp DE$	① Given
② ∠D and ∠E are right angles	② Definition of ⊥ lines
③ ∠D ≅ ∠E	③ Right angles are congruent
④ ∠DBC ≅ ∠EBA	④ Vertical angles are congruent
⑤ △DBC ~ △EBA	⑤ AA ~ Postulate
⑥ $\frac{BC}{BA} = \frac{DC}{EA}$	⑥ Corresponding sides of ~ triangles are in proportion.

Score Point 4

The student demonstrates a clear understanding of how to prove that distances are proportional by including all steps in the proof. All statements are logically sequenced, and all reasons are correctly aligned with these statements. The abbreviations used are acceptable. The proof is accurate and complete.

Appendix C

(Percentage of Students Receiving Each Score on the 1988-89 Geometry Proof Field-test)

1988-89 Geometry Proof Field Test

Quadralateral (Common Proof)

Percentage of Students Receiving Each Score

Area	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	N
State	8.4	5.5	19.2	9.4	11.1	8.6	12.9	7.2	17.6	43,926
Region 1	6.0	5.3	18.0	10.1	12.9	9.6	13.4	6.6	18.1	2280
Region 2	8.1	4.6	20.4	10.7	12.9	8.6	11.1	6.7	17.0	5090
Region 3	7.1	4.6	17.8	9.0	10.9	8.4	13.7	7.6	20.7	7286
Region 4	11.3	8.4	22.6	10.6	10.8	8.1	10.1	5.8	12.3	5147
Region 5	8.1	5.0	18.1	8.6	11.3	10.1	13.8	8.0	17.1	8256
Region 6	10.5	6.0	21.7	9.3	9.8	6.7	13.1	6.7	16.2	7942
Region 7	6.2	4.7	16.9	9.2	10.6	9.4	14.1	8.3	20.6	4248
Region 8	7.3	5.2	16.0	8.7	11.8	8.7	14.0	7.6	20.6	3676

Perpendicular Bisector (A)

State	6.7	25.7	25.5	20.3	21.7	11,177
Region 1	2.9	24.1	30.0	15.5	27.7	582
Region 2	5.1	26.7	28.[illegible]	20.4	19.3	1294
Region 3	6.8	20.4	25.7	19.6	27.5	1846
Region 4	8.1	31.5	25.0	14.8	20.5	1310
Region 5	5.8	24.2	24.2	22.3	23.5	2098
Region 6	9.0	29.0	25.7	20.6	15.7	2027
Region 7	4.2	22.0	25.3	26.2	22.3	1083
Region 8	9.2	28.6	21.8	20.5	19.9	936

Three Dimensional (B)

Percentage of Students Receiving Each Score

Area	0.0	1.0	2.0	3.0	4.0	N
State	**15.5**	**43.4**	**32.4**	**7.7**	**1.0**	**11017**
Region 1	12.6	48.0	30.4	8.2	0.9	573
Region 2	19.2	51.4	23.0	5.8	0.6	1279
Region 3	13.8	33.8	41.6	9.6	1.1	1827
Region 4	21.4	53.8	21.5	3.3	0.1	1287
Region 5	13.3	42.8	32.5	9.8	1.5	2066
Region 6	15.9	43.4	32.6	6.9	1.2	1999
Region 7	13.2	40.7	36.3	9.2	0.7	1057
Region 8	14.3	38.2	38.8	7.6	1.1	929

Parallel Lines (C)

Area	0.0	1.0	2.0	3.0	4.0	N
State	**28.6**	**29.7**	**12.7**	**9.4**	**19.7**	**10925**
Region 1	29.1	29.5	14.3	9.7	17.5	567
Region 2	28.3	32.7	11.4	9.5	18.1	1263
Region 3	29.1	27.1	11.7	9.9	22.2	1814
Region 4	35.3	35.5	10.4	5.9	12.9	1284
Region 5	26.2	28.4	15.2	10.4	19.7	2061
Region 6	31.9	28.4	11.0	8.5	20.3	1970
Region 7	19.0	29.8	14.6	12.2	24.4	1059
Region 8	28.2	27.8	14.1	9.3	20.6	907

Similar Triangles (D)

Percentage of Students Receiving Each Score

Area	0.0	1.0	2.0	3.0	4.0	N
State	17.3	37.5	21.6	15.6	8.1	10,807
Region 1	13.6	36.4	23.8	16.7	9.5	558
Region 2	14.8	43.1	20.0	16.3	5.9	1254
Region 3	14.9	36.1	19.7	19.4	10.0	1799
Region 4	22.1	45.3	19.0	9.2	4.4	1266
Region 5	17.2	31.9	24.3	16.7	9.9	2031
Region 6	20.2	39.1	21.0	12.3	7.5	1946
Region 7	17.5	31.2	25.5	16.9	9.0	1049
Region 8	14.7	38.7	20.4	18.1	8.1	904

Appendix D

(Descriptive Statistics for Proof Scores and Grades and Geometry Proof Focused-Holistic Score Scale Distribution)

Descriptive Statistics for Proof Scores and Grades

Variable	Number of Students	Mean	Standard Deviation	Minimum	Maximum
Common Proof Rating 1	43,926	2.143	1.327	0	4.0
Common Proof Rating 2	43,926	2.152	1.331	0	4.0
Variable Proof Rating 1	43,926	1.705	1.253	0	4.0
Estimated Geometry Grade	43,400	2.058	1.347	0	4.0
Estimated Proof Grade	43,103	1.848	1.300	0	4.0

Geometry Proof Focused-Holistic Score Scale Distribution

	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
Common	8.4%	5.5%	19.2%	9.4%	11.1%	8.6%	12.9%	7.2%	17.6%
				(N=43,926)					
A	6.7%		25.7%		25.5%		20.3%		21.7%
				(N=11,177)					
B	15.5%		43.4%		32.4%		7.7%		1.0%
				(N=11,017)					
C	28.6%		29.7%		12.7%		9.4%		19.7%
				(N=10,925)					
D	17.3%		37.5%		21.6%		15.6%		8.1%
				(N=10,807)					

Appendix E

(A Summary of Teacher Assigned Proof Grades and Geometry Proof Scores: Common and Variable Proofs)

TEACHER ASSIGNED PROOF GRADES AND GEOMETRY PROOF SCORES: CORE PROOF

Proof Grades

Proof Scores		F %	D %	C %	B %	A %	All %
	0.0	23.5	10.2	4.5	1.4	.5	8.3
	0.5	12.9	8.1	3.6	1.3	.4	5.5
	1.0	33.1	28.6	18.1	7.3	2.7	19.2
	1.5	10.7	13.8	11.1	6.3	2.1	9.5
	2.0	9.0	13.0	14.4	10.5	5.9	11.1
	2.5	4.5	8.8	11.2	10.3	6.7	8.6
	3.0	3.9	9.7	16.1	19.0	16.6	13.0
	3.5	1.0	3.1	6.8	12.6	16.2	7.2
	4.0	1.4	4.7	14.4	31.5	49.0	17.6

chi-square = 17927.79 $p<.001$
$r = .60$

TEACHER ASSIGNED PROOF GRADES AND GEOMETRY PROOF SCORES: CORE PROOF

:Variable Proof A
Perpendicular Bisector

Proof Grades

		F %	D %	C %	B %	A %	All %
Proof Scores	0	20.7	8.0	3.0	.9	.8	6.7
	1	50.0	39.9	22.4	7.6	2.7	25.8
	2	22.3	31.3	31.8	22.4	12.4	25.4
	3	5.3	13.4	23.8	31.0	29.3	20.2
	4	1.7	7.3	19.0	38.2	54.8	21.8

:Variable Proof B
Three-Dimensional

Proof Grades

		F %	D %	C %	B %	A %	All %
Proof Scores	0	39.21	18.9	10.4	4.5	1.7	15.6
	1	50.9	55.7	49.0	32.9	16.8	43.4
	2	9.7	24.4	36.3	48.6	47.2	32.3
	3	.2	1.1	4.0	12.9	29.3	37.7
	4	.0	.0	.3	1.1	5.0	1.0

:Variable Proof C
Parallel Lines

Proof Grades

		F %	D %	C %	B %	A %	All %
Proof Scores	0	56.8	38.5	24.2	9.6	3.1	28.6
	1	36.1	42.3	33.7	18.6	7.2	29.9
	2	5.1	12.2	17.2	18.1	7.9	12.6
	3	1.1	3.9	11.3	16.7	16.9	9.3
	4	.9	3.1	13.6	37.0	64.9	19.7

:Variable Proof D
Similar Triangles

Proof Grades

		F %	D %	C %	B %	A %	All %
Proof Scores	0	42.5	23.5	11.6	4.3	1.0	17.3
	1	47.9	53.4	41.3	24.4	9.2	37.6
	2	7.7	16.8	27.4	31.1	23.2	21.5
	3	1.5	5.6	14.9	26.0	38.4	15.6
	4	.4	.8	4.9	14.2	28.2	8.0

Appendix F

(Sample School System Disaggregate Report)

REGION [redacted]
SYSTEM [redacted]

NORTH CAROLINA END-OF-COURSE TESTING PROGRAM

GEOMETRY PROOFS --- 1989

SYSTEM REPORT

VARIABLE PROOFS

		NUMBER TESTED	PERPENDICULAR BISECTOR					THREE DIMENSIONAL					PARALLEL LINES					SIMILAR TRIANGLES					COMMON PROOF								
SCORE POINTS			0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
ALL STUDENTS TESTED																															
	STATE	43926	7	26	26	20	22	16	43	32	8	1	29	30	13	9	20	17	37	22	16	8	8	5	19	9	11	9	13	7	18
	REGION	7279	7	20	26	20	27	14	34	42	10	1	29	27	[illegible]	10	22	15	36	20	19	10	7	5	18	9	11	8	14	8	21
	SYSTEM	851	6	16	28	29	21	9	30	45	13	1	23	29	[illegible]	16	19	11	44	26	15	5	6	4	17	8	12	8	17	9	20
SEX																															
MALE	STATE	19291	7	24	25	21	22	15	40	34	9	2	26	27	14	11	22	17	36	22	16	9	8	5	17	9	11	9	13	8	19
	REGION	3242	5	20	24	21	29	13	32	43	11	2	24	23	13	13	27	14	34	20	22	9	6	4	16	8	11	9	14	9	23
	SYSTEM	385	3	12	29	32	24	8	29	50	11	2	30	22	9	20	19	12	39	29	17	3	4	4	17	6	14	8	16	9	22
FEMALE	STATE	22799	6	26	26	20	22	14	45	31	8	1	29	32	13	9	18	16	38	22	16	9	8	5	20	10	11	8	13	7	17
	REGION	3761	6	20	28	19	27	13	36	40	10	1	30	29	12	8	20	14	38	19	18	11	7	4	19	10	11	8	14	7	20
	SYSTEM	433	7	19	26	30	18	10	31	42	16	1	18	33	18	13	18	10	48	22	14	7	6	3	18	11	11	9	16	8	18
PARENTAL EDUCATION																															
LESS THAN 8TH	STATE	249	9	25	33	7	25	27	43	28	0	2	28	46	[illegible]	7	13	20	47	10	15	7	12	4	24	13	9	7	13	6	12
	REGION	40	14	29	57	0	0	33	44	22	0	0	40	30	10	0	10	8	67	0	17	8	13	0	23	20	13	5	10	3	15
	SYSTEM	1																0	**	0	0	0	0	0	0	0	0	**	0	0	0
8TH TO 12TH	STATE	2466	8	35	26	17	14	19	50	26	5	0	37	35	11	6	11	23	45	19	9	4	11	8	24	11	12	7	10	5	11
	REGION	386	11	37	28	12	13	21	38	36	6	0	42	35	11	4	8	29	48	15	5	2	13	8	27	13	11	5	9	5	9
	SYSTEM	20	14	29	14	29	14	0	0	0	**	0	57	29	14	0	0	0	80	20	0	0	5	5	25	15	15	5	15	5	10
HIGH SCHOOL	STATE	9953	7	29	28	18	18	17	47	29	6	1	31	34	13	7	14	19	42	21	13	5	10	6	21	11	12	8	12	6	14
	REGION	1514	8	25	30	18	20	15	39	38	8	0	36	32	12	8	12	18	44	17	17	4	9	5	21	11	12	9	13	6	14
	SYSTEM	133	7	17	29	33	14	21	34	38	7	0	38	35	12	8	8	8	58	17	8	8	7	5	19	10	13	8	15	9	14
MORE THAN 12TH	STATE	29188	6	23	25	22	24	14	40	34	10	2	25	28	14	11	23	15	35	23	17	10	7	5	18	9	11	9	14	8	20
	REGION	5013	4	17	25	22	32	11	32	43	12	2	23	24	13	12	28	12	32	21	22	13	5	4	16	8	11	9	15	8	25
	SYSTEM	665	4	15	28	31	23	7	30	47	14	2	19	27	15	18	21	12	39	27	18	4	5	3	17	6	12	8	17	9	22

NOTE: FOUR FORMS OF THE GEOMETRY PROOFS TEST WERE ADMINISTERED IN EACH CLASSROOM. EACH STUDENT TOOK ONE COMMON PROOF AND ONE OF FOUR VARIABLE PROOFS. THE NUMBERS IN THE TABLE REPRESENT THE PERCENTAGES OF STUDENTS ATTAINING EACH SCORE POINT. 100% IS REPRESENTED BY '**'. PERCENTAGES FOR ALL STUDENTS TESTED WERE OBTAINED DIRECTLY FROM THE SCORE DATA. PERCENTAGES BY SUBGROUP WERE OBTAINED FROM DATA CODED ON THE MULTIPLE-CHOICE ANSWER SHEET.

REGION [redacted]
SYSTEM [redacted]

GEOMETRY PROOFS --- 1989

SYSTEM REPORT

VARIABLE PROOFS

		NUMBER TESTED	PERPENDICULAR BISECTOR					THREE DIMENSIONAL					PARALLEL LINES					SIMILAR TRIANGLES					COMMON PROOF								
SCORE POINTS			0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
GRADE IN SCHOOL																															
NINE	STATE	7820	2	8	18	29	43	4	27	45	21	4	10	15	13	17	45	4	18	26	32	21	2	1	7	5	9	8	16	13	38
	REGION	1532	2	6	15	28	49	6	17	49	26	2	10	13	11	18	48	5	19	22	34	20	2	1	7	5	8	7	16	13	41
	SYSTEM	157	0	3	23	38	38	3	11	47	37	3	2	7	18	34	39	0	17	34	31	17	1	1	3	5	8	4	20	14	45
TEN	STATE	19998	6	23	27	21	22	14	45	33	7	1	25	31	15	10	19	14	38	25	15	8	6	5	18	10	12	9	14	8	18
	REGION	3186	4	19	27	21	28	11	35	44	8	2	24	29	15	10	22	12	35	23	19	11	5	4	16	9	12	9	16	8	21
	SYSTEM	400	6	14	24	35	21	8	32	40	11	0	18	33	17	14	19	8	39	33	16	4	4	4	14	8	13	11	18	9	19
ELEVEN	STATE	11103	10	37	28	14	11	22	50	24	3	0	40	36	11	6	7	25	48	16	8	2	13	8	26	11	12	8	10	4	8
	REGION	1757	11	30	32	13	14	17	47	32	4	0	44	33	11	6	6	23	49	15	11	2	11	7	27	12	12	8	11	4	8
	SYSTEM	216	7	27	39	16	11	7	42	44	7	0	39	32	12	10	7	21	60	10	9	0	10	3	31	11	13	7	13	5	7
TWELVE	STATE	3162	12	37	29	14	7	24	48	26	3	0	41	38	11	4	6	33	44	13	8	2	16	9	26	11	11	7	9	4	7
	REGION	510	7	34	39	11	9	23	35	40	2	0	44	34	10	5	7	23	50	10	13	3	14	7	26	10	13	7	10	4	8
	SYSTEM	45	0	36	45	18	0	38	38	25	0	0	50	50	0	0	0	0	83	17	0	0	13	13	31	9	18	7	7	2	0
OTHER	STATE	109	6	16	13	19	45	0	15	37	26	22	21	5	5	11	58	13	27	13	17	30	10	7	6	2	6	7	6	15	40
	REGION	47	9	0	0	18	73	0	7	29	43	21	11	0	11	11	67	9	9	9	18	55	9	2	0	0	2	9	4	15	60
	SYSTEM	7	0	0	0	**	0	0	0	50	0	50	0	0	0	**	0	0	0	**	0	0	14	0	0	0	0	29	14	14	29
ETHNIC GROUP																															
AMER. INDIAN	STATE	436	11	31	37	14	6	16	67	16	1	0	36	39	9	8	8	24	51	17	6	2	11	9	24	12	14	9	8	5	7
	REGION	25	40	0	40	20	0	9	64	27	0	0	40	20	20	0	20	50	0	50	0	0	20	16	16	8	12	8	12	4	4
	SYSTEM	0																													
BLACK	STATE	10089	10	36	27	13	13	23	49	23	4	0	42	37	9	5	7	27	44	17	8	4	14	8	27	11	11	8	9	4	8
	REGION	2134	11	31	30	14	14	23	41	30	5	0	44	33	10	5	8	26	47	14	10	4	12	7	26	12	11	9	10	4	8
	SYSTEM	206	9	26	29	19	17	16	40	32	12	0	37	35	17	8	4	17	50	24	7	2	9	8	24	13	14	8	10	5	10
WHITE	STATE	30681	5	21	25	23	25	12	40	36	10	2	23	28	15	11	23	13	35	24	18	10	6	4	16	9	11	9	14	8	21
	REGION	4669	3	15	25	23	34	8	30	47	13	2	20	24	14	13	29	9	32	22	24	13	4	3	14	7	11	9	16	9	27
	SYSTEM	582	4	11	28	36	22	7	28	49	14	2	19	25	14	18	23	9	42	25	19	5	5	2	15	7	12	9	19	10	23
OTHER	STATE	854	6	17	24	20	34	10	38	32	17	4	14	25	11	15	34	15	27	24	24	10	7	4	13	7	9	8	15	11	25
	REGION	164	3	8	22	11	57	0	37	37	21	5	3	14	11	19	53	13	21	21	36	10	4	4	7	7	10	5	18	12	33
	SYSTEM	29	0	20	20	40	20	0	0	75	25	0	0	27	9	27	36	13	25	38	13	13	3	0	17	7	10	3	7	7	34

NOTE: FOUR FORMS OF THE GEOMETRY PROOFS TEST WERE ADMINISTERED IN EACH CLASSROOM. EACH STUDENT TOOK ONE COMMON PROOF AND ONE OF FOUR VARIABLE PROOFS. THE NUMBERS IN THE TABLE REPRESENT THE PERCENTAGES OF STUDENTS ATTAINING EACH SCORE POINT. 100% IS REPRESENTED BY '***'. PERCENTAGES FOR ALL STUDENTS TESTED WERE OBTAINED DIRECTLY FROM THE SCORE DATA. PERCENTAGES BY SUBGROUP WERE OBTAINED FROM DATA CODED ON THE MULTIPLE-CHOICE ANSWER SHEET.

N C END-OF-COURSE TESTING PROGRAM: 1988-89
CLASS ROSTER FOR GEOMETRY PROOFS

REGION 6
SCHOOL ANSON JUNIOR H.S.
CODE 40304
SYSTEM ANSON COUNTY SCHOOLS

TEACHER RANDALL P CLASS PERIOD 1

NOTE: CODE THE PROOF SCORES ON THE APPROPRIATE STUDENT ANSWER SHEETS ACCORDING TO THE DIRECTIONS ON PAGES 13 & 14 OF THE TEST ADMINISTRATOR'S MANUAL FOR THE GEOMETRY TEST. CODE THE COMMON PROOF SCORE IN COLUMNS K AND L, THE VARIABLE SCORE IN COLUMNS M AND N, AND THE FORM IN COLUMN O. ONLY THE COMMON PROOF SCORE SHOULD BE USED IN DETERMINING STUDENT GRADES. THE VARIABLE PROOFS VARIED IN DIFFICULTY AND WILL BE USED FOR SCHOOL AND SCHOOL SYSTEM REPORTING. FOR YOUR INFORMATION, THE PROOFS FOR EACH FORM WERE AS FOLLOWS: A=PERPENDICULAR BISECTOR, B=THREE DIMENSIONAL, C=PARALLEL LINES, D=SIMILAR TRIANGLES. THE STATEWIDE SCORE DISTRIBUTIONS FOR ALL PROOFS ARE GIVEN BELOW. COMMON PROOF SCORES ARE BASED ON TWO INDEPENDENT READINGS WHICH PRODUCE SOME MID-POINT SCORES. VARIABLE PROOF SCORES ARE BASED ON ONE READING.

STATEWIDE DISTRIBUTION OF SCORES

PROOF	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	TOTAL
COMMON	8.4%	5.5%	19.2%	9.4%	11.1%	8.6%	12.9%	7.2%	17.6%	43926
A	6.7%		25.7%		25.5%		20.3%		21.7%	11177
B	15.5%		43.4%		32.4%		7.7%		1.0%	11017
C	28.6%		29.7%		12.7%		9.4%		19.7%	10925
D	17.3%		37.5%		21.6%		15.6%		8.1%	10807

STUDENT	COMMON PROOF	FORM	VARIABLE PROOF
	2.5	A	3.0
	1.0	A	3.0
	4.0	D	2.0
	3.5	B	3.0
	1.5	A	3.0
	1.0	B	2.0
	1.0	C	4.0
	3.5	C	4.0
	4.0	D	3.0
	2.0	A	2.0
	3.0	D	3.0
	1.5	C	2.0
	0.0	B	2.0
	2.0	D	4.0
	2.0	B	1.0

Appendix G

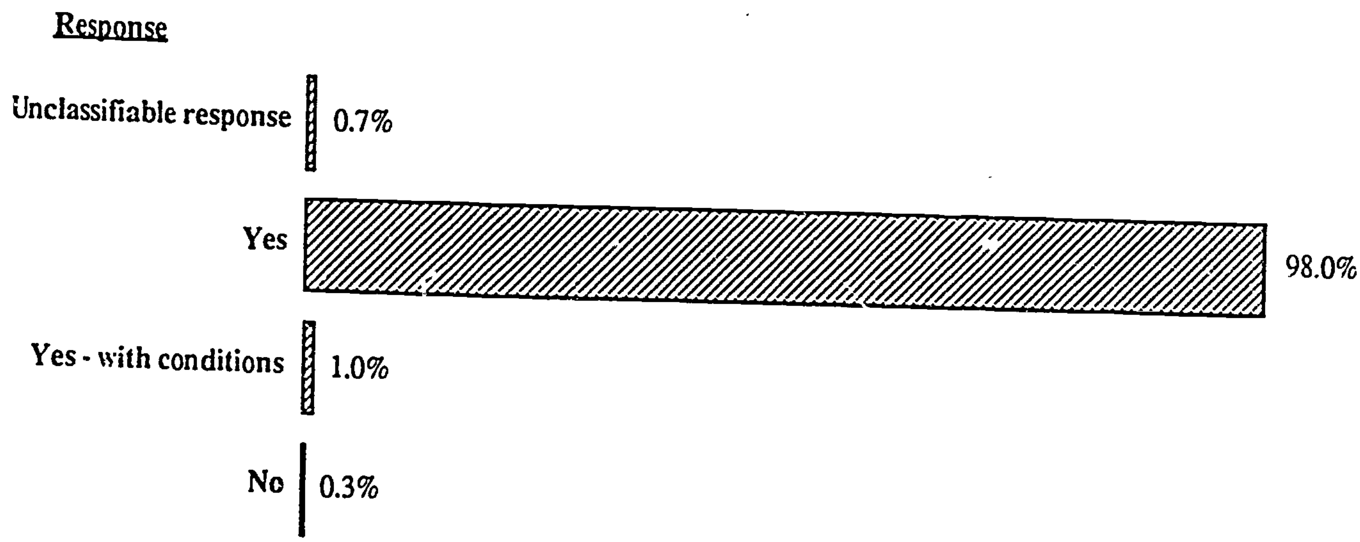
(Teacher Survey Data)

Responses to Evaluation of Geometry Field Test

Question 1: Test Administration

Question 1a: Was there sufficient time to do two proofs in one period?

Responses to 1a:



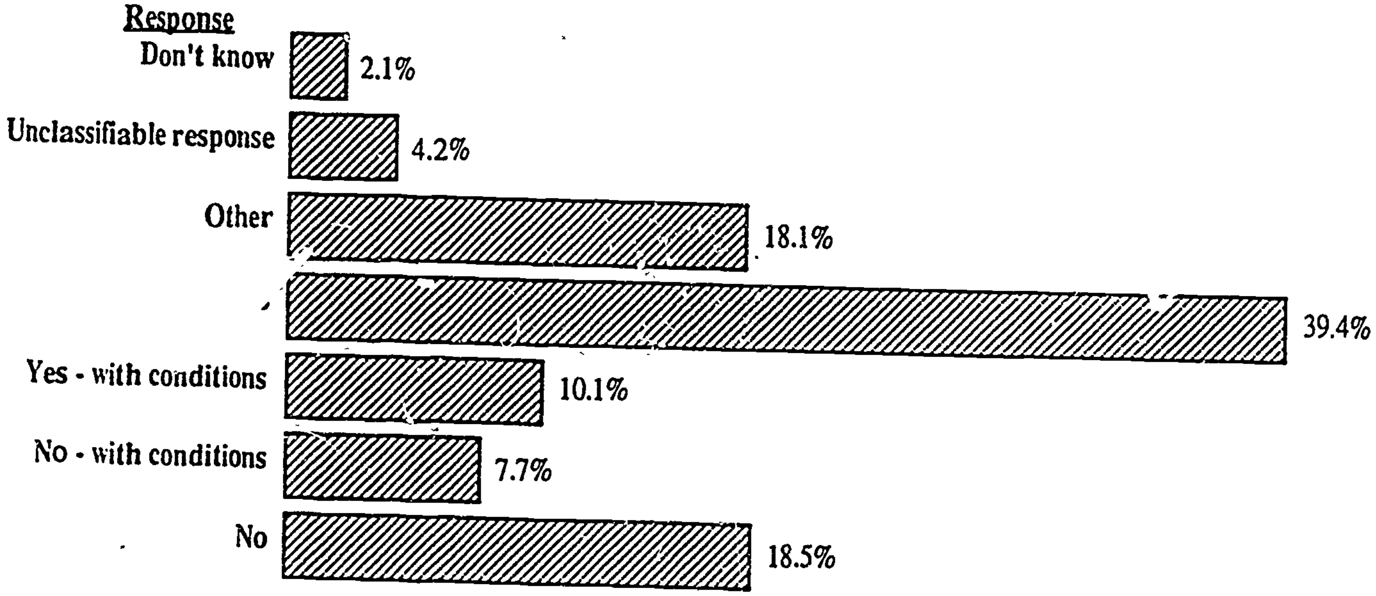
Response	Count
No	1
Yes-- with conditions	3
Yes	297
Unclassifiable response	2
Missing	12

Summary of 1a:

This question had by far the most response; furthermore, the response was focused more directly at a single answer--*yes*. The three conditional affirmative responses were concerned with administrative duties. There was only one *no* without explanation.

Question 1b: Should all testing be done first period?

Responses to 1b:



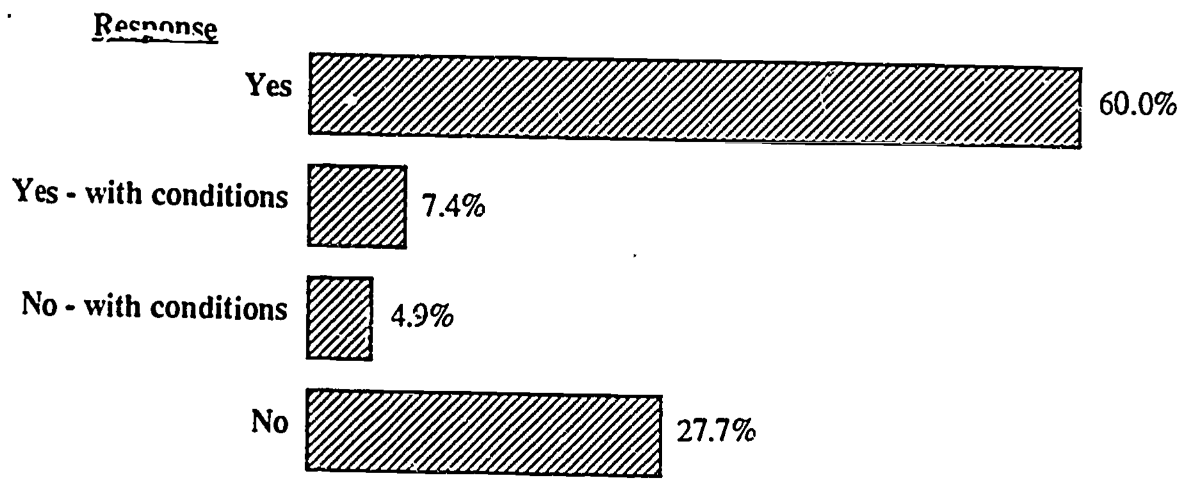
Response	Number
Don't know	6
Unclassifiable response	12
Other	52
Yes	113
Yes - with conditions	29
No - with conditions	22
No	53
Missing	28

Summary of 1b:

Yes was reported more than twice as much as any other response. However, the *other* response showed a different measure; these teachers expressed a concern that all testing should be done during the same period for all students all over the state. Many of the teachers reported that students who were in their classes late in the day already knew which proofs were going to be on the common te[illegible] The *no* and *no--with conditions* that had explanations were generally concerned with late buses or tardy students. But most did think that testing in the morning was a good idea. Some, preferring to take the test in their regular class, explained their *no* responses because of problems with rescheduling other classes.

Question 1c: Is the end of March an appropriate time?

Responses to 1c:



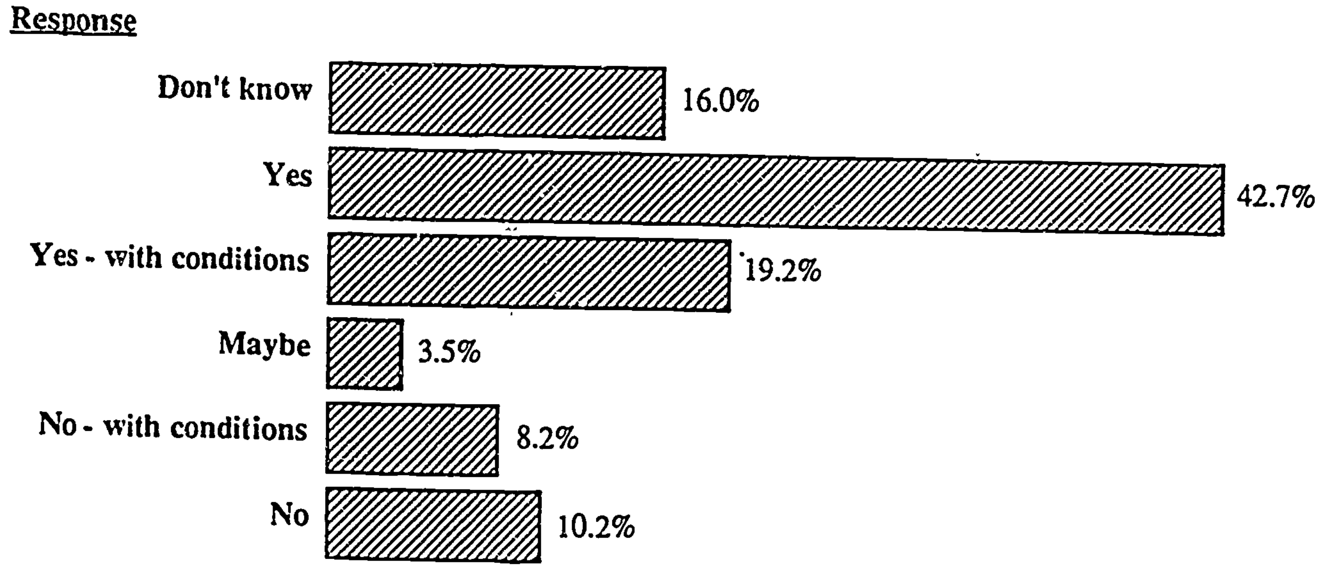
Response	Number
No	79
No--with conditions	14
Yes--with conditions	21
Yes	171
Missing	30

Summary of 1c:

Again, *yes* received twice as many responses as any other answer. In general, the affirmative responses do not have explanations because the question sums up the response. The negative responses, on the other hand, usually are followed by reasons; in this case, the *no* responses were reported because the test was given later than the teachers thought it should have been. Many teachers explained that they taught proofs in the first semester; therefore, testing should be done earlier. Many reported that the end of first semester would be a better time. Much of the conditional response was due to conflicts with spring break and other End of Course testing.

Question 1d: Did students reveal their true abilities?

Responses to 1d:



Response	Number
No	26
No--with conditions	21
Maybe	9
Yes--with conditions	49
Yes	109
Don't Know	41
Missing	60

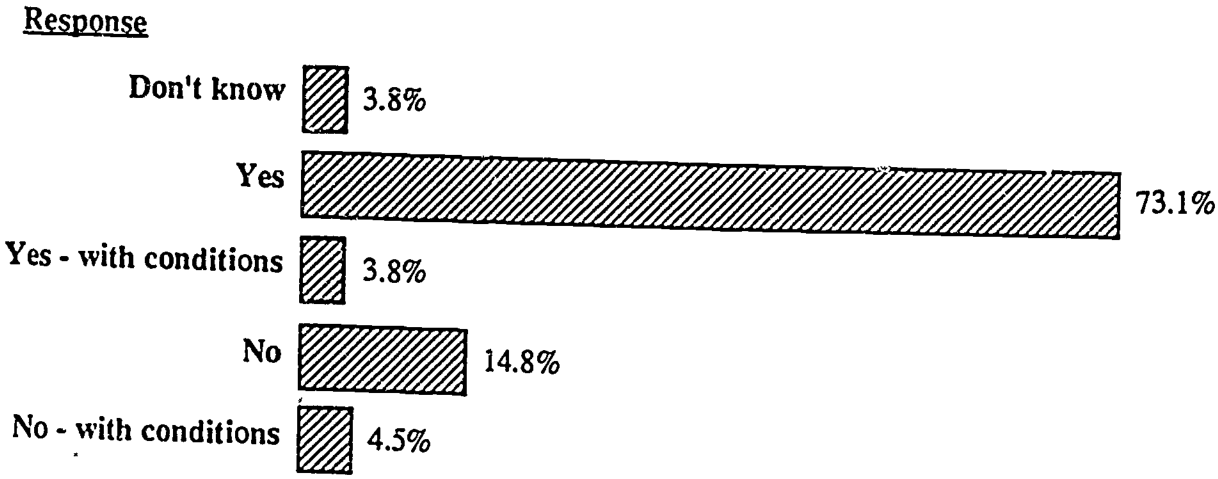
Summary of 1d:

Again, *yes* was twice as popular as any other response. The responses with phrases such as "most did" and "probably did" were categorized as *yes--with conditions*. Also in this category are the responses with "background interference" such as nervousness. Many teachers wrote that they would not know the answer to this question until they know their students' scores. The negative responses are founded upon explanations of students' apathy, nervousness, and, from the response to question 1b, the apparent cheating.

Question 2: Proofing Skills

Question 2a: Is it important to test "proofing" in addition to multiple choice testing?

Responses to 2a:



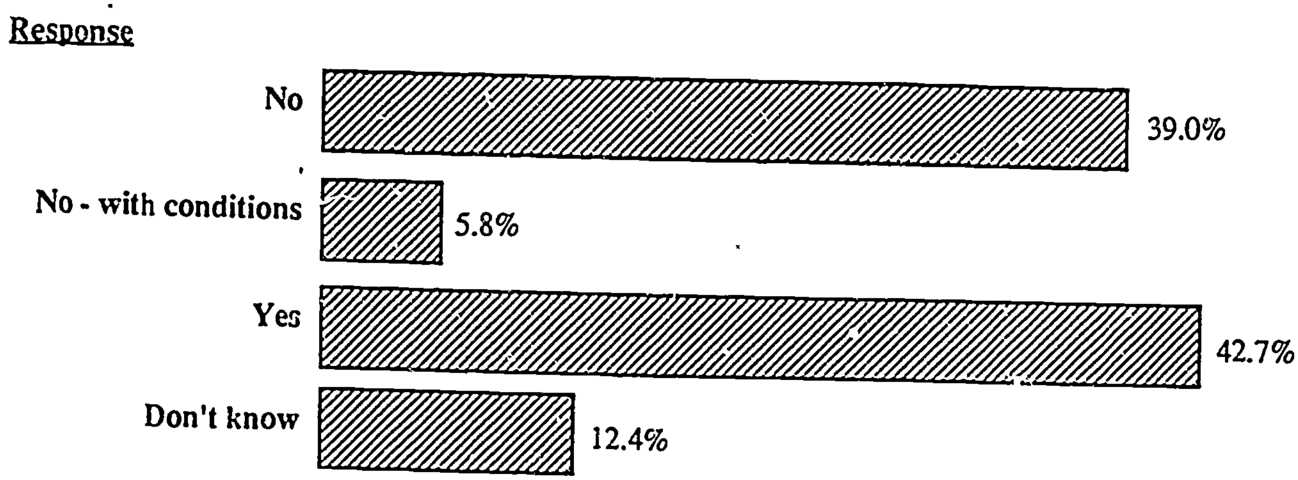
Response	Number
No	39
No--with conditions	12
Yes--with conditions	10
Yes	193
Don't Know	10
Missing	51

Summary of 2a:

By far, the majority of teachers' responses fits into the *yes* category. The *missing* category is comprised largely of responses that did not answer this question, but did answer question 2b with responses such as "it's not worth it." Many teachers suggested a proofing test in the form of a multiple choice test with choices of statements and reasons.

Question 2b: Is it worth the General Assembly providing $3 per student or would sampling be acceptable?

Responses to 2b:



Response	Number
No	94
No--with conditions	14
Yes	103
Don't Know	30
Missing	74

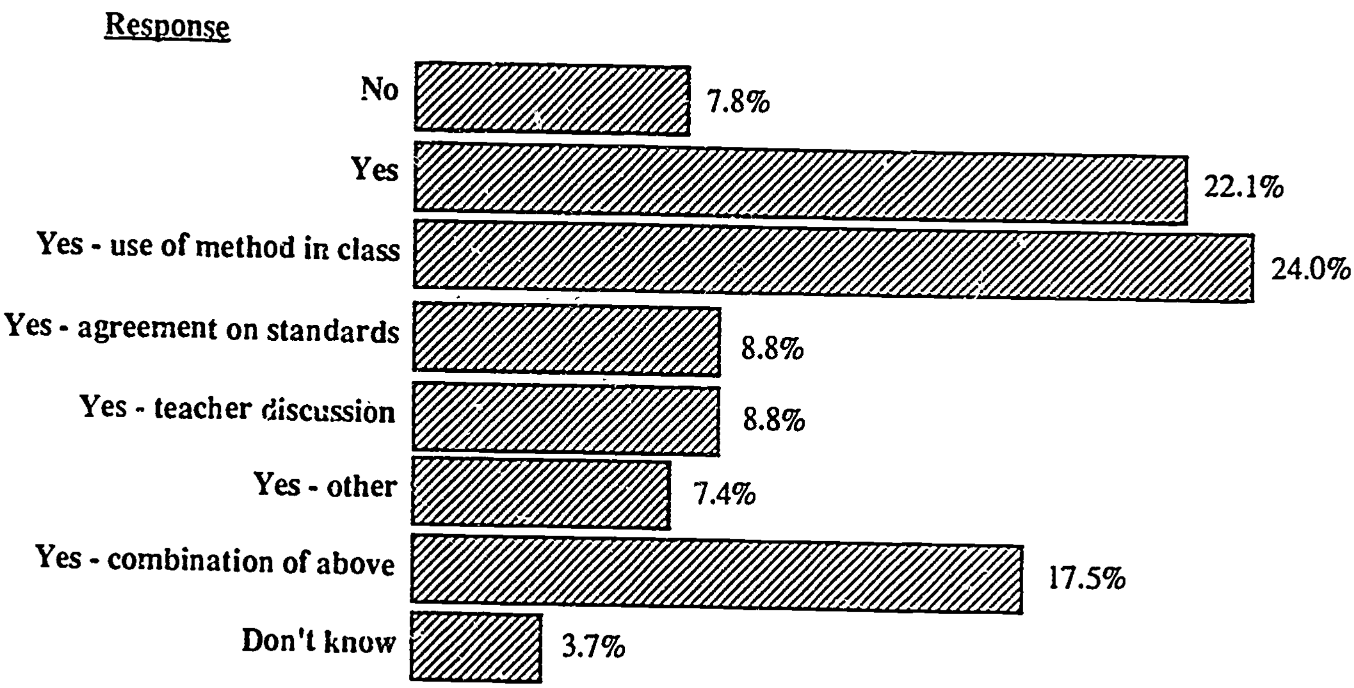
Summary of 2b:

Yes and *no* were reported in almost the same frequencies in response to this question. Many did not know whether or not it was worth the cost. The teachers who responded *yes* to this question thought that it was worth it to test all students, while some of the *no* responses favored sampling.

68

Question 2c: Have there been any benefits beyond scores for students and curriculum information-- e.g., use of scoring method in class, agreement on standards, discussion among geometry teachers?

Response to 2c:



Response	Number
No	17
Yes	48
Yes - use of method in class	52
Yes - agreement on standards	19
Yes - teacher discussion	19
Yes - other	16
Yes - combination of above	38
Don't Know	8
Missing	98

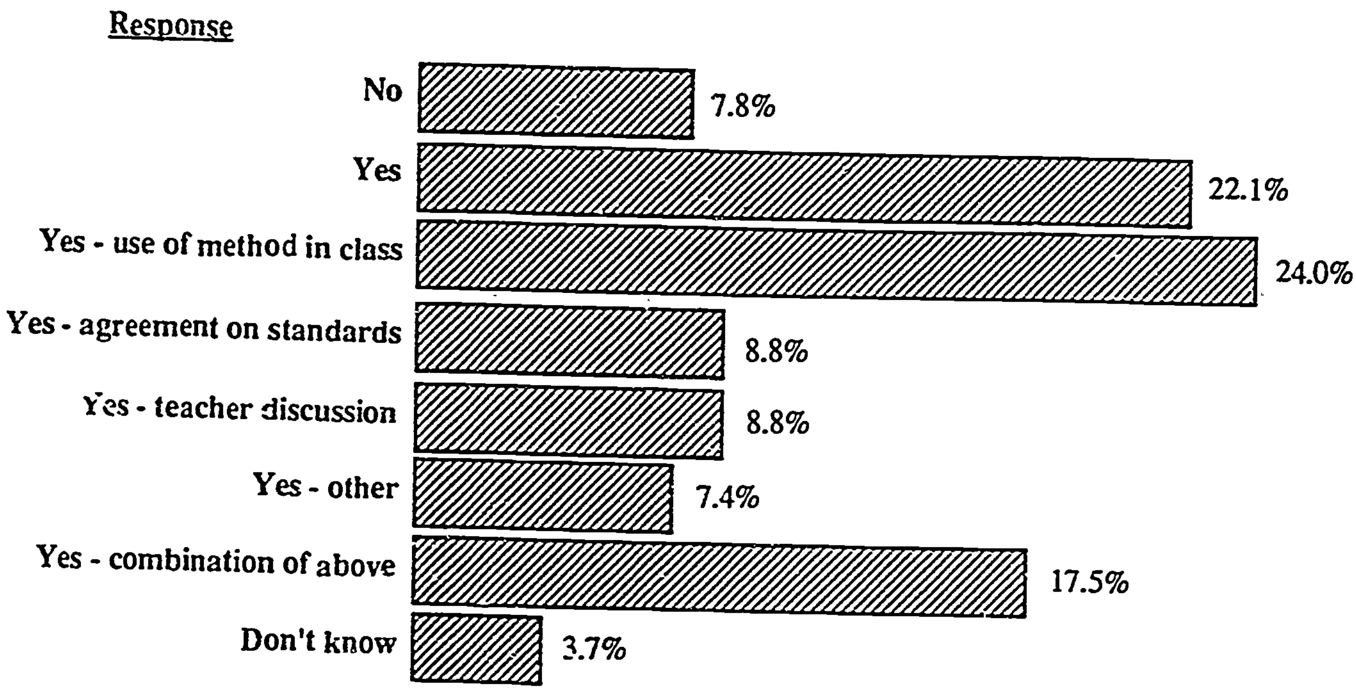
Summary of 2c:

Only 5% said *no*. Of the benefits, *use of the method* was reported the most often. Teachers generally liked the method for scoring proofs. Almost everyone who responded agreed that the method was useful. Another commonly reported benefit was *agreement on standards*; many teachers did not know expectations of their curriculum.

Question 3: Scoring of Proofs

Question 3a: Should lead, practicing teachers continue to be scorers?

Responses to 3a:

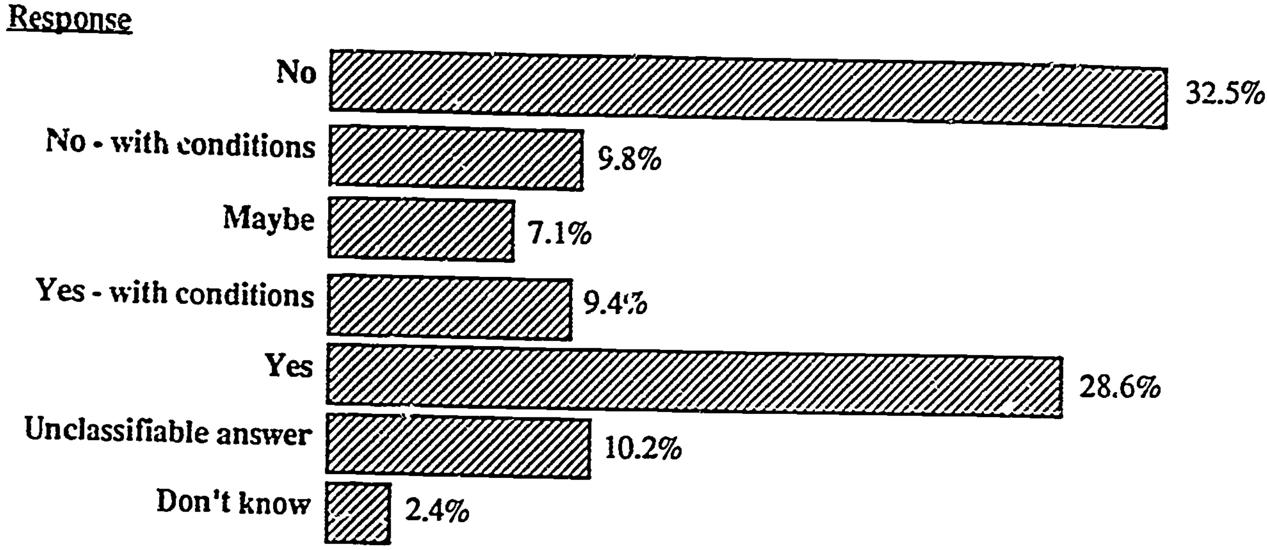


Response	Number
No	14
No--with conditions	4
Yes--with conditions	6
Yes	241
Unclassifiable response	14
Missing	36

Summary of 3a:

Teachers agreed that geometry teachers should continue to be scorers. Some of those teachers in the *no* category suggested that there be a professional group of readers. This question was often answered by the next question which indicated that all geometry teachers should have the experience of grading proofs in this manner. The positive responses were not necessarily directed at the benefit for scoring, but rather at the improvement in teaching skills of individual teachers which results from the experience of scoring.

Question 3b: Should different teachers score each year?



Responses to 3b:

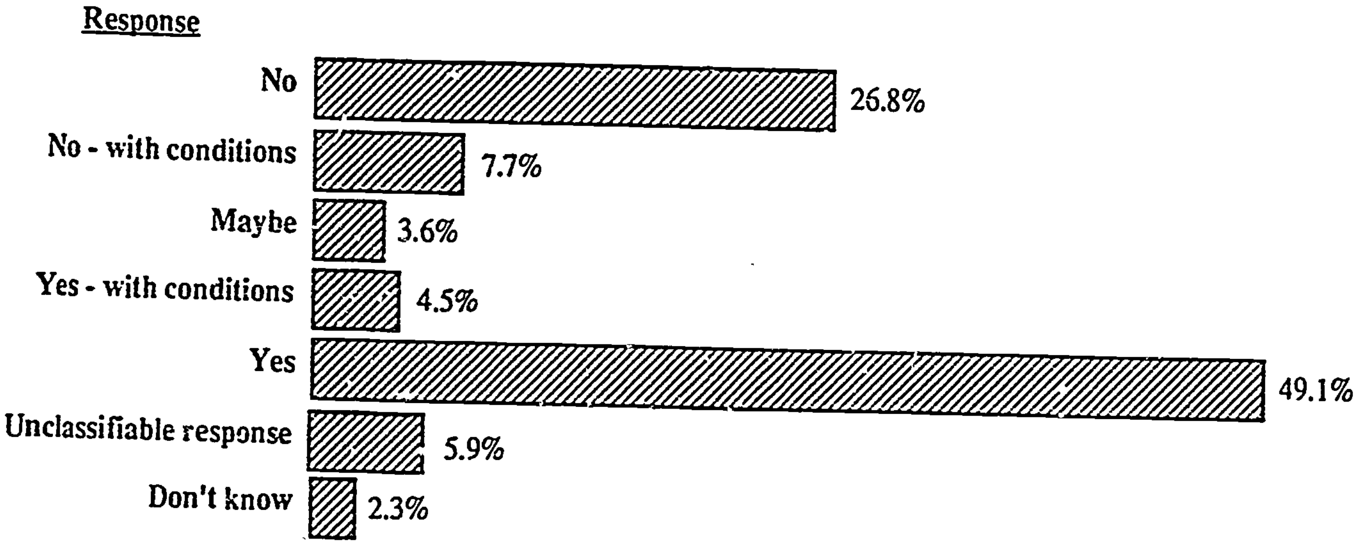
Response	Number
No	83
No--with conditions	25
Maybe	18
Yes--with conditions	24
Yes	73
Unclassifiable Answer	26
Don't Know	6
Missing	60

Summary of 3b:

The negative responses were based on concerns about cost; the positive responses were based on concerns about quality of teaching. The conditional responses and *maybe* category represent a large number of unsure evaluators.

Question 3c: Can scoring move to school systems in the next five years?

Responses to 3c:



Response	Number
No	59
No--with conditions	17
Maybe	8
Yes--with conditions	10
Yes	108
Unclassifiable response	13
Don't Know	5
Missing	95

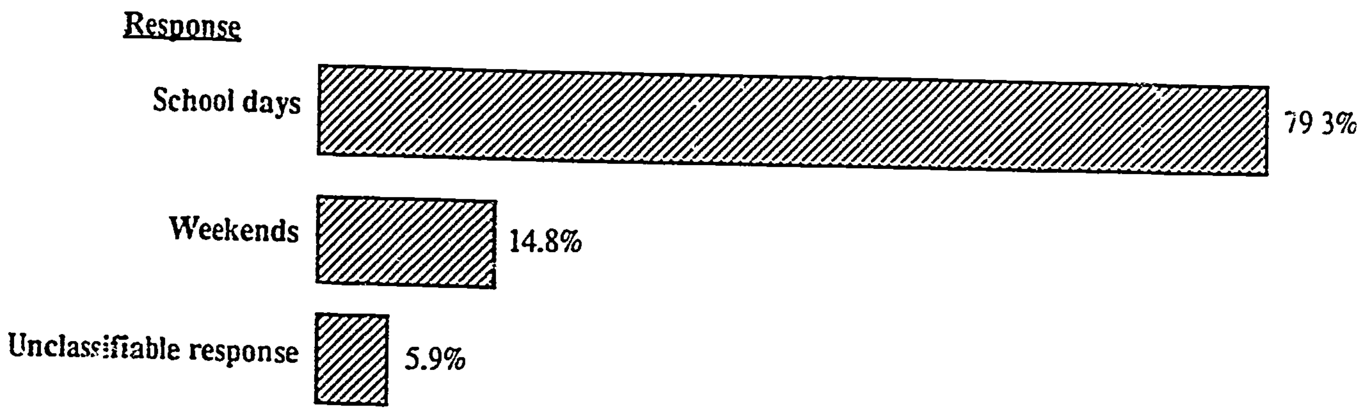
Summary of 3c:

Teachers responded *yes* to this question in general. Those teachers who did respond *no* questioned whether their school system or other school systems would be honest when it came to scoring their own students. Many of the conditional responses showed concern for this matter.

72

Question 3d: Should we use school days or weekends if scoring is done at the regional level?

Responses to 3d:



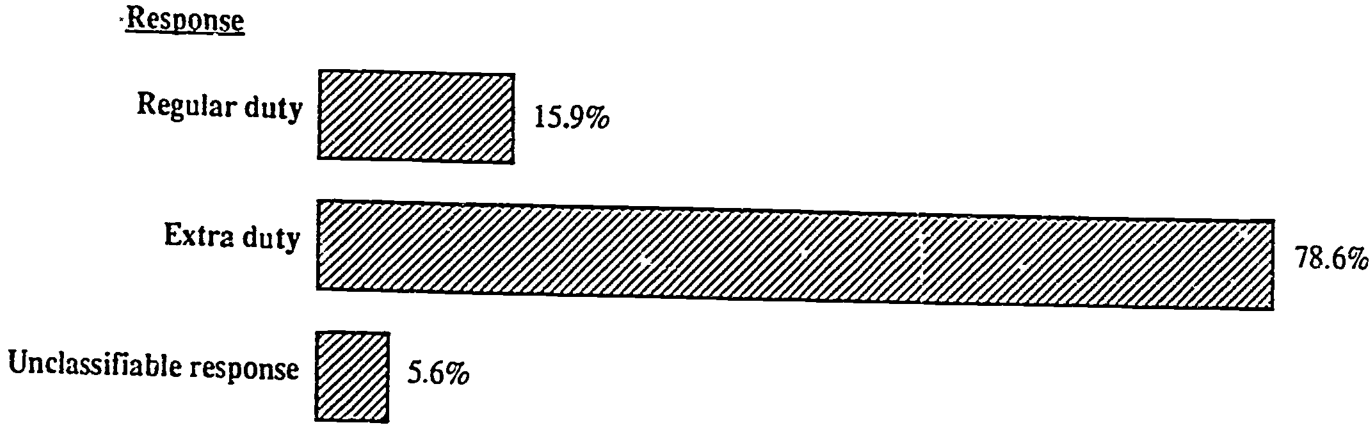
Response	Number
School Days	203
Weekends	38
Unclassifiable response	13
Missing	59

Summary of 3d:

School days was by far the most popular answer to this question. Some were strongly opposed to weekends. Many teachers were opposed to testing taking away from class time; these teachers answered *weekends* or suggested using teacher workdays. The issue of reimbursement for this scoring came up repeatedly in responses to this question; some wanted extra pay for weekends-- some would not score on weekends even for extra pay.

Question 3e: Is this scoring regular duty or extra duty?

Responses to 3e:



Response	Number
Regular duty	40
Extra duty	198
Unclassifiable Response	14
Missing	63

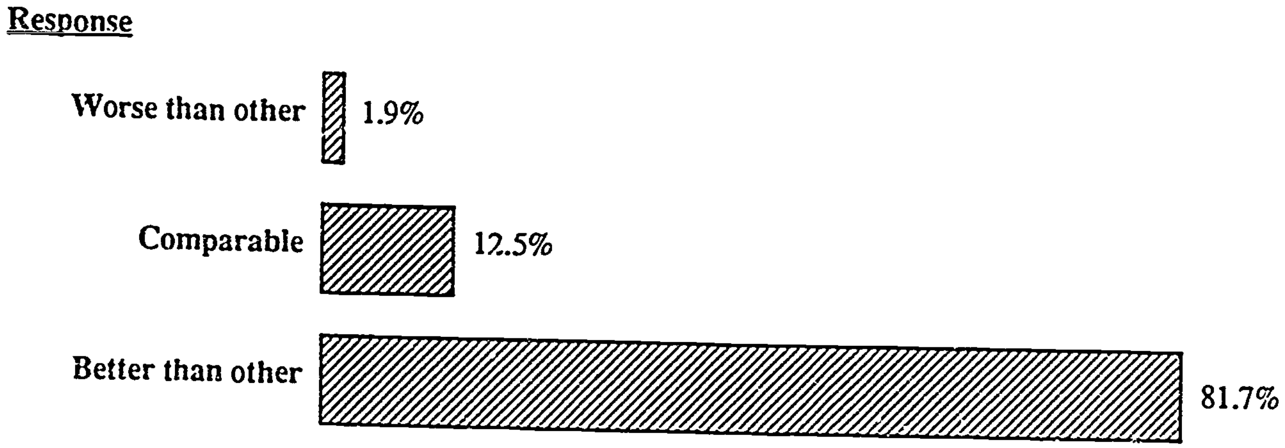
Summary of 3e:

Extra duty was the winner here. Again, many included payment as an issue.

Question 4: Training on Holistic Scoring

Question 4a: Rate the awareness and scoring sessions compared to other staff development.

Responses to 4a:



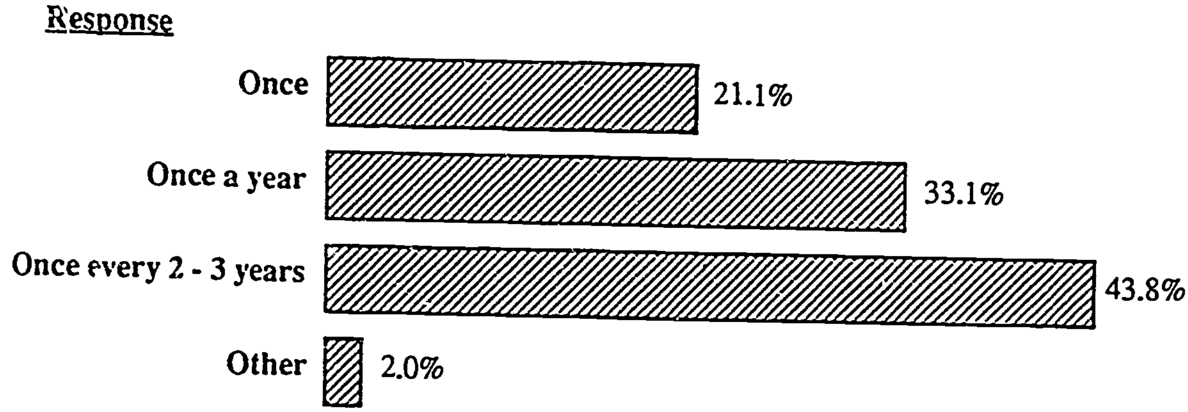
Response	Number
Worse than other	2
Comparable	13
Better than other	85
Missing	211

Summary of 4a:

Most teachers did not respond at all to this question. But those who did respond responded favorably.

Question 4b: How often should a person receive this training-- once, every three years, never?

Response to 4b:



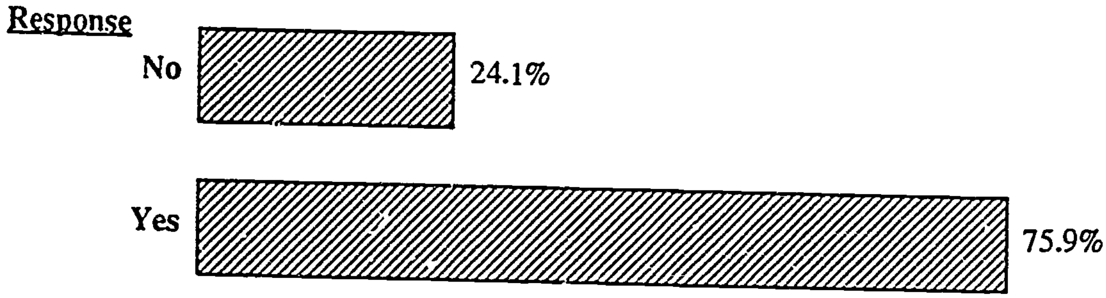
Response	Number
Once	53
Once a year	83
Once every 2-3 years	110
Other	5
Missing	64

Summary of 4b:

Once a year and *once every 2-3 years* were the most prevelant responses. A lot of teachers favored one intense training session with a brief review every testing period. One suggestion as to how this plan could be accomplished was by mailing the proofs with scoring for various sample answers. This question also received more of the "all geometry teachers need it" response.

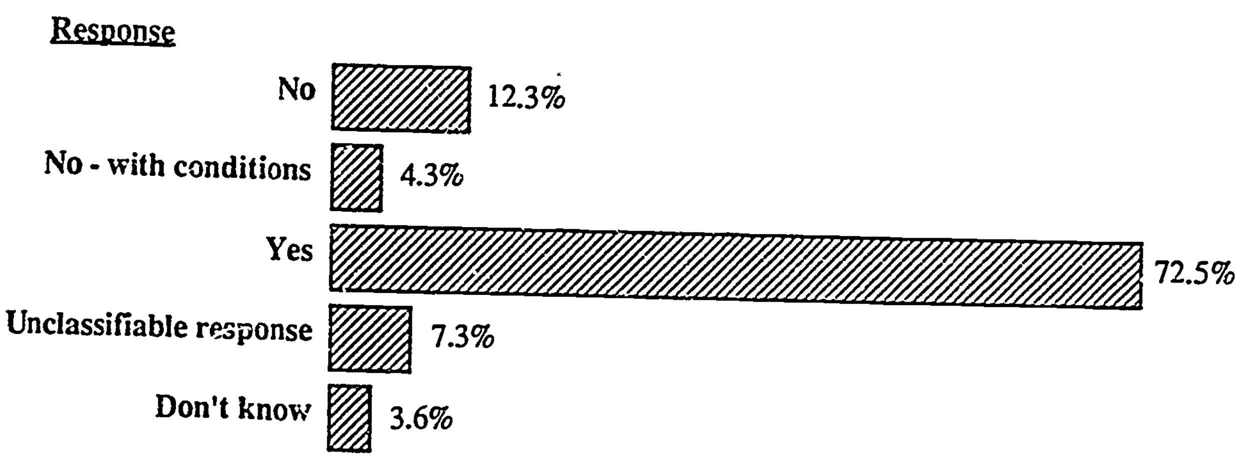
Question 4c: Have you (1)used this technique in class- (2)has it improved instruction?

Response to 4c(1):



Response	Number
No	51
Yes	161
Missing	103

Response to 4c(2):



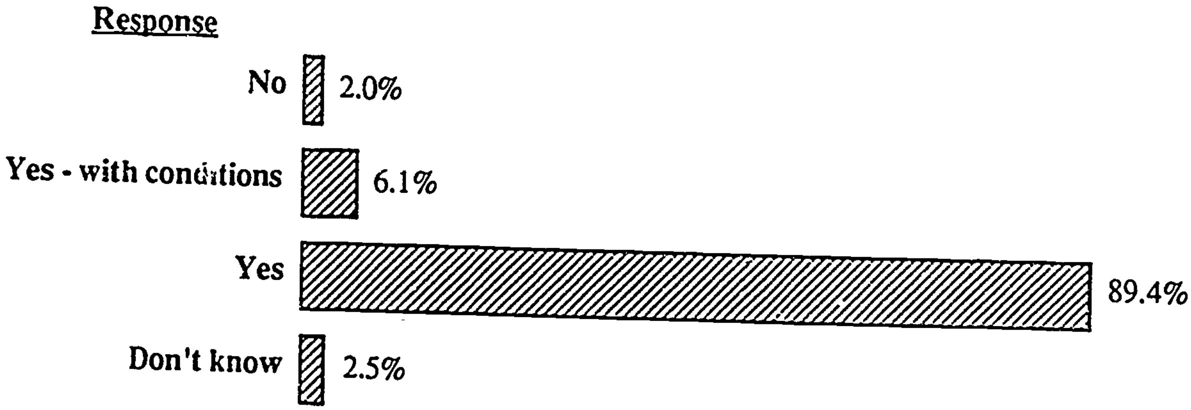
Response	Number
No	17
No--with conditions	6
Yes	100
Unclassifiable response	10
Don't Know	5
Missing	177

Summary of 4c:

Most teachers used the technique to grade proofs in their classes. Of those teachers who responded *no*, many are going to use it next year; in many cases teachers were no longer teaching proofs beacause the test was given late in the year. Those teachers who used the technique reported improved instruction for the most part. They reported improved awareness on the students' part of the method used to grade the proofs. Some said it was too early to tell.Many others reported that it shortened their time to grade the tests.

Question 4d: Would you trust scores from others?

Response to 4d:



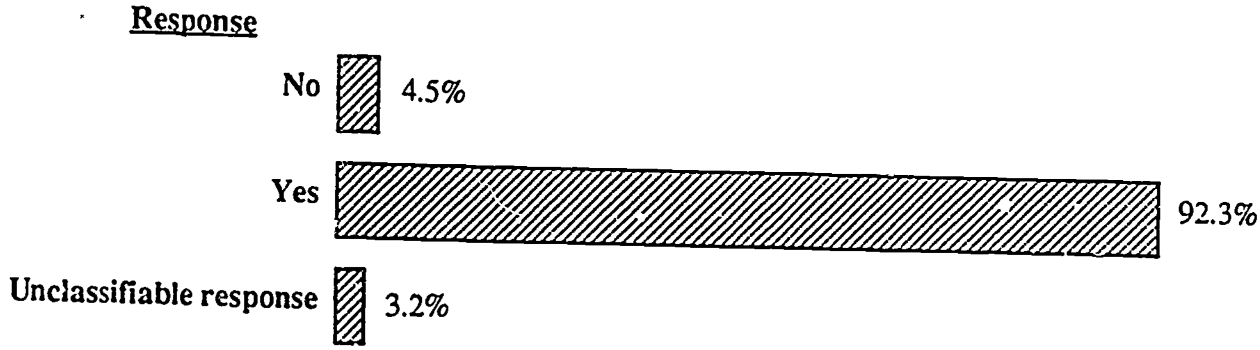
Response	Number
No	4
Yes--with conditions	12
Yes	177
Don't Know	5
Missing	117

Summary of 4d:

A typical response-"yes, if trained as I."

Question 4e: Should awareness and/or scoring be continued?

Response to 4e:



Response	Number
No	7
Yes	143
Unclassifiable response	5
Missing	160

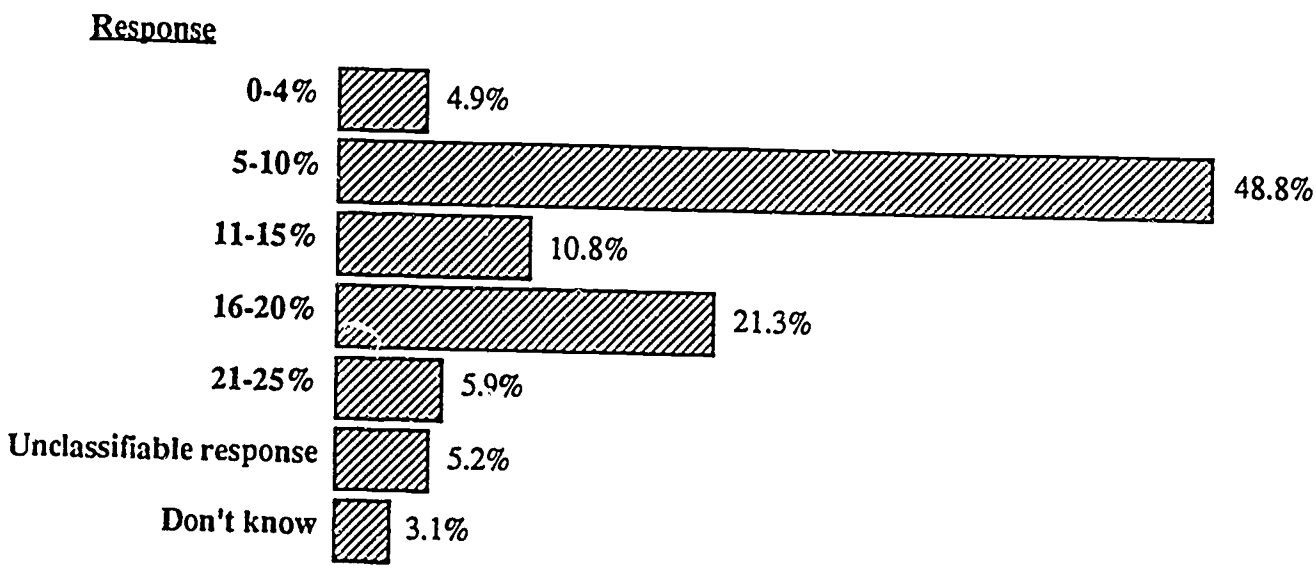
Summary of 4e:

Half of the teachers left this question blank. The response was almost all positive.

Question 5: Grading

Question 5a: What percentage of a "geometry" total score should a single focused holistic score count?

Responses to 5a:



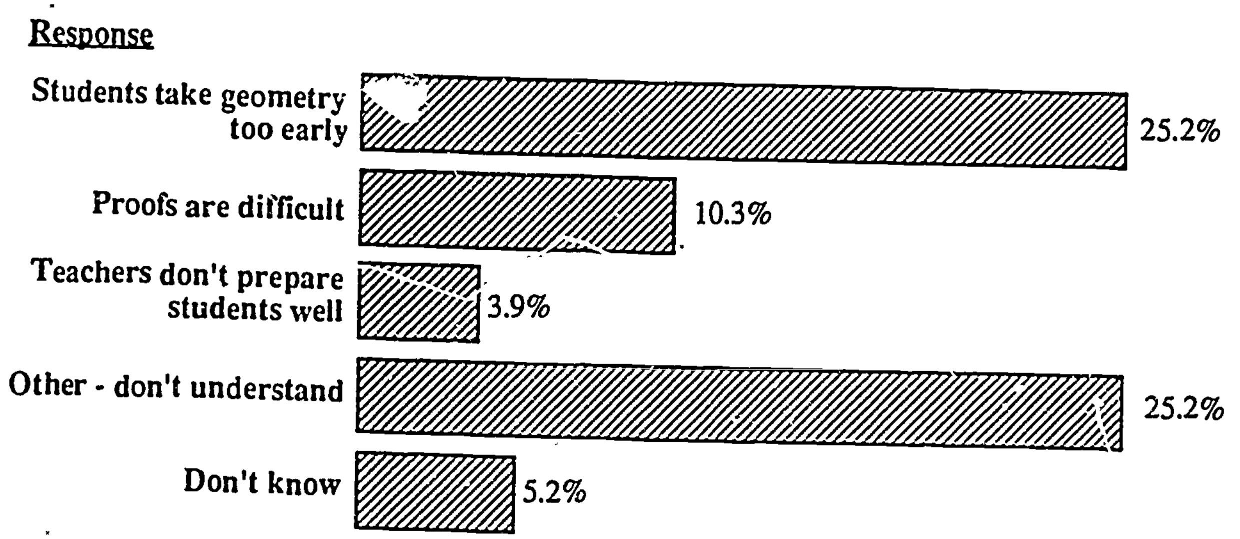
Response	Number
0-4 %	14
5-10 %	140
11-15 %	31
16-20 %	61
21-25 %	17
Unclassifiable response	15
Don't Know	9
Missing	28

Summary of 5a:

10 % was by far the most common answer.

Question 5b: Comment about the apparent results that one-third of students are receiving 0 or 1 scores when only half of students even take geometry.

Responses to 5b:



Response	Number
Students take geometry too early	39
Proofs are difficult	16
Teachers don't prepare students well	6
Other- Don't understand question	39
Don't Know	8
Missing	160

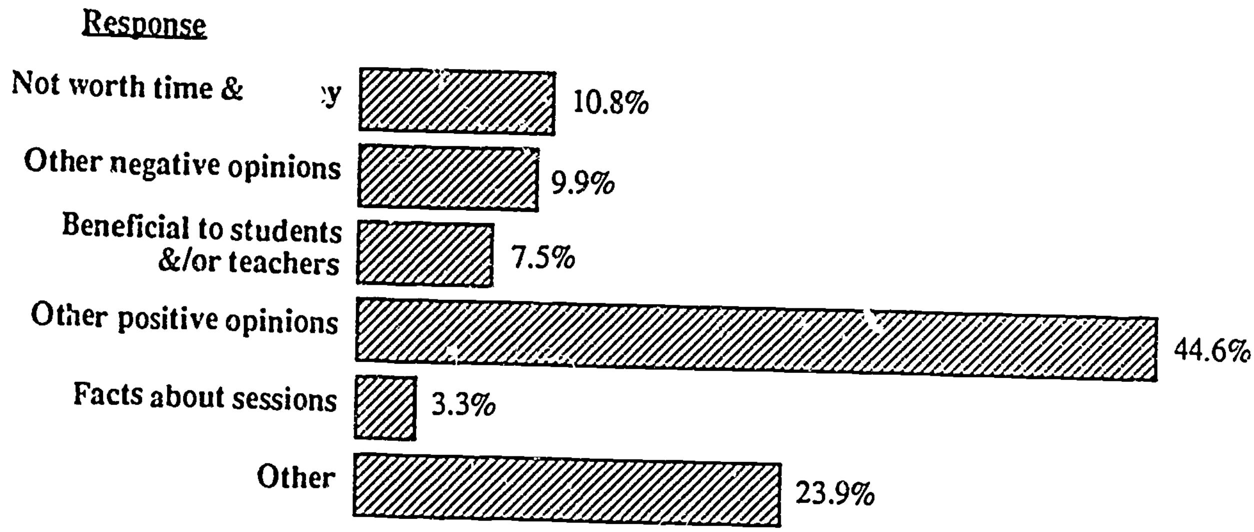
Summary of 5b:

The majority of teachers did not understand what they were supposed to be answering. Other teachers reported that the logic necessary for proofs is not yet developed in their students. There were several suggestions of classes without proofs for these students; the advanced students would be taught and tested on proofs in these classes.

Question 6: Other

Question 6: What would or do you plan to tell your principal, superintendent, or legislator about this effort?

Responses to 6:



Response	Number
Not worth time &/or money	23
Other negative opinions	21
Beneficial to students &/or teachers	16
Other positive opinions	95
Facts about sessions	7
Other	51
Missing	102

Summary of 6:

The responses to this question were positive in general. The most common response was that this testing should continue. Another theme that occurred frequently in response to this question was that the scoring was very exhausting but worth the benefits to teachers' instruction. Other teachers reported lost instruction time due to the scoring and due to testing as a whole.