DOCUMENT RESUME

ED 301 838 CS 211 614

AUTHOR Greenburg, Karen, Ed.; Slaughter, Ginny, Ed.

TITLE Notes from the National Testing Network in Writing. Volume VIII, November 1988.

INSTITUTION City Univ. of New York, N.Y. Office of Academic Affairs.

SPONS AGENCY Fund for the Improvement of Postsecondary Education (ED), Washington, DC.

PUB DATE Nov 88

NOTE 34p.; Abstracts of papers presented at the National Testing Network in Writing Conference (1988).

PUB TYPE Collected Works - Conference Proceedings (021) -- Collected Works - Serials (022)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Abstracts; Computer Uses in Education; Essay Tests; Holistic Evaluation; Scaling; *Scoring; *Writing Evaluation; *Writing Research

ABSTRACT

This newsletter contains 32 abstracts of approximately 1000 words each of papers presented at the 1988 conference the National Testing Network in Writing. Abstracts, listed with their authors, include "Instructional Directions from Large Scale K-12 Writing Assessments" (C. Chew); "Portfolio Assessment across the Curriculum: Early Conflicts" (C. Anson and others); "Revamping the Competency Process for Writing: A Case Study" (D. Holdstein and I. Bosworth); "We Did It and Lived: A State University Goes to Exit Testing" (P. Liston and others); "Proficiency Testing: Issues and Models" (G. Gadda and M. Fowles); "Creating, Developing, and Evaluating a College-Wide Writing Assessment Program" (S. Groden); "How to Organize a Cross-Curricula Writing Assessment Program" (G. Hughes-Wiener and others); "Presenting a Unified Front in a University Writing and Testing Program" (L. Silverthorne and P. Stephens); "Evaluating a Literacy across the Curriculum Program: Designing an Appropriate Instrument" (L. Shohet); "Validity Issues in Direct Writing Assessment" (K. Greenberg and S. Witte); "Reliability Revisited: How Meaningful Are Essay Scores?" (E. White); "Establishing and Maintaining Score Scale Stability and Reading Reliability" (W. Patience and J. Auchter); "Training of Essay Readers: A Process for Faculty and Curriculum Development" (R. Christopher); "Discrepancies in Holistic Evaluation" (D. Daiker and N. Grogan); "Problems and Solutions in Using Open-Ended Primary Traits" (M. C. Flanagan); "The Implications of Using the Rhetorical Demands of College Writing for Placement" (K. Fitzgerald); "Using Video in Training New Readers of Assessment Essays" (G. Cooper); "WPA Presentation on Evaluating Writing Programs" (R. Christopher and others); "Developing and Evaluating a Writing Assessment Program" (L. Boehm and M. A. McKeever); "The Changing Task: Tracking Growth over Time" (C. Lucas); "Assessing Writing to Teach Writing" (V. Spandel); "Reader-Response Criticism as a Model for Holistic Evaluation" (K. Schnapp); "The Discourse of Self-Assessment: Analyzing Metaphorical Stories" (B. Tomlinson and P. Mortensen); "The Uses of Computers in the Analysis and Assessment of Writing" (W. Wresch and H. Schwartz); "Legal Ramifications of Writing Assessments" (W. Lutz); "Some Not So Random Thoughts on the Assessment of Writing" (A. Purves); "Research on National and International Writing Assessments" (A. Purves and others); "Teaching Strategies and Rating Criteria: An International Perspective" (S. Takala and R. E. Degenhart); "Effects of Essay Topic Variations on Student Writing" (G. Brossell and J. Hoetker); "What Should Be a Topic?" (S. Murphy and L. Ruth); "Classroom Research and Writing Assessment" (M. Meyers); and "Computers and the Teaching of Writing" (M. Ribaudo and L. Meeker). (RS)

VOLUME VIII

NOVEMBER 1988

NOTES FROM THE NATIONAL TESTING NETWORK IN WRITING

A Project of The City University of New York and The Fund to the Improvement of Postsecondary Education

The National Testing Network in Writing, now in its seventh year, numbers 3,000 members across eleven countries. We are busier than ever collecting, cataloging, and disseminating information and data on measures and procedures used to assess students' writing skills. We are grateful for your help in sending us materials from testing programs--which, amazingly, are still proliferating, as questions about when, how, and whether to test writers continue to plague teachers and school administrators.

From the beginning, NTNW has sought to help members find answers to these questions. Our eight issues of Notes, the book, Writing Assessment: Issues and Strategies, and our annual conferences have attracted teachers, administrators, and assessment specialists from institutions around the world to examine models, and to explore the impact of assessment on pedagogy, curricula, and students.

The 1989 conference will be international in scope, featuring noted researchers from eight countries who will lead workshops and present their latest findings. The conference, co-sponsored by Dawson College. will take place Sunday, April 9th through Tuesday, April 11th (to allow for a weekend in Quebec) at the Centre Sheraton Hotel in Montreal, Canada. A new feature will be pre- and post-conference workshops. (See the centerfold of this issue for more information and a registration form.)

This issue of *Notes* continues the tradition of publishing abstracts from the annual conferences. The 1988 conference was co-sponsored by the University of Minnesota under the coordination of Chris Anson. Here are the abstracts of all of the workshops and panels, grouped according to themes: the first nine describe models of successful writing assessment programs, followed by eight that focus on models of scales and scoring; nine abstracts examine the impact of writing assessment on students, faculty, and curricula; and the final six examine current research on writing assessment.

The theme of the upcoming 1989 conference is "Writing Assessment Across Cultures." We hope you will join us in Montreal in April.

Karen Greenberg and Ginny Slaughter

We wish to thank the people who graciously contributed to this issue of *Notes*. First we thank all of the Recorders whose summary reports make up the bulk of the issue. We also thank Chris Anson of the University of Minnesota, Twin Cities for organizing a wonderful conference and Leslie Denny for carefully overseeing all of the local arrangements. And we are grateful to Haeryung Shin and Roddy Potter for their invaluable editorial assistance. Finally, our thanks to Harvey Wiener, University Associate Dean for Academic Affairs, and Director of CUNY's Instructional Resource Center, for supporting the production of NTNW *Notes*.

NOTES is published by the Instructional Resource Center, Office of Academic Affairs
The City University of New York, 535 East 80th Street, New York, New York 10021

CONTENTS

Page

MODELS OF EFFECTIVE WRITING ASSESSMENT PROGRAMS

Instructional Directions from Large Scale K-12 Writing Assessments: CHARLES CHEW 4

Portfolio Assessment Across the Curriculum: Early Conflicts: CHRIS M. ANSON, ROBERT L. BROWN, & LILLIAN BRIDWELL-BOWLES 6

Revamping the Competency Process for Writing: A Case Study: DEBORAH HOLDSTEIN & INES BOSWORTH 7

We Did It and Lived: A State University Goes to Exit Testing: PHYLLIS LISTON, JOHN MATHEW, & LINDA PELZER 7

Proficiency Testing: Issues and Models: GEORGE GADDA & MARY FOWLES 8

Creating, Developing, and Evaluating a College-Wide Writing Assessment Program: SUZY GRODEN 9

How to Organize a Cross-Curricula Writing Assessment Program: GAIL HUGHES-WIENER, SUSAN JENSEN-CEKALLA, MARY THORNTON-PHILLIPS & GERALD MARTIN 10

Presenting a Unified Front in a University Writing and Testing Program: LANA SILVERTHORNE & PATRICIA STEPHENS 10

Evaluating a Literacy Across the Curriculum Program: Designing an Appropriate Instrument: LINDA SHOHET 12

MODELS OF SCALES AND SCORING

Validity Issues in Direct Writing Assessment: KAREN GREENBERG & STEPHEN WITTE. 13

Reliability Revisited: How Meaningful Are Essay Scores?: EDWARD WHITE . 14

Establishing and Maintaining Score Scale Stability and Reading Reliability: WAYNE PATIENCE & JOAN AUCHTER 14

Training of Essay Readers: A Process for Faculty and Curriculum Development: ROBERT CHRISTOPHER 15

Discrepancies in Holistic Evaluation: DONALD DAIKER & NEDRA GROGAN 16

Probems and Solutions in Using Open-Ended Primary Traits: MICHAEL C. FLANAGAN 17

The Implications of Using the Rhetorical Demands of College Writing for Placement: KATHRYN FITZGERALD 17

Using Video in Training New Readers of Assessment Essays: GEORGE COOPER 20

THE IMPACT OF WRITING ASSESSMENT ON STUDENTS, FACULTY, AND CURRICULA

WPA Presentation on Evaluating Writing Programs: ROBERT CHRISTOPHER, DONALD DAIKER, & EDWARD WHITE 21

Developing and Evaluating a Writing Assessment Program: LORENZ BOEHM & MARY ANN MCKEEVER 22

The Changing Task: Tracking Growth Over Time: CATHERINE LUCAS 23

Assessing Writing to Teach Writing: VICKI SPANDEL 24

Reader-Response Criticism as a Model for Holistic Evaluation: KARL SCHNAPP 25

The Discourse of Self-Assessment: Analyzing Metaphorical Stories: BARBARA TOMLINSON & PETER MORTENSEN 25

The Uses of Computers in the Analysis and Assessment of Writing: WILLIAM WRESCH & HELEN SCHWARTZ 26

Legal Ramifications of Writing Assessment: WILLIAM LUTZ 27

Some Not So Random Thoughts on the Assessment of Writing: ALAN C. PURVES 28

CURRENT RESEARCH ON WRITING ASSESSMENT

Research on National and International Writing Assessments: ALAN C. PURVES, THOMAS GORMAN, & RAINER LEHMANN 29

Teaching Strategies and Rating Criteria: An International Perspective: SAULI TAKALA & R. ELAINE DEGENHART 29

Effects of Essay Topic Variations on Student Writing: GORDON BROSSELL & JIM HOETKER 30

What Should Be a Topic? SANDRA MURPHY & LEO RUTH 31

Classroom Research and Writing Assessment: MYLES MEYERS 32

Computers and the Teaching of Writing: MICHAEL RIBAUDO & LINDA MEEKER 33

INSTRUCTIONAL DIRECTIONS FROM LARGE SCALE K-12 WRITING ASSESSMENTS

Speaker: *Charles Chew*, New York State Department of Education

Introducer: *Marie Jean Lederman*
Baruch College,
CUNY and NTNW

It is now generally agreed that (1) direct assessment of writing should, if possible, approximate what we expect when students write; (2) learning to write is a process which takes place over time, sometimes recursively; and (3) requiring students to write whole discourses is a better assessment tool than the objective test of discrete skills. In 1979, when New York State first instituted a writing competency test including three writing samples, it was virtually alone in its attempt to assess students' writing ability through multiple writing samples. Five years have passed since the initiation of the first Regents Competency Test in Writing. The program has grown to encompass not only the eleventh grade but the eighth grade and fifth grade as well. Since September of 1985, G.E.D. diploma candidates must also write an essay.

The Writing Test for New York State Elementary Schools administered at grade 5 comes very close to approximating the composing process. Students are required to write two different pieces for the test on two different days. A prewriting section precedes the writing sample and is not evaluated. Students draft a response and then redraft. In the tests at the secondary level, the Preliminary Competency Test in Writing at grade 8 requires students to write three pieces, as does the Regents Competency Test in Writing which is administered in grade 11 and is a requirement for graduation. The Comprehensive Examination in English, which is administered usually to average or above average students requires two writing samples.

If tests are to approximate the reality of the writing process and are to have a positive effect on instruction, they need to require many types of writing . The outline below shows the types of writing assessed in New York State Writing Assessment programs.

Types of Writing

Writing Test for New York State Elementary Schools

Personal Narrative	The writer recounts an experience which he/she had.
Personal Expression	The writer recounts a feeling or an emotion.
Description	The writer describes a person, object, or place.
Process	The writer explains how to do something.
Story Starter	The writer completes a story which is started in the writing prompt.

Preliminary and Regents Competency Test in Writing

Business Letter	The writer at eighth grade writes a letter ordering something. At grade 11, the writer composes a letter of complaint and suggests how to remedy the situation.
Report	The writer takes data supplied and prepares a report for another person or the class.
Persuasive Discourse	The writer attempts to persuade the reader to take some action by stating the action to be taken and giving reasons why such action should be taken.

Comprehensive Examination in English

Essay	The writer, using literature which has been read, responds to a given question which is generic in the sense that a wide variety of literature could be used in the response.
Composition	The writer can choose to do one question from among eight. Two of these are situations which provide a purpose and audience. The other six are discrete topics which require a full rhetorical invention by the writer.

The methods of evaluation used in this testing program also speak to instruction. Students' writing samples are evaluated holistically at grade 5, and modified holistic scoring is used with all the other tests. This rating procedure delivers a message to anyone in the state who is involved with the testing program. The idea that the whole piece of writing may be worth more than any one single feature is an important message to teachers who have for years spent an inordinate amount of time "red-penciling" errors in students' work. Many of the criteria used to evaluate the writing samples are virtually the same for grades 5-11, indicating that these elements are seen as essential in a competent piece of writing. The fact that the criterion focusing on mechanics is not at the top of the list reminds teachers that mechanics, although important, are not the "be all and end all" of written discourse. The fact that the tests are unlimited in time and that length is merely suggested delivers additional messages about the teaching of writing.

Any student who fails below the State Reference Point on the writing tests is required by the state to receive additional or remedial instruction. Parents must be notified of the student's grade and must be informed of the remedial program established for the student. These programs must begin no later than one semester after the administration of the test. Students can be removed from remediation if it can be documented that deficiencies have been overcome. At the senior high school level, students must pass the Regents Competency Test in Writing in order to receive a diploma.

To meet the needs of educators at the local school level in rating the tests and to devise instructional strategies to meet student needs in writing, the Bureau of English and Reading Education developed a two-year in-service program. The first phase of the program identified fifty key teachers or supervisors, representing geographic areas of the state, who came to Albany for a two and one-half day intensive training program. This program focused on rating procedures, developing a workshop agenda, and actually simulating the role of workshop leader. When these fifty people returned to their local areas, they in turn trained teachers from the local schools who were involved directly with students affected by the writing tests. The success of the program was confirmed by the evaluations done by workshop participants. The sampling by the Bureau of ratings of test papers done locally attested to the reliability of local rating.

Because of the success of our assessment program, there is a reluctance to make changes. In New York we have sensed a need to change the examinations for a number of years. After extensive discussion, pretesting and field testing, changes in the examinations will begin in the 1988-89 school year. Part III of the Preliminary Competency Test will be changed to reflect the revised composition curriculum for New York State, and the purposes for writing will rotate among those covered in this curriculum material. Evaluation of the samples will no longer require model answers. Although rating will be done in much the same way, raters will rely on criteria only. This change goes into effect for both the PCT and RCT. In January, 1989, the format of the business letter on the Regents Competency Test in Writing will change and information needed by the test taker will be in note or outline form. This change will require the student to process the demands of the task and formulate a response rather than simply reword the task. In January 1992, Part III of the RCT will change to follow the change begun in the PCT.

I conclude by pointing out some problems, questions, and concerns which still need to be addressed by test makers and others interested in improving students' writing ability. These are as follows:

(1) Samples of students' writing for evaluation and instructional purposes must be obtained throughout the school year, not only at test time.

(2) Research needs to focus on the development of writers over time.

(3) Research needs to determine if skills differ appreciably for various types of writing in a test situation.

(4) Research needs to ascertain the relationship between writing done during a test and that done by the student at other times.

(5) Writing prompts may not tap the experience of the writers.

(6) Instruction can be limited to test items. Students may spend an inordinate amount of time writing business letters and structured responses to literature.

(7) Evaluators using holistic scoring may not appreciate the fact that more must be done with student papers to plan instruction.

(8) Once common elements of competent writing have been identified, instructional strategies need to be developed which will enable teachers to focus on these elements in a total language approach.

(9) In-service programs connected to a test may be limited when compared to extensive needs of teachers.

Those of us involved in the assessment of writing know how much more we know today than we knew just a few years ago, but there is still much to be learned.

PORTFOLIO ASSESSMENT ACROSS THE CURRICULUM: EARLY CONFLICTS

Speakers: *Chris M. Anson, Robert L. Brown, Jr., and Lillian Bridwell-Bowles*, University of Minnesota

Introducer: *Virginia Slaughter*, CUNY

Faced with a mandate to begin assessing students' writing at the University of Minnesota, members of the Program in Composition and Communication there finally convinced an interdisciplinary task force that a cross-curricular portfolio assessment would be the only way to bring about large-scale changes in the quantity and quality of writing instruction beyond their own writing program. In this session, the speakers shared pieces of an ongoing cultural critique that focuses on the political, curricular and ideological contexts in which they are struggling to turn a potentially damaging process into a method for empowerment, enrichment, and educational change.

Currently, the University of Minnesota plans to require applicants to submit a high school portfolio as part of the admission requirements. These portfolios require samples of writing from several subject areas as a way to encourage writing across the curriculum in the high schools. Throughout their college years, students will continue to build on their portfolio until they are juniors, at which time their major department will be responsible for assessing the quality of their writing for exit from junior-year status. Increased attention to writing, including new composition courses, writing-intensive courses across the curriculum, and trial assessments before the junior year, will provide support for the assessment program. Composition faculty will take on a greater consultative role to help departments incorporate writing into their curriculum and to help them establish methods for the portfolio assessment.

Chris Anson described the University's plans for this assessment as these are outlined in the 1987 report of the Task Force on Writing Standards. Reactions to the report were solicited from departments and colleges at the University of Minnesota, from 143 secondary school teachers and administrators across the state, and from assorted other readers, including personnel at the Minnesota State Department of Education and local professionals. Anson's close reading of these readers' responses to the report revealed a more positive attitude toward instruction among secondary teachers than among teachers at the University itself. In comparison to college faculty, the secondary teachers showed a deeper understanding of the relationship between testing and teaching, expressed fewer fears about increased workload, and worried more about the potential hazards of testing when it does not support enhanced instruction. Using quotations from several responses, Anson showed how faculty members' views of writing assessment are not only saturated by their tacit endorsement of the surrounding academic values of their institution, but also by the more specific ideological perspectives of their discipline.

Anson explained the resistance to portfolio assessment among the college faculty by describing the institutional ethos at Minnesota, a university that privileges research and publication and de-emphasizes undergraduate education. After exploring some of the ideological reasons why university faculty resist rich types of assessment and accept simplistic types (such as multiple-choice tests of grammar skills), Anson argued that before a writing assessment program can be implemented successfully, administrators must study and understand the academic culture that surrounds the planned assessment. Armed with this knowledge, administrators can plan ways of implementing rich assessment programs without facing the sort of resistance that can lead to impoverished tests and instructional decay.

Central to these understandings is an awareness of the relationship between writing programs and the larger academic culture. Composition teachers and administrators in radical writing programs are change agents, whose political praxis must be consciously grounded in theory or run the risk of becoming ineffectual, or worse, of merely reinscribing the ideologies they seek to change. Beginning with this premise, Robert Brown set out to raise theoretical questions central to such praxis. An adequate theory, he claimed, would be hermeneutic, and might take as its text the university itself, in its several manifestations: the behaviors of its members, its constituting texts, and its organizational structures. The university-as-text speaks of knowing and knowledge: their nature, value (economic and otherwise), creation and social utility. We might profitably read this text through the reciprocal processes defined in radical ethnography. If we do, we can simultaneously explicate the bureaucratic forces we encounter in attempting to build genuine literacy programs, and our own culture-specific ideologies.

Creating change requires ongoing dialogue across the curriculum about such issues as standards vs. individuality and creativity; program assessment vs. individual growth; and the place of writing instruction in the rise of the new professionalism vs. the liberal arts education. Arguing that change is possible with the right incentives for faculty, Lillian Bridwell-Bowles concluded the session by outlining some of the assessment activities underway at Minnesota. These include a study of "strong, typical and weak" writing samples across the undergraduate curriculum, studies of writing in "linked courses" which combine composition instruction with content learning, and planning the implementation of portfolios as a requirement for admission. The newly endowed Deluxe Center for interdisciplinary Studies of Writing will provide ongoing research funds for faculty interested in five categories: the status of writing ability during the college years; characteristics of writing across the curriculum; the functions of writing in learning; characteristics of writing beyond the academy; and curricular reform in undergraduate education. Other efforts to improve the context for the planned assessment include early pilot projects for portfolio assessment that have been conducted in 18 Twin Cities Metropolitan school districts, and collaborative writing assessment projects that are part of the Alliance for Undergraduate Education, a consortium of 13 public research universities.

REVAMPING THE COMPETENCY PROCESS FOR WRITING: A CASE STUDY

Speakers: *Deborah Holdstein*, Governors State University, Illinois
Ines Boswoth, Educational Testing Service, Illinois

Introducer/ Recorder: *Lu Ming Mao*, University of Minnesota

Deborah Holdstein began by describing Governors State University, a junior, senior, and graduate institution with diversified student body. The University used to have a writing competency test for prospective juniors and seniors. Accompanying this competency test were a set of grading standards and a pass-and-fail system, both of which had continually drawn criticism from test readers and scorers alike, because they were vague and not academically sound. According to this set of standards, a passing essay must (1) respond to the stated topic; (2) have a clearly stated thesis; (3) show clear, logical organization of ideas in organized, well-developed paragraphs; (4) include supporting details; (5) demonstrate one's editing ability.

Holdstein was asked to change this system. She observed that the process of revamping an assessment program was as political as it was academic. One misperception bandied around a lot was that the English teachers were determined to flunk students. Holdstein recalled that they needed someone from outside, an expert with no stake, political or otherwise, in the system, to help teachers revamp the system. Ines Bosworth from Educational Testing Service was brought in as a consultant. Bosworth emphasized that as a neutral observer, she was able to get different opinions from faculty in different departments. These discussions became extremely useful because they enabled faculty to articulate their concerns about possible changes in the testing program. Out of these discussions--and the Provost's unfailing support--came the new scoring criteria, which have four major areas: focus, organization, elaboration (support) and conventions (mechanics). These are scored with a 6-point scale, 6 being superior and 1 being seriously inadequate. This scale replaced the old pass/fail scale.

Holdstein noted that the number of questions on the test was reduced from 5 to 3 (although the test time is still 60 minutes) in response to students' complaints that the number of tasks on the old test forced them to spend a lot of time reading and figuring out questions instead of actually composing. One of the three new questions reads as follows:

> "Matrimony is a process by which a grocer acquired an account the florist had." What does this quote say about the transition from single to married life? Is it accurate? How so -- or how not? Again, be sure to formulate a thesis with your point of view, and use specific examples to back up your points.

Both speakers noted that one of the many merits of this new competency test is that readers can more easily score each essay according to the criteria. Moreover, the new system is fairer than the old one. Under the old system, whenever there was a split in "failing" or "passing" decisions, a 3rd reader was consulted. Under the new system, each essay receives four readings, and readers do not know whether they are the 3rd or 4th reader; thus, a lot of political heat is removed. Readers also provide students with an "analytic checklist," which informs them of the criteria used, the weaknesses in their essays, and comments from the readers.

Bosworth commented that the interrelater reliability of the new test is 92% (as opposed to 73% in the old test) and that more students have passed the new competency test than before. However, Holdstein pointed out that most questions in the new test tend to be too content-laden and that the scoring criteria are too heavily weighted toward content. Nevertheless, both speakers noted that the new test has proven to be far more effective than the old one and has fostered faculty collaboration.

WE DID IT AND LIVED: A STATE UNIVERSITY GOES TO EXIT TESTING

Speakers: *Phyllis Liston, John Mathew, Linda Pelzer*, Ball State University

Introducer/ Recorder: *Joyce Malek*, University of Minnesota

In Fall 1987, Ball State University (Muncie, Indiana) implemented exit testing for writing competency as a prerequisite for graduation. The three member panel responsible for establishing the rubrics and coordinating the testing and holistic grading discussed what they learned during this first year. Participants were given hands-on experience with the exam by writing briefly in response to a sample writing test essay assignment and discussing the process we went through to begin answering the essay question. They then ranked actual essays and were led through the process the panel uses to develop the rubric.

Phyllis Liston began by describing what the exam coordinators learned in the process: (1) implementing, coordinating and gaining community-wide acceptance for exit exams is "a lot harder than it looks"; (2) communication at all levels is essential; (3) low-level mistakes can cause high-level difficulties; money when needed is found; and (4) holistic grading works well. In addition, the exam needs full administrative and faculty support. As the director of the writing competency exam, Liston found the administrative duties to be a full-time responsibility requiring personnel assistance.

Liston explained Ball State's "3/3/3" exam process. Students sign up for the exam three weeks before the exam date and are given an instruction sheet detailing the exam process, the exam question, how to prepare for the exam, and where to go to receive help preparing for the exam. On the exam date, students are given three hours to write approximately three pages in response to the exam question. Students must pass the test to graduate. After two attempts, they are required to enroll in--and repeat until they pass--an upper division writing course. The second opportunity to take the exam constitutes an automatic appeal. Exit from the course is by portfolio prepared by the students with the help of their instructors. Portfolios are evaluated by two or more readers other than the classroom teacher/coach. No student takes the exam more than twice.

John Mathew explained the training process for holistic graders by taking participants through a mini grading workshop. We read and ranked three sample essays high, middle and low. Then we read, ranked and integrated into the previous essays three more, and did the same for two additional essays. We then discussed our ranking of one of the essays in terms of its strengths and weaknesses. Finally we were presented with the six-point rubric developed by the panel for the particular exam and were asked to rate the essay.

In an actual reading, graders read ten papers at a time, assess, record and score, and pass the papers on to a second reader. Papers with scores that do not match are given to a third reader. All pass decisions are made by the University Provost under the advisement of the panel and other administrators after all exams for the quarter have been scored. The panel acknowledges a high reader calibration and suggests a main reason for it is that readers do not know the cut-off point for failing, and therefore are more objective and not sympathetically influenced to pass a border-line paper.

Linda Pelzer described the rubric design process. The panel develops a new rubric for each exam by reading and sorting all essays written for the exam into high, middle and low categories. After sorting, they discuss the categories and write about them, and they draft a six-point rubric--one that is quite detailed and descriptive and that includes specific examples from student papers to illustrate the rubric's categories. A six-point rubric is used because it eliminates a middle score and because a four-point rubric would not be specific enough to encompass the aspects of the writing they wish to assess. The panel takes care and time in designing the rubric to make it clear and specific in order for readers to reach consensus and to withstand criticism from students, parents and faculty. Rubrics are kept on file at the University library. One indicator of the success of the rubric is that students who fail the exam and wish to contest it usually reach agreement after examining the rubric and evaluating their own writing against it.

Although the writing competency examination project is bigger than the panel first anticipated, they agree that it is worth the work.

PROFICIENCY TESTING: ISSUES AND MODELS

Speakers: *George Gadda*, University of California, Los Angeles
Mary Fowles, Educational Testing Service, New Jersey

Introducer/ Recorder: *Adele Hansen*, University of Minnesota

George Gadda opened the discussion with a statement concerning general issues in developing a proficiency testing program. Proficiency testing, like achievement testing, measures success in a particular domain. There are several motivations for proficiency testing: to certify individual achievement exclusive of grades, to validate a program's effectiveness, or to screen before certification of passing to the next level of instruction. The choice of purpose governs the rest of the assessment program. Proficiency tests may be used to exempt students from further work; to prove value added in a course program; to permit passage, graduation or certification; or to identify those who need further instruction.

Gadda noted that test-makers should define the domain of the test by describing the kind of written ability being assessed and that we should make a public statement concerning the criteria used for judgment. Tests used for advancement should be a well-defined part of the curriculum, with samples and grading criteria clearly described. Ideally, scorers should be those people who are testing and using the results. In addition, we need to determine what will happen to those who don't pass. Gadda noted that proficiency tests should not be a "roadblock." He concluded by stating that we should strive for high reliability and validity in our testing because proficiency tests need to withstand legal challenges.

Mary Fowles remarked that we need an increased understanding of what is to be tested and that the "community" must share the same standards. She referred to a project in Rhode Island, where a state administrator decided to work on literacy beginning in the third grade. ETS was asked to construct a test that encouraged good writing. They worked :·:th local administrators and teachers from every school district in the state to formulate a writing test which was administered to all 3rd graders. The test featured a pre-writing section and then an essay test. It also included an editing phase, where students were given specific questions about content.

Fowles described how scorers were trained· Every district in the state was represented in training sessions, and benchmark papers were identified and then used to train local raters. After the results were tabulated, the teachers returned to the classroom and showed examples of good papers to the students and discussed the scoring criteria. Next, the state decided to develop a portfolio of such "assignments" to validate the scores on the "test" and to enhance teaching.

In the discussion that followed, questions were raised concerning the "read and respond" type of test. Gadda agreed that such a test does assess reading as well as writing, but that there is a connection and such tests are useful to determine the students basic ability to do university level work. :He added that such tests seem most fair, because all students begin with the same information and the students can then better understand the testing situation. He cautioned that such tests should always be pre-tested to discover if the reading is "accessible and interesting" and if the assignment elicits more than one response, because this can affect raters' evaluations.

CREATING, DEVFLOPING, AND EVALUATING A COLLEGE-WIDE WRITING ASSESSMENT PROGRAM

Speaker: *Suzy Groden*, University of Massachusetts, Boston

Introducer/ Recorder: *Geoffrey Sirc*, University of Minnesota

In this session, Suzy Groden reported on the University of Massachusetts' writing assessment program, on-going since 1978. She described how it was developed, changed, and validated. The exam is a "rising junior exam," required of students after 68 credits (or within first semester for transfer students). Called a test of writing proficiency, the exam really tests reading, writing, and critical thinking because students have to respond to questions on texts (or "reading sets") with which they are provided one month prior to the exam. Students are either judged proficient or must remediate their writing skills.

The idea behind the test is to teach students what the faculty want them to know in various core curriculum courses, courses designed to include elements of critical analysis and reading/writing associated with that discipline. Each reading set is 20 pages and concerns a controversial topic associated with a specific discipline. Students chose one of three sets, from natural sciences and mathematics, the social sciences, or the humanities, and they read the set for a month After the first exam was given in 1978, a sample for students became available and a student manual was developed.

Groden stated that one problem in the exam is the lack of a penalty for those who fail. The exam is graded by readers who are trained in one morning and then read exams all afternoon. A student needs two readers to agree in order to pass, and three readers to agree in order to fail. But there are actually no practical penalties now associated with failing the exam: students can still take upper-division courses if they fail, and there is now an alternate way to demonstrate proficiency--a portfolio.

During the course of subsequent years, changes occurred in the context of the exam. After Groden and the university's ESL Director became involved, an interest in writing and the acquisition of language found its way into the readings. Policies surrounding the implementation of the test were gradually loosened. The use of the portfolio alternative was extended, particularly to ESL students. Also, the range of writing samples included in the portfolio was expanded to include more than just the traditional analytical paper: lab reports, for example, would be accepted. Students were allowed three hours to write the exam, rather than just two. And in one of her more striking findings, Groden found students wrote much more easily when they switched from the standard-size blue books to the larger, 8- and-1/2 inch size (that being the standard in which they most frequently composed). The exam committee also spent more time thinking about readings and questions; the exams became more complicated, involving ideas about the nature of knowledge. What ultimately evolved were two possible questions, one for the non-intellectual and one for the more challenging intellect. Finally, they also offered an evening session for taking the exam.

There were also many changes over the years which Groden termed losses. Faculty involvement waned, with more and more responsibility for grading falling to the exam committee. The school changed, taking in fewer freshmen and more transfer students, with the exam becoming a kind of graduation test. Funding dried up, causing the university to retreat from its core curriculum and limit the number of its core courses, and, hence, severing the relationship between the curriculum and writing proficiency.

One area in which the Massachusetts exam developers were successful was in establishing grading criteria. Sending exam samples to national experts, Groden received strong validation and agreement on their criteria. The experts, however, were critical of the number of questions, feeling they made different cognitive demands and were unfair. The committee is continuing to revise the exam.

HOW TO ORGANIZE A CROSS-CURRICULA WRITING ASSESSMENT PROGRAM

Speakers:	*Gail Hughes-Wiener, Susan Jensen Cekalla, Gerald Martin , Mary Thornton-Phillips,* Minnesota Community College System
Introducer/ Recorder:	*Julienne Prineas.* University of Minnesota

The speakers began the session by describing how, with the support of a Bush Foundation Grant, the Minnesota Community College System (comprised of 18 two-year college scattered across the state) has been engaged in a three-year project to assess the effects of their Writing Across the Curriculum (WAC) program on faculty and students, especially on student learning. The four speakers described their separate but overlapping roles in the project, with a view to communicating the complexity of implementing this type of project. Mary Thornton-Phillips' role has been to design and establish the broad structure of the project. Susan Jensen-Cekalla, as the WAC Program Coordinator, has served as a bridge between the evaluation project and the faculty out int he colleges. Gail Hughes-Wiener's role as Evaluation Coordinator is to ensure that all of the components of the evaluation-- such as surveys, interviews, essay exams and such--are designed, coordinated, implemented, analyzed and communicated. Gerald Martin, as research analyst, is in charge of processing the data.

Hughes-Wiener pointed out the need to budget for an immense amount of administrative, interpersonal, and program development required prior to any actual data analysis or report writing. Her experience has been that no one, including consultants prominent in the field of program evaluation, anticipated the amount of work and time needed for this preparatory work. The scope of the project demonstrates its complexity: data must be collected on faculty attitudes, student attitudes, and student learning. The project required the careful development of questionnaires and surveys, the effective training of interviewers, and the preparation of holistic scoring terms. In addition, the project leader had to build credibility and trust among program participants and become knowledgeable about all needed information.

Thornton-Phillips commented that their progress has been aided by a clear sense of direction, despite uncertainty as to how to achieve their goals. Allowing for flexibility and change within a general framework has proved necessary and productive. For example, faculty involvement was essential to the success of the project. Faculty had to become trained, knowledgeable participants who understood the research and their role in it. Thus, Thornton-Phillips' first challenge was to assess the needs and interests of faculty in an attempt to generate strong staff commitment and to develop a core faculty able to provide leadership for the program. The task was hindered by the voluntary nature of staff development in the Community College System and by the tendency to cut funds for such development during budget crises. Thornton-Phillips found the catalyst for the change needed in a dedicated Joint Faculty/Administrative Staff Development Committee and in a small group of faculty who had worked together for several years on implementing "Writing Across the Curriculum." Jensen-Cekalla joined the team as program coordinator, leaving Thornton-Phillips free to work on budgeting, staffing, and scheduling aspects. Together, they refined the assessment component and developed reliable approach to reassure faculty.

In her role as the most direct connector between faculty and the evaluation project, Jensen-Cekalla's foremost concern has been that all participants work together and coordinate their efforts. A cornerstone of the project is a summer workshop, which brings together faculty from all eighteen colleges. Follow-up meetings during the year provide the support and opportunity for exchange of information needed to maintain a united WAC teaching approach, and the grant provides all teachers with funds for a variety of supportive measures such as tutors, materials and supplies for the Learning Centers, and outside and in-house consulting. Jensen-Cekalla has also had to organize the data flow of information and resources from the evaluation out to people in the colleges.

Martin's roles have included data analyst, data processor, in-house statistical research consultant, and resident skeptic. With the project now in its fourth year, the time has arrived to renew the research grant and inform the granting agency of the progress made. Martin noted that a project of this type raises many issues along the way. Its original purpose was to look at student outcomes, such as specific changes in writing proficiency and the learning of subject matter. However, several other desirable outcomes not in the original proposal have become obvious. Hughes-Wiener noted that they have learned, for example, that faculty enthusiasm for WAC can be generated and that inroads into the organizations at both campus and administration levels can be made.

PRESENTING A UNIFIED FRONT IN A UNIVERSITY WRITING AND TESTING PROGRAM

Speakers:	*Lana Silverthorne,* University of South Alabama *Patricia Stephens,* University of South Alabama
Introducer/ Recorder:	*Gail A. Koch,* University of Minnesota

How can we foster institutional consensus about undergraduate writing in a university? Lana Silverthorne and Patricia Stephens answered this question by focusing on university-wide participation and dialogue. They described a multilateral commitment to undergraduate writing that has grown incrementa'ly over the past seven years at the University of South Alabama. The primary agent of this progress has been the continuous participation of faculty from various disciplines, especially in the construction of an upper-level writing across the curriculum (WAC) program .

The impetus for ongoing development of the upper-level WAC program has been a week-long summer seminar for faculty across the disciplines. including one representative from each of the undergraduate departments. It has been repeated annually since 1981. The seminar work is guided by Director of the University Writing Program and by an outside consultant. The participants write and talk about the purpose of writing in their junior and senior courses. They get acquainted with the practice of continuous "writing-to-learn" and with its potential uses in their courses. They put together a proposal for a sequence of "writing-to-learn" assignments to be tried and revised in their own courses over several quarters, and they review each others' WAC proposals. They learn ways of responding to students' efforts to "write-to-learn."

According to Silverthorne, the WAC seminar, first conceived as a means to convert, has by now become a forum for faculty leadership. Participants become the teachers of upper-level content courses designated as writing courses. By now, at least half of the faculty are teaching such courses. (Students are now required to take two such courses, one in their major, and there are now about 70 such courses available each quarter.) WAC-experienced faculty influence the criteria by which a content course can be designated as a writing course. They give precedence to continuous writing in content courses over production of the "one-shot" term paper, and they sanction "discovery" writing which encourages students to "bring their own experiences to bear upon subject matter."

Silverthorne noted that holistic assessment of essays composed by transfer students who have had Freshman English elsewhere has provided a second opportunity for building consensus at the University of South Alabama. Piloted in 1983, the test has recently become a requirement. The test prompt mirrors the emphasis on personal writing in the University's first quarter of lower-level composition and on the exit test given at the end of this first quarter of writing. Students are given a choice of three prompts. They have two hours to write with dictionaries and handbooks. Students are informed of the general criteria by which their essays will be judged. Each paper gets three readings, and the evaluation determines whether or not a tested transfer student starts in the first of the University's writing courses. Since 1983, about 75 percent of the students have passed the test The transfer test essays are assessed by cadre of faculty readers from various disciplines who teach the upper-level content courses designated as writing courses. Their decision is to pass or fail an essay. If an essay arouses irresolvable ambiguity in one reader, it is passed on to two additional readers for the pass/fail decision.

Records on this assessment process bear out the claim of active university-wide participation of faculty. Between the fall of 1986, fifty-six faculty have served as readers, about 71% of them from the professorial ranks. Their distribution by department or discipline shows variety: 7% Business; 34% English; 9% Humanities and the Arts, 18% Medical Sciences and Nursing; 14% Natural Sciences and Engineering; 18% Social Sciences and Education. The records also show high inter-reader agreement. Figures over twelve quarters between the fall of 1983 and the fall of 1987 show the average rate of agreement to be 87.2% in the first year. The local reading was tested against the judgment of external readers. With the help of the NTNW, a study was conducted to compare the assessments of five local readers to that of three readers at CUNY. The rate of agreement between he two groups overall was nearly 80%.

Patricia Stephens took up the matter of the reasons for the high degree of consensus in this assessment process. She cited the quality of the WAC seminars, the credibility of the program director,and administrative support and incentives.Faculty who are developing a new upper-level writing-designated content course are released from teaching one course, and the enrollment in their writing-designated content course is reduced to 25. A participant in the week-long WAC seminar is paid $400; a reader for the transfer test essay who, on an average, judges 35-40 papers, receives an honorarium of $50. Stephens stressed the importance of the faculty's common concern for students' development as effective writers, underscoring Silverthorne's contention that drawing upon faculty from various disciplines creates a university-wide sense of responsibility for the quality of students' writing and fosters a continuing university-wide dialogue about writing standards.

The continuing dialogue is crucial. Stephens described "calibration sessions". In these sessions, readers consider their common purpose of helping students to improve their writing and discuss the general criteria or qualities by which they decide to pass or fail a test essay in relation to this common goal. There are four qualities, a number kept small on purpose, to head off a penchant"read for everything we know in our various disciplines." The naming of the criteria, too, is kept simple and true to the holistic assessment principle of reading for general impression: Invention (Has the writer of the essay been thoughtful, reflective, candid?) Arrangement (Has the writer achieved wholeness, made a piece of it?) Development (Has the writer recognized and fleshed out the point of the essay, giving it credibility and validity?) Style (Does the essay have clarity, give evidence of the writer's own voice, the writer's own crafting, and editing?).

The dialogue amongst faculty continues through instructional use of carefully kept records. Results of inter-reader reliability and validity studies are shared with readers to help them evaluate their own reading performance in relation to that of the others. Readers are given detailed information about the results of their own decisions, a statistical summary of each reading session, and a cumulative summary of all reading sessions. In addition, the readers are rated and their ranking reported to them. They are rated on three bases: experience, reliability, and validity (or the fit between their judgments and other information about students such as GPAs and ACT scores). In short, readers have regular, informed opportunities to reflect upon the relative fit of their judgment with the consensus.

One last piece of information about the consensus reported by Silverthorne and Stephens is that the membership of the transfer-test reading group is stable, the chief movement being the addition each year of two new members from the summer WAC seminar. Once having assumed the role, very few have ever repudiated it. Stephens pointed out that it is in the faculty's interest to be involved: reading the test essays serves as a useful means by which faculty who teach writing-designated junior and senior courses can gauge students' readiness to deal with the "writing-to-learn" orientation of their courses.

EVALUATING A LITERACY ACROSS THE CURRICULUM PROGRAM: DESIGNING AN APPROPRIATE INSTRUMENT

Speaker:	*Linda Shohet*, Dawson College, Montreal
Introducer/ Recorder:	*L. Lee Forsberg*, University of Minnesota

Linda Shohet, director of the Literacy Across the Curriculum Center at Dawson College, has taught Canadian literature and writing at Dawson since 1973. She began developing the Literacy Across the Curriculum program in 1984; the Center now provides instructional and consultation services to English high schools and colleges throughout the province. She began the seesion by reviewing the language-related political issues in Quebec, and then she sketched the development of the center and discussed the evaluation of the program scheduled this spring at Dawson.

Quebec is a unilingual province in a bilingual country. French (first language) speakers comprise 24.6 percent of the population of Canada and 83.5 percent of the population in Quebec. English (first language) speakers comprise 68.2 percent of the population in Canada and 12.7 percent of the population in Quebec. French speakers see the maintenance of their language politically, as the survival of their culture. Consequently, language awareness is high.

Dawson College is a two-year, English language community college; all students going on for a University degree must first complete community college. The Literacy Across the Curriculum program was initiated by the faculty development committee, not the English Department, and its administration remains in the faculty development office. Keeping it out of the English department gives the program a broader base of support and institutional commitment, Shohet said. The program, originally intended as internal, soon started receiving requests from English-language high schools and other colleges, asking for ideas and resources. As the program expanded to meet those needs, costs rose. The only source of additional funding was the government, which required evidence that the program was relevant to the entire community, including French-language schools and colleges. The Dawson College administration had ordered an evaluation of the program to show its value to its own faculty before supporting expansion.

The bureaucratic demand, Shohet said, is to ask how much student literacy has increased as a result of a program; her response consisted of showing how faculty have responded and how classroom activities have changed. She also commented that the outcomes of a literacy across the curriculum program are not limited to reading and writing instruction. At Dawson, faculty members began attending more faculty development seminars, interacting across departments, and volunteering to develop classroom projects (when previously, they had been embarrassed to be seen at a writing workshop). Faculty publications also increased.

The design of an evaluation instrument began with a model developed at San Diego State College, which helped define an evaluation that would be particular to Dawson. The college also employed an outside consultant, Shohet noted, which gave the evaluation additional objectivity. She cautioned against using generic evaluation instruments; each program develops with its own goals and philosophy, and must be evaluated on that basis.

The Dawson evaluation focused on particular questions about classroom activities before and after workshop attendance: class time spent writing; time spent talking about writing; writing assignments; use of journals; working with drafts; oral communication assignments; use of library resources. In categories covering writing, reading, speaking, and listening skills, the evaluation attempted to determine what changes instructors had made in their classes, what goals they had for center programs, and whether participation in center programs had promoted educational exchange with other faculty members or increased levels of theoretical knowledge. One section defines the program's objectives and asks faculty to evaluate objectives as appropriate and applicable; this type of inquiry not only helps refine program goals for future planning, but reinforces awareness of program objectives among faculty members who respond, Shohet said.

The center ran a pilot study, distributing evaluation forms to 150 randomly selected faculty members, chosen from those who had attended workshops. About 100 responses were returned; some questions were refined. The evaluation in its final form will be distributed to faculty members across the curriculum this spring. Shohet will report on the results at next year's NTNW Conference, which will be held at Dawson College. (Shohet is the Conference Co-Coordinator.)

VALIDITY ISSUES IN DIRECT WRITING ASSESSMENT

Speakers: *Karen Greenberg*, NTNW and Hunter College, CUNY
Stephen Witte, Stanford University

Introducer/ Recorder: *Joanne Van Oorsouw*, College of St. Catherine, Minnesota

Karen Greenberg began with what she deemed a radical statement: "I have examined more than 600 writing tests and have yet to see one that I would consider to be a valid one." She went on to state that it seems impossible for writing tests, with their narrow subjects, implausible audiences and severely restricted time frames, to reflect the natural processes of writing in either academic or personal contexts.

Greenberg explained her position by pointing out that writing consists of the ability to discover what one wishes to say and to convey one's message through language, content, syntax and usage that are appropriate for one's audience and purpose. In light of this, she said, it is particularly distressing to note that teachers at many institutions find themselves administering tests that bear little resemblance to this definition or to their curricula and pedagogy. For example, many schools still use multiple-choice tests of writing even though this type of testing does not elicit the cognitive and linguistic skills involved in writing.

She stated that writing sample tests, on the other hand, can assess writing capacities that cannot be measured by existing multiple-choice tests. They, however, also, have flaws, and many problems result from our reliance on single-sample writing tests for placement and proficiency decisions. She warned that a single writing sample can never reflect a student's ability to write on another occasion or in a different mode. Yet, according to surveys conducted by NTNW and CCCC, thousands of schools across the country continue to assume that "writing ability" is stable across different writing tasks and contexts and continue to use a single piece of writing as their sole assessment instrument.

Greenberg then went on to suggest what those involved in large-scale scale direct assessment of writing should do about validity. The first step in establishing a test's validity is to determine its purpose: what information is needed by which people and for what purposes? The next step is developing a clear definition of the writing competence that is being assessed, one that will vary according to the purpose and context of the assessment. Developing this definition is a critical step in creating a valid assessment, but it is easier said than done for there is as yet no adequate model of the various factors that contribute to effective writing in different contexts. Finally, after coming to agreement on their definition of writing competence, faculty need to establish consensus about the writing tasks that are significant in particular functional contexts.

Greenberg noted that she deliberately chose to talk about faculty rather than test developers, for she believes that the people who teach writing should be the ones who develop the assessment instruments. Faculty need to work together to develop tests, to shape an exam they believe in so that they can be sure its principles infuse curriculum and classroom practice. Even when faculty work together, however, Greenberg said that definitions of competent writing may vary dramatically. Locally-developed essay tests show incredible variability in the skills measured, due to difference in the range of skills assessed and the criteria used to judge those skills. For example, faculty often differ about the range of discourse structures that they should teach and that a test should assess. One way to sample students' ability to write different types of discourse is to use the portfolio method, in which writers select three or four different types of drafts and revisions for evaluation. This kind of assessment reflects a pedagogy that emphasizes process over short, unrevised products. Thus, this kind of test stimulates writing teachers and programs to pay more attention to the craft of composing.

Greenberg's final point was phrased as a question: What is the relationship between what we teach and what we test? We cannot, and should not, separate testing from teaching, and we as a profession must be more concerned with the validity of both of these efforts.

Steve Witte summarized a study begun in 1982 which sought to answer two research questions: (1) Do writing prompts that elicit different types of writing and that elicit written texts of the same quality cause writers to orchestrate composing in different ways? and (2) Do comparable prompts that elicit the same type of writing and elicit written texts of the same quality cause writers to orchestrate composing in different ways? Witte stated that although this study did not investigate naturally occurring discourse, this type of experimental study can inform the kinds of conceptualizations we can make beyond the experimental study.

The first step in conducting this research was to create two comparable writing tasks of two types:

expository and persuasive. Prompts were created after consultation with students, writing teachers, high school teachers, and pre-service high school teachers. Those prompts found to be comparable by these groups were then pretested and were found to elicit comparable ranges of writing quality. The subjects were 40 volunteer college freshmen at the University of Texas who were randomly assigned to one of the four tasks. Think-aloud protocols and rough drafts were collected and analyzed according to a coding scheme developed by the experimenters. The results of a multivariate ANOVA showed that 16 variables distinguished between the persuasive and expository tasks; these variables included generating ideas, setting content goals, reviewing text. Writers tended to set more content goals and generate more ideas for the expository tasks and set more rhetorical goals for the persuasive tasks. A discriminant analysis was done to determine which variables distinguished among all four tasks. Eleven variables were found to do this.

Witte stated that findings indicate that writers engage in different kinds of processes for different kinds of tasks. In terms of writing assessment, each prompt we use to assess ability will be measuring different dimensions of that ability. The obvious conclusion, then, is that there is no way to assess writing ability with only one task or prompt. We do not yet know how many prompts or tasks might be needed. Witte also noted that this study should make us question models or the writing process that are based on protocols from just one task. More research of the type presented here--studies that examine the effects of context on process--are needed. In Witte's study, context was limited to the writing prompt, a part of the context important to writing assessment. He said that we need more research that will help us identify how writing processes are circumscribed by other aspects of context.

RELIABILITY REVISITED: HOW MEANINGFUL ARE ESSAY SCORES?

Speaker:	*Edward White*, California State University, San Bernardino
Introducer/ Recorder:	*Karen Greenberg*, NTNW and CUNY

Ed White began the session by offering a clear definition of reliability: it is the consistency of measurement over different test situations and contexts. He explained the various types of reliability and discussed their origins in agricultural research. He briefly discussed validity in educational research and noted that reliability is "the upper limit for validity" (i.e., no test can be any more valid than it is reliable).

Next, White discussed "true scores," the "standard error of measurement," and uncertainty in measurement. The true score of a test is a Platonic ideal--it is the mean score of repeated attempts at the test under identical conditions. Since we can never determine a student's true score on a test, we need to calculate the test's standard error of measurement (a statistical estimation of the standard deviation that would be obtained for a series of measurements of the same student on the same test). White pointed out that because of the error in all measurement, no single score is reliable enough to be used as the sole determinant of any particular ability or skill.

Next, White explained the problems in essay test reliability. He compared the reliabilities of holistic scoring, analytic scoring, and multiple-choice scoring; and he discussed the difference between inter-rater reliability (agreement between different raters) and intra-rater reliability (agreement of a rater with him/herself at different points in time). White commented that rater disagreements over the quality of holistically-scored essays do not constitute "errors." The traditional psychometric paradigm of reliability cannot help us with a phenomenon such as subjective judgment, which may be better determined through rater <u>disagreements</u> rather than through their agreements. This led White to a discussion of "generalizability theory" and its implications for the reliability of essay test scores. He noted that our goal should be a reduction in the number of rater disagreements of more than two scale points (these should occur no more than 5% of the time in any scoring session).

White ended with suggestions for increasing the reliability of essay testing. Essay test administrators should reduce the sources of variability in test contexts (by controlling as many variables as possible), should keep the scoring criteria constant, should pre-test and control test prompts, should control essay reading and scoring procedures, and should always try to use multiple measures to assess students' skills.

ESTABLISHING AND MAINTAINING SCORE SCALE STABILITY AND READING RELIABILITY

Speakers:	*Wayne Patience*, GED Testing Service *Joan Auchter*, GED Testing Service
Introducer/ Recorder:	*Anne Aronson*, University of Minnesota

Wayne Patience and Joan Auchter presented the procedures used by the General Education Development Testing Service (GEDTS) to evaluate essay exams required as part of the GED Test for individuals seeking high school equivalency diplomas. They described and illustrated the methods employed by GEDTS to establish and maintain stability or consistency of scoring, and reliability among readers, despite the decentralized nature of their evaluation program.

Patience explained that the notion of equivalency derives from: (1) defining the content of the GED Tests so as to reflect the community expected outcomes of completing a traditional high school program of study and

(2) defining passing scores relative to the actual demonstrated performance of contemporary graduating seniors. Only those examinees who receive scores that are better than 30% of high school seniors are awarded the diploma. The job of the GED staff is rather to describe the skills and content knowledge that characterize the work of high school seniors than to prescribe levels of achievement.

The recent addition of an essay exam to the Writing Skills Test created questions about how to establish and maintain reliability. The GEDTS's first activity was to develop a scoring scale that would have the same criteria regardless of time or place. By administering an essay exam to thousands of high school seniors, sorting those essays into six stacks, and describing the characteristics of each stack, the Writing Committee of GEDTS was able to develop a holistic scoring scale that has been used successfully in hundreds of sites nationwide.

Auchter then reported on how GEDTS insures stability and reliability in the use of the scoring guide. A permanent GEDTS Writing Committee, consisting of practicing language arts professionals, selects the topics and the papers that are used in training, certifying, and monitoring site trainers and readers. The Writing Committee chooses and tests "expository" topics that do not require students to have any special knowledge or experience. The next step is for GEDTS to train and certify Chief Readers who are responsible for insuring that the GED scoring standards are applied uniformly. During the 2 1/2 day training, Chief Readers learn to overcome personal biases (e.g., responses to handwriting) that may influence scoring, and to use the language of the scoring guide alone to describe and evaluate papers. Sets of training papers contain a range of papers for each point, to illustrate the fact that there is no "perfect" paper for each point, but that there is typically a distribution of high, medium, and low papers. Training packets also include problematic papers (e.g., a paper written in the form of a rap song). Since the national average for high school essays scores is 3.25 and for GED scores is 2.7, training sets contain a disproportional number of 2, 3, and 4 papers. After working with training papers, GEDTS trainees are required to evaluate several sets of papers to determine whether or not they are currently certifiable as Chief Readers.

The same training and certifying procedure is carried out at the various decentralized testing sites, with the Chief Readers responsible for training and certifying readers. Auchter noted that language arts teachers trained through this process feel better about teaching writing and about using holistic grading in the classroom.

Further steps to insure score scale stability and reliability are site certification and monitoring. Each scoring site must demonstrate the ability to score essays in accord with the standards defined by the GED Testing Service. Essays used for site certification must receive at least 80 percent agreement in scoring among Writing Committee members. Although some sites may achieve high inter-reader reliability, a site cannot pass certification unless it achieves at least 90 percent agreement with GEDTS essay scores (or 85 percent for a provisional pass). Three procedures are used to monitor testing sites: (1) the Chief Reader does third readings of discrepant scores and records each time a reader is off the standard; (2) readers evaluate a set of "recalibration" papers at the beginning of each scoring session in order establish reliability for that day; and (3) GEDTS conducts site monitoring using the same procedures as are used in site certification.

TRAINING OF ESSAY READERS: A PROCESS FOR FACULTY AND CURRICULUM DEVELOPMENT

Speaker:	*Robert Christopher*, Ramapo College, New Jersey
Introducer/ Recorder:	*Mary Ellen Ashcroft*, University of Minnesota

Robert Christopher emphasized the imperative of assessment, pointing out that assessment has always been intrinsic to the classroom experience, but it has now become extrinsic. He noted that many faculty fear writing assessment efforts because they represent an intrusion on their methods for evaluating students. He stated that faculty fears can be countered by several arguments: assessment helps students, it facilitates faculty and institutional research, and it is a professional activity.

Christopher went on to suggest ways of building faculty consensus for assessment. In a training readers, he suggested starting with a loyal, supportive core. This group's primary responsibility would be the development of an instrument for assessment, a task which should take six months to a year. He suggested that good readers are people who are task oriented, are good collaborators, are preferably not new faculty members (who might not have a sense of writing at the institution), and who work with "all deliberate speed." Good readers must not be "Matthew Arnolds" before whose standard everything fails. It is important, according to Christopher, that a large pool of readers be developed, so that a small loyal group will not wear out.

The next step, according to Christopher, is conducting a reading to build consensus. The initial reading should consist of 500 to 1000 essays, so that readers get a sense of the range of writing abilities of students at the institution. The reading must be conducted "blind" with each paper read and assessed twice. Essays are identified as "strong," "weak," and "in-between"; readers discuss each essay and slowly evolve into an interpretive community.

In terms of curricular implications, Christopher pointed out that placement assessment is easier to accomplish and has been more fully studied than proficiency assessment. He also noted that assessment can be used for students to learn to talk about their writing in small groups and in conferences, so that students learn to to better readers and editors. Assessment can also be used to encourage collaborative or group teaching, said Christopher. As faculty members relinquish some control to group or collaborative situations in the assessment process, they learn from one another and share techniques and materials.

In answer to conferees questions about developing the holistic process, Christopher suggested that the English Department provide a core of expert readers which should eventually grow to become interdisciplinary. He noted that in any two-day reading of essays, there is always the need for reliability checking ("Let's all read this essay and make sure we're on track"). Finally Christopher pointed out that students benefit from holistic essay assessment because their writing skills are evaluated by a team of teachers. This kind of assessment program, Christopher says, works on behalf of students.

DISCREPANCIES IN HOLISTIC EVALUATION

Speakers:	*Donald Daiker*, Miami University, Ohio *Nedra Grogan*, Miami University, Ohio
Introducer/ Recorder:	*Sandra Flake*, University of Minnesota

Donald Daiker presented the goals of the sessions: to share the conclusions and a tentative evaluation of his and Nedra Grogan's examination of discrepancies in holistic evaluation. Noting that discrepancies in holistic evaluation have been a problem from the beginning, he raised two questions: What accounts for discrepancies in holistic evaluation if the "quirky" reader is ruled out? And is there such a thing as a discrepant essay?

Daiker and Grogan sought to answer these questions using an annual holistic grading session for Miami University's Early English Composition Assessment Program (EECAP), a program in which 10,000 essays written by high school juniors in a controlled setting are evaluated for diagnostic purposes. The setting was one in which students, using a prompt, wrote for 35 minutes in a high school composition class. The time limitation was dictated by the constraints of a single class period. The goal of the holistic evaluation was essentially diagnostic, with a scoring scale of 1 to 6. Grades of 5 or 6 indicated clearly above average papers demonstrating strengths in all of the rating criteria. Grades of 3 or 4 indicated papers ranging from slightly below to slightly above average, with combined strengths and weaknesses in the criteria or under development. And grades 1 or 2 indicated clearly below average papers failing to demonstrate competence in several of the criteria, often because the paper was too short. A grade of 0 was used only for papers which were off the topic of the prompt. Evaluators gave each paper a single holistic rating, and additionally rated criteria in four categories (ideas, supporting details, unity and organization, and style).

The participating high school teachers (who were the evaluators) were trained through a process of rating and discussing sample papers, so that the rating criteria would be internalized. Participants in the session were then provided with the writing assignment or prompt, the scoring scale, the rating criteria, a rater questionnaire, and one of the papers.

To locate possible discrepant papers, Daiker looked for three-point gaps in scoring by two evaluators and gave such papers to both a third and fourth evaluator. If those evaluators also disagreed on the rating of the paper, he identified it as a potentially discrepant paper. Through this process, four potentially discrepant papers were identified, and those four papers were given to all 61 of the evaluators in a session at the end of the second weekend of evaluation. Participants in our session then read and evaluated one of the potentially discrepant papers, using a rater questionnaire, scoring scale, and rating criteria. The rating of the participants were tabulated: 1 person assigned the the paper a 6, 16 assigned a 5, 28 assigned a 4, and 4 assigned a 3.

Following the participant evaluation and some discussion, Grogan presented the result of the evaluation by 61 trained raters who rated the paper at the end of the second weekend of evaluation, with 26 of the raters (42.6%) giving an upper range (5-6) rating, 34 of the raters (55.8%) giving a middle range (3-4) rating, and 1 (1.6%), giving a lower range (1-2) rating.

Because of the clear division between the 5-6 and the 3-4 rating, Grogan and Daiker believe that the paper did qualify as a discrepant paper. Daiker reported that discussion following the rating by the trained evaluators suggested a correlation between the depth of emotional response to the paper and the highness of the score. Following some discussion about whether or not the paper was truly discrepant, a conferee asked whether the problem was really caused by discrepant readers who could not be objective because of the depth of their emotional response. Daiker argued that reader objectivity was more complicated issue and further argued that precisely because the paper provokes a range of responses to the emotional content, it could be defined as a discrepant paper.

The implications of evaluating discrepant paper were then summarized by Grogan, who raised the issue of the role of holistic evaluation of a single essay that receives discrepant scores. She concluded that in such cases a single essay should not determine the fate of the writer, and that an appeals process clearly needs to be a significant part of a holistic evaluation program. Discussion throughout the session focused on some of the limitations of holistic evaluation of writing produced under a time constraint, on problems in establishing clear criteria and scales, and on problems of reader objectivity.

PROBLEMS AND SOLUTIONS IN USING OPEN-ENDED PRIMARY TRAIT SCORING

Speaker:	*Michael C. Flanigan*, University of Oklahoma
Introducer/ Recorder:	*Chris Anson*, University of Minnesota

Michael Flanigan began by outlining his university's plan for five years of experimental research on the teaching and testing of writing. Much of this research will replicate published research studies, but original research will also be conducted. All of the studies will be controlled experimental studies so that the researchers can be fairly faithful to the original ones and can analyze any differences between the original and new research.

Flanigan discussed one study, already completed, in which his colleague David Mair and he combined the strategies of two studies by George Hillocks, an experimental study involving teaching extended definition using inquiry and models and a descriptive study dealing with "modes of instruction" (both of which Hillocks discusses in some detail in this book Research on Written Composition). In the replicated study at Oklahoma, all twenty classes consisted of university freshmen; for nine of the ten teachers it was their second semester of teaching, and approximately 500 students were involved.

Flanigan pointed out he chose Hillocks' studies because both dealt with significant areas in teaching and writing. Extended definition represents a kind of discourse that permeates almost all thinking and writing. The researchers believe that by replicating such an important study they could get inside the problems of the earlier research, and come to understand it better. The experimental extended definition study also used Hillock's open-ended primary trait scoring technique because the researchers wanted to learn to use and understand it better.

After reporting the findings from a small sampling of the data, Flanigan described some problems that he and his colleagues faced as they attempted to use Hillocks' open-ended primary trait scoring system and he discussed the modifications they made in it to obtain reliable results. He pointed out that with an open-ended primary trait scoring scale theoretically there is almost no limit to what students can score. Most scoring scales range from 1 to 6 (as in the holistic score for the ECT), 2 to 8 (as in CLEP), 1 to 5 ([illegible] CORE scoring) and so forth. In open-ended primary trait scoring, the limit for a talented student is probably dictated by time and the variation and limitations imposed by the writing called for. In the papers scored in this study, the top score was 28.

The traits for which students could receive scores were: (1) properly putting an item in a class; (2) creating criteria for the class; (3) giving examples; and (4) providing contrastive examples to clarify and limit each criterion. Points were not given for differentiae as in Hillocks' original study; instead, class and differentiae were combined (on the advice of Hillocks when the study was set up). Hillocks' scorers had had problems reaching agreement on this point. Students could receive 2 points for the class, 2 for each criterion, 2 for each example, and 2 for each contrastive example. Obviously the more criteria, examples and contrastive examples students could come up with, the higher their score. In initial training, scorers had problems staying close together in the higher ranges, so Flanigan modified his tolerance of acceptability by allowing scores in the range 1 to 10 to differ by 1 point, 11 to 20 to differ by 2 points, and 21 up to differ by 3 points. Scores within that range were averaged; scores that did not meet acceptable standards were read by a third reader. If the third reading fell within range of either of the other two readers, then those scores were averaged. If there still was no agreement, a fourth and fifth reader scored the paper, and the paper and the range of scores were given to the researchers and a score was determined. For example, one paper was scored 6 and 8; a third reader gave it 10; the fourth reader gave it 9, and the fifth reader gave it 7. Its final average was an 8. Only seven papers required the fourth and fifth reader. Often, readers had problems keeping clearly in mind the kinds of criteria the writers were developing. To simplify the process, any one clear criterion could be accompanied by a number of examples and contrastive examples. If no criterion was given, only one example could be counted. If an undeveloped example or string of general examples was given, a score of 1 was given.

Flanigan concluded that open-ended primary-trait scoring offers real promise, for it allows for a kind of differentiation that closed, limited systems do not. However, researchers who use the system will probably have to modify it to get consistent, reliable scores. They will also have to plan their research so that the traits they are describing and scoring are clear, well-defined, and fully conceptualized by their scores. The session ended with the speaker giving participants six papers that had been scored by three readers and leading participants through a guided scoring session.

THE IMPLICATIONS OF THE RHETORICAL DEMANDS OF COLLEGE WRITING FOR PLACEMENT

Speaker:	*Kathryn Fitzgerald*, University of Utah
Introducer/ Recorder:	*Linda Jorn*, University of Minnesota

Kathryn Fitzgerald gave participants attending this session a chance to analyze student writing in terms of rhetorical evaluation criteria developed at the University of Utah. These criteria are intended to do the following:
(1) describe the rhetorical situation college students face when asked to write an essay that will be assessed and used to place them in a freshman writing course; (2) assert that the evaluation of the rhetorical situation provides valid criteria for placement of students into various levels of freshmen writing courses; and (3) shape the discussion

THE SEVENTH ANNUAL CONFERENCE ON WRITING ASSESSMENT is a national conference designed to encourage the exchange of information about writing evaluation and assessment among elementary, secondary, and postsecondary administrators, teachers, and test developers through forums, panels, and workshops.

DISCUSSION TOPICS:

Research on Writing Assessment ⊠ Writing Assessment Across Cultures and Languages ⊠ Developing Essay Tasks ⊠ New Models of Scoring Essays ⊠ The Impact of Testing on Curricula and Pedagogy ⊠ Problems in Exit and Proficiency Testing ⊠ Computers and Writing Assessment ⊠ Writing Program Evaluation ⊠ National Standards for Writing ⊠ Portfolio Assessment

Note: **Springboards**, Quebec's annual language arts conference, will take place April 13th and 14th in Montreal. For information, call Fran Davis at (514) 484-7646. E.S.T

PRE- AND POST-CONFERENCE WORKSHOPS:

To provide in-depth exploration of selected issues, four workshops have been added to this year's program. Enrollment for each is limited, so register as soon as possible. The fee for each workshop is $40 US, $50 Canadian (which includes lunch, coffee breaks, and all materials).

Pre-Conference: Saturday, April 8
9:30 a.m.-4:00 p.m.

A. Computers in Writing
Leaders: Helen Schwartz & Michael Spitzer (30 participants)

B. Large-Scale Writing Assessment
Leader: Edward White
(50 participants)

Post-Conference: Wednesday, April 12
9:30 a.m.-4:00 p.m.

C. Portfolio Assessment
Leaders: John Dixon & Peter Elbow
(50 participants)

D. Writing Assessment Across Cultures
Leaders: Alan Purves and colleagues
(50 participants)

CONFERENCE SCHEDULE

Sunday, April 9, 1989

6:00 p.m. - 7:30 p.m.
Conference Opening

SPEAKERS:
Carolynn Reid-Wallace, Vice Chancellor, CUNY
Gerrard Kelly, Director General Dawson College
Rexford Brown, Education Commission of the States

7:30-9:30 p.m. Reception

Monday, April 10, 1989

9:00 a.m. - 10:30 a.m.
Opening Plenary Session

SPEAKERS:
Joseph S. Murphy, Chancellor, The City University of New York
John Dixon, Author of Growth Through English

10:30 a.m. - 5:30 p.m.
Concurrent Panels/Workshops

Tuesday, April 11, 1989

9:00 a.m. - 10:15 a.m.
Second Plenary Session

SPEAKER:
Janet White, Deputy Director, NFER, England

10:30 a.m. - 11:45 a.m.
Concurrent Panels/Workshops

12:00 p.m. - 2:00 p.m.
Luncheon and Closing Session

SPEAKER:
Bernard Shapiro, Deputy Minister of Education, Ontario

2:15 - 5:00 p.m.
Concurrent Panels and Workshops

REGISTRATION INFORMATION

Registration is limited to 700 people. Before March 9, the registration fee of $100 US or $125 Canadian per participant includes:

- opening night reception (April 9)
- 2 continental breakfasts (April 10,11)
- conference luncheon (April 11)
- coffee and tea between sessions
- all conference materials

After March 9, 1989, the registration fee is $120 US or $150 Canadian.

For further information, call Linda Shohet 514-931-8731 or Karen Greenberg 516-766-8099

HOTEL RESERVATIONS

Please write or call the Centre Sheraton Hotel directly before March 7, 1989 to receive our special conference rates:

Single:	$115 Canadian*
Double:	$125 Canadian*

*There is no tax on hotels in Montreal. The United States/Canadian exchange rate varies: the cost of the room in US dollars will be 20% - 25% less than the Canadian rates above.

Address: Le Centre Sheraton Hotel
1201 Dorchester Blvd. West
Montreal, Quebec
Canada H3B 2L7

Phone: (514) 878-2063

THE SEVENTH ANNUAL CONFERENCE ON WRITING ASSESSMENT

Registration Form

NAME____
(last name) *(first name)*

TITLE____

INSTITUTION____

MAILING ADDRESS____________________________________

HOME PHONE (____)______________ **WORK PHONE** (____)____________

Conference Registration Fee: Before March 9, 1989: $100 US; $125 Canadian
After March 9, 1989: $120 US; $150 Canadian

Workshop Fee: If you are registering for a pre- or post-conference workshop, please write a separate check for each workshop (so that we may return your check if the workshop has already reached its limit). Each workshop is $40 US, $50 Canadian. Please circle the letter of each workshop for which you are registering:

Sat. April 8: A__ B__ Wed. April 12: C__ D__

Please make all checks payable to: National Testing Network 1989

Please mail completed registration form and check(s) to:

Linda Shohet
3040 Sherbrooke Street W.
Montreal, Quebec, Canada H3Z 1A4

of the student writing by providing a holistic view of writing.

Before handing out samples of student writing, Fitzgerald discussed the theoretical background for developing the rhetorical criteria. She also reviewed some of the common problems of holistic scoring, emphasizing the fact that holistic scoring does not consider the different purposes of writing (for example, persuasive vs. self expressive writing). The rhetorical evaluation criteria developed at the University of Utah were designed to alleviate some of the problems encountered when using holistic scoring. The criteria help readers consider the purpose of students' writing and identify the internal and external purposes of the writing situation.

Fitzgerald pointed out that students' internal and external purposes complicate the writing situation for them. At the University of Utah, faculty feel that the purpose for students' writing needs to come from the students (i.e., internal), but in academia the purpose often comes from the instructor and is motivated by grades (external). The student has to think up his or her purpose for writing and must shape this purpose to serve the academic external purpose. Therefore, the student's purpose is always dual. These internal and external purposes are in essence the rhetorical situation and they need to be taken into account when faculty evaluate writing, particularly when this evaluation is used to place freshmen into English courses. Students' ability to handle this complex rhetorical situation informs instructors of the students' readiness for college writing.

Next Fitzgerald described how the rhetorical expectations of University of Utah professors were determined and used to develop the rhetorical evaluation criteria. These criteria consist of the following categories:

> **Category 1:** The writer's relationship to college readers and writers. Expectations: The most proficient writers recognize that any single piece of college writing is part of an ongoing written discussion about a topic and that they are expected to make a contribution to the discussion. They recognize that an authority (i.e, professor, test giver) identifies issues for discussion.
>
> **Category 2:** The writer's relationship with his or her subject matter. Expectations: College writers control their subject matter, pressing it into service to support their internal and external purposes.
>
> **Category 3:** The writer's relationship to the conventions of the genre. Expectations: College writers employ syntactical units appropriate to their thought, precise vocabulary, and the mechanics and spelling of standard written American English.

University of Utah students are given placement essay directions that explain the external rhetorical situation; and they have 45 minutes to plan, write, and revise their essays.

After reviewing the theoretical background and the criteria, participants used these criteria to evaluate and discuss some student writing. Fitzgerald pointed out that readers are told to pay attention to content and reasonability, that there are no hard and fast rules, and that judgment is a balancing act of various criteria and expectations of each institution. Readers at the University of Utah look at the quantity of student writing as relative to every piece of writing. In summary, Fitzgerald stated that these rhetorical evaluation criteria force readers to evaluate writing for its purpose, help readers define good college writing, and address the need to teach students about the effect that the rhetorical situation has on their writing.

USING VIDEO IN TRAINING NEW READERS OF ASSESSMENT ESSAYS

Speaker: *George Cooper*, University of Michigan
Introducer/
Recorder: *Terence Collins*, University of Minnesota

Large scale testing programs face a recurring problem of reader consistency and reliability. In this presentation, George Cooper demonstrated how the English Composition Board at the University of Michigan uses a video presentation of reader "standardization sessions" for self-monitoring within the reader cadre, for training new readers, and for disseminating information about the ECB's procedures to various campus constituencies. While Cooper presented alone, his remarks were prepared with Liz Hamps-Lyon.

In its placement readings, members of Michigan's ECB teams are guided by statements of criteria clustered under three headings: "structure of the whole essay," "smaller rhetorical and linguistic units," and "conventions of standard English surface features." Students write essays in response to prompts that define a situation and provide several choices of opening sentences. Two important characteristics of the 6000 student essays, then, are that topic choice is limited and orientation toward the topic is guided through provision of choices for essay openings. Further, the essays are rated for placement: recommendations fall into one of the following categories: exempt (7%), Introductory Composition (82%), and tutorial (11%). These recommendations reflect scores of 1, 2-3, and 4. While criteria for quality are outlined to readers, no specific calibration of trait content for the four point range is provided.

NOTES from the
NATIONAL TESTING NETWORK IN WRITING

The City University of New York
and the Fund for the Improvement of
Postsecondary Education

Directors, NTNW
Karen Greenberg
Harvey Wiener
Richard Donovan

Editors, *NOTES*
Karen Greenberg
Virginia Slaughter

Published by the Instructional Resource Center, CUNY Office of Academic Affairs
Harvey Wiener, Director

Scoring in this system depends on achieving what Cooper calls a "community of values" among readers. The video of reader standardization sessions grew out of one summer's experience in which this community of values has been lost as Cooper put it, "readers were using an unimaginable range of criteria by which to evaluate essays" and "had become entrenched in their own perspectives." The original motive for the video was self-examination. Through videotaping daily standardization sessions in which papers receiving "split" scores were the focus of discussion, Cooper's team of readers sought to capture the articulation of values giving rise to the discrepancies and to record the process of moving to agreement on application of criteria. This led the team to analyze and communicate important characteristics of their standardization sessions and our assessment as a whole. Also, this procedure modeled a process of "give-and-take" that was helpful in training new readers and in explaining the placement process to various departments.

From ten hours of session tapes, the team assembled thirty five minutes of actual exchanges interspersed with explanation and highlighting. The standardization discussion presented in the tape enacts what Cooper calls "positive sharing": talk marked by the various readers' attempts to recognize the qualities in an essay that lead to divergent scoring, each reader's comments leading to further discussion and finally to agreement. Such discussion (whether on the tape or in person at the start of a reading session) reminds participants of the criteria governing scoring. It serves the further purpose of helping group members realize the vitality of the act of reading, placing an apparently perfunctory reading act (in the context of reader-response theory) into the full context of extra-textual factors that shape readings in open view. The importance of reflecting on the evaluator as reader--co-creator of a text--rests in the capacity of texts to sway a reader-evaluator when they embody positions to which the reader might be favorably inclined or which the reader might find repugnant.

Cooper asserted that the taped standardization sessions play the key role of "forming individual consciousness into a community consciousness." The video record of this work in progress puts flesh on the abstraction and models the process for beginners in order to cultivate a community of readers who will evaluate not only the student essays, but who will also study their own responses, keeping in mind the relationship of their responses to the criteria.

WPA PRESENTATION ON EVALUATING WRITING PROGRAMS

Speakers: *Robert Christopher*, Ramapo College, New Jersey
Donald Daiker, Miami University, Ohio
Edward White, California State University, San Bernardino

Introducer/ Recorder: *John Schwiebert*, University of Minnesota

This session was organized by the National Council of Writing Program Administrators (WPA), and the panelists wished to share their experiences as writing program evaluators and to address salient issues of writing assessment as they pertain to writing program evaluation.

Upon request, consultant-evaluators from the National Council of Writing Program Administrators will conduct a writing program assessment for a college or university. To prepare both themselves and the WPA evaluators (usually a team of two) for the assessment, schools are asked to complete a narrative "self-study" of their writing program at least one month before the WPA team visits. Robert Christopher distributed copies of the self-study guidelines, which can be obtained from the address given at the end of this abstract. The purpose of the assessment is to help faculty and administrators develop more effective writing programs appropriate to their institutions' needs. Donald Daiker and Edward White described occasions when the WPA service assisted writing faculty on a campus to enlist high-level administrative support for innovative reforms in their writing programs.

Most of the session focused on the topic of testing, which, it was emphasized, is only one dimension of an overall program assessment. To be effective, institution-wide programs of assessment should be appropriate to the particular needs, demographics, and aims of the individual school. The challenge of deciding what is appropriate underscores the relevance and value both of the WPA assessment and of the self-study a school does before the WPA visit. Panel members discussed some of the key issues involved in each of the following kinds of testing: admissions, placement, equivalency, and course exit. Rising junior and value-added tests were also mentioned but could not be discussed in detail in the time allotted. Key points about each type of test are below:

Admissions Tests: Discussing the purposes of the SAT verbal exam, White stressed that the SAT assesses verbal aptitude and not writing ability. As such, it is useful as a criterion for admissions but should not be a basis for exempting students from freshman composition.

Placement Tests: Before actually developing a placement test, a school should decide if it needs one. Many institutions do need such exams to assure that individual students receive writing instruction appropriate to their abilities and experience. After a need has been determined, a school should develop a test based upon its own curriculum--specifically, upon what is taught in freshman composition. Some schools borrow or adopt tests that fail to mesh with their own institutional needs. Only by examining its curriculum can an institution rationally decide what it is testing for.

Equivalency Tests: These tests provide a special service to students, and they differ fundamentally from placement exams. The basic message of an equivalency is: "Show us that you (i.e., the student) are in control of what we do in freshman comp and we'll let you out of it." As such, equivalency tests must be based firmly on the school's curriculum. Given its special purpose, the testing instrument must also be more complex than one used for placement.

Course Exit Tests: The course exit exam is a common test that all students must pass in order to complete a course (freshman composition or other). Noting that such tests can discriminate against students who write well but who are poor drafters or test takers, White urged against tests being the only basis for exit. A good exit exam covers materials and processes which students have addressed in their class. White observed that the greatest potential benefit of an exit test derives less from the test itself than from the incentive it can provide for departmental and interdepartmental faculty discussions of writing and curriculum.

Institutions desiring more information on the WPA consultant-evaluator service should write to Professor Tori Haring-Smith at the following address:

Rose Writing Fellows Program, Box 1962,
Brown University, Providence, RI 02912.

DEVELOPING AND EVALUATING A WRITING ASSESSMENT PROGRAM

Speakers: *Lorenz Boehm*, Oakton Community College. Illinois,
Mary Ann McKeever Oakton Community College, Illinois

Introducer/
Recorder: *Marion Larson*, Bethel College, Minnesota

Lorenz Boehm and Mary Ann McKeever addressed issues of designing, implementing, and evaluating an essay test currently being used by three Chicago-area community colleges. This test is designed both to place students in appropriate composition courses and to determine if students in developmental or ESL composition courses are prepared to move on to Freshman Composition.

Although the test has been used since 1984, preparations for its implementation began in 1982, and evaluation and refinement of test questions and procedures is ongoing. This test replaced an objective test of grammar and usage that was being used at the time. During the planning process, prompts were developed and pilot-tested, evaluation criteria were discussed, and reader training methods were developed. In addition, those developing the test sought to gain campus-wide support and involvement from faculty, staff, and administration.

In the test, students are given two argumentative topics from which to choose. With each topic, they are give a context for writing and an audience for whom they are told to formulate an essay arguing their position. They are given 50 minutes to plan and write their essay. Efforts are made to be fair to ESL students: topics are as "culture free" as possible, prompts are worded simply, ESL (and Learning Disabled) students are given an additional 20 minutes to write their essay, and specially-trained readers evaluate ESL and LD responses.

Each essay is holistically scored on a 6-point scale by three readers, two of whom must agree in their assessment. In cases of disagreement, an additional reader may be used, and an appeals procedure is available to students. These readers come from across the college and all of them participate in frequent, extensive training to be sure that the understand and agree upon criteria for the essays they will be asked to evaluate. In training, as well as before actual evaluation sessions, agreement among readers is reached by examining, rating, and then discussing sample essays; discussing criteria for scoring; and then rating more sample essays.

Many benefits have come from Oakton's use of this writing placement test. Primary among them is the greatly increased dialogue among faculty, administrators, staff, local high schools, parents, and students about writing. Such cooperation is essential to the test's success, because it has helped short-circuit potential disagreement and has made members of the college community more receptive to what the composition faculty are trying to accomplish. It has also greatly fueled writing across the curriculum efforts on campus.

This test is continually being evaluated by Boehm, McKeever, and their colleagues to ensure that it is placing students appropriately, that the different prompts are eliciting responses of comparable quality, and that agreement among readers is high. The results thus far are quite positive: composition teachers are very satisfied that students are being placed in the courses they need. Pilot testing prompts in composition classes and then carefully monitoring the ratings given to essays written in response to these prompts has helped ensure that different versions of the test are comparable; and evaluation criteria are kept consistent by frequent, ongoing training of essay raters.

THE CHANGING TASK: TRACKING GROWTH OVER TIME

Speaker:	*Catharine Lucas*, San Francisco State University
Introducer/ Recorder:	*Hildy Miller*, University of Minnesota

Catharine Lucas explained that traditional writing assessment is designed to determine whether student writing improves on a given specified task, whereas what we need is a new kind of assessment that focuses on how students change the task as they grow as writers. She noted that we know that as writers develop, they formulate new structures to represent tasks, and that they may be awkward in their initial attempts at working with new structures. For example, writers may experiment with complex argumentative structures, abandoning the simpler narrative structures at which they may be more skilled. Ideally, writing assessment should recognize and reward their attempts at more sophisticated formulations, even when performance falls short, rather than constraining the writing task in a way that only measures their ability at what Moffett calls "crafting to given forms."

To debunk the myth that writing is a unitary measurable construct and to show instead the impact of a student's maturing task representation, she provided samples of one student's writing that were submitted in response to a longitudinal portfolio assessment of his writing abilities from ninth to twelfth grade. During each of the four years, the student was asked to produce an essay as part of a school-wide assessment program. Four readers then rank-ordered the four papers to determine the writer's best and weakest work. While we would assume that his ninth grade essay would be weakest and the twelfth grade version the best, instead a different pattern emerged: raters consistently rated the twelfth grade effort the worst.

The reason for this surprising result was found through closer inspection of the writer's choices in task representation. In the three papers he submitted in grades 9, 10, and 11, the writer used the narrative form, a structure that develops comparatively early, since 6th graders are typically sophisticated story tellers. These essays were successful, in part, because he was using a familiar form. However, in the 12th grade essay he chose to represent the task with an argumentative form, usually a later developing skill, and one in which he was as yet inexperienced.

Thus, Lucas concluded, we need a way to take a writer's growth into account in assessment. Writers experimenting with new structures face a harder task, one which is likely to cause the writer initially to produce new errors. Evaluators of writing, like judges of figure skating, diving and other "performance sports," need to develop systematic ways of taking into account the difficulty level of what the performer is attempting. In order to account for changes in what is attempted we need to study how writers develop both across and within discourse domains. This will require a common language for identifying domains and a way of charting what carries over and what changes when writers move form one to another. All discourse theorists polarize fictional and non-fictional writing, or as Britton terms it, poetic and transactional writing. As a result, we tend to assume that the two are mutually exclusive: fiction writers rarely include essays in fiction and in academia we rarely allow poetic expression. In addition to these polar ends of the discourse continuum, Lucas posits a middle category, which draws freely on both fictional and academic styles, and includes autobiography, belles letters, the New Journalism, and the personal reflection essay widely used in classrooms and school assessments. While it is relatively easy to chart a writer's development within either the literary or the discursive domains, growth in this middle domain is sometimes marked by shifts from fictional techniques to extended abstract discourse, as in the case presented. Whether students are moving within the mixed domain, or from the literary end of the spectrum to the discursive end, even when teachers recognize the second piece as representing a later effort, they recognize that the text is often less successful in what it attempts than the earlier piece. This difference diminishes, of course, as the student gains skill in handling discursive, transactional writing.

To make possible more careful comparisons of what changes as students move within and across domains, Lucas has developed a method of defining tasks that draws on work done by Freedman and Pringle ("Why Children Can't Write Arguments") based on Vygotsky's (Thought and Language) distinctions between focal, associational and hierarchical arrangements, as well as on Coe's (Toward a Grammar of Passages) method of charting relations between propositions in a text. Lucas's system distinguishes between four text patterns: (1) the chronological core in which the student tells a story, providing commentary at end--a sign the writer is moving toward abstraction; (2) the focal core in which the title provides the subject of focus, with each sentence relating to it--a sign that some notion of related ideas is emerging; (3) the associational core in which we see chains of associations forming, often with a closing commentary; and (4) the hierarchical core, in which long-distance logical ties supplement short-range connections between complexly interrelated ideas, in a pattern typical of advanced exposition in Western cultures. Using this system, we may begin to see how writers build new schema within these different domains, and begin to reward them for these promising signs of growth in our assessments of their writing abilities.

ASSESSING WRITING TO TEACH WRITING

Speaker: *Vicki Spandel*, Northwest Regional Education Laboratory
Introducer/ Recorder: *Alice Moorhead*, Hamline University

Rarely are the lessons learned from large-scale writing assessment translated into terms that make them relevant for and useful to the classroom teacher. Yet many of those lessons show how teachers can use systematic writing assessment--especially when teaching writing as a process. Large scale, district-wide writing assessment is a costly process (at least 2.5 days for training/assessing and between $2-$8 a writing sample); however, as part of professional development programs, most districts could justify the necessary time and budget.

In this presentation, Vicki Spandel discussed her efforts, along with those of Richard Stiggins', to link writing assessment and instruction through their work in the Portland area for Northwest Regional Education Laboratory. Spandel's current assessment method focuses on using an analytic rating guide. She argues that although it is difficult to separate form from content in assessment, one can assess the features of writing, thus her interest in an analytic guide that can be used holistically to assess and to teach writing. Since teachers are often afraid of assessment, using the rating guide can ensure that what teachers value gets assessed and then gets translated into practice.

As an assessment tool, Spandel's analytic rating guide was generated from writing samples rather than developed as a guide to impose upon writing. The guide captures a more complete profile of the writing samples when used along with holistic assessment. It distinguishes six features of writing: ideas and content; organization; voice, word choice; sentence structure; writing conventions. Each feature is described and ranked by degrees for a score of 5 or 3 or 1. Not only does this analytic rating guide objectify expectations for writing but it also offers a more defensible version of the subjective process of writing assessment.

Using this guide with the holistic assessment process, particularly as in-service workshop for professional development, has two key advantages:

(1) The assessment process promotes "real" agreement among teachers and professional raters about strengths and weaknesses in writing.

(2) Teachers can re-enter the classroom to teach writing more explicitly on what "counts" in writing and know this instruction is in concert with and reinforced by others.

Not only can teachers use the analytic guide but so can students. In peer review groups, students can focus their writing efforts more directly with the six feature guide as "revision stations" for students to visit for specific feedback on their writing. In Spandel's experience, teachers welcome the use of this analytic guide for assessment and for teaching writing. Many teachers claim: "I'll never teach or think or writing in quite the same way."

READER-RESPONSE CRITICISM AS A MODEL FOR HOLISTIC EVALUATION

Speaker: *Karl Schnapp*, Miami University
Introducer/ Recorder: *Ann Hill Duin*, University of Minnesota

Karl Schnapp's session focused on the application of reader-response theory to large and small scale holistic assessment. Schnapp began by citing the work of Stanley Fish, David Bleich, and Norman Holland as working models for the holistic evaluation of student writing. He then said that his own work is also based on Edward White's theories of composition as a socializing and individualizing discipline. From these theorists, Schnapp concluded that the best composition pedagogy views students' writing from both social and individual perspectives. In short, the interpretation and evaluation of writing depends on qualities of the community in which the writing was created and was evaluated.

Schnapp then described his specific project. His model is based on three reading theories that lead to a model for the holistic evaluation of writing. The first theory is the "top-down" model of reading as discussed by Holland and Bleich, the second is the "text-reader interaction" theory (from information-processing theory) as discussed by Rosenblatt, and the third is the "communal association" theory as discussed by Fish. Schnapp described his model in detail. Then he asked conferees to fill out a survey identical to that used in his study. The survey asked us to complete questions regarding our perceptions and understanding of composition/language arts. Next we read an essay written by a freshman student and rated the student essay. Finally, we completed a second survey in which we gave information on the criteria we employ when holistically evaluating student writing. As with Schnapp's results, we had about 75% agreement in terms of the common goals of the composition instructors present. Schnapp stated that his research shows that writing teachers see writing as helping students on more of a practical level than on an aesthetic level.

The remainder of the presentation was a discussion between Schnapp and the conferees. Key points that emerged included: the need to ask readers about what influences them as they evaluate papers; the need to determine the evaluative standards for one's discourse community; and the extent to which readers are influenced by what they are thinking about while evaluating writing.

THE DISCOURSE OF SELF- ASSESSMENT: ANALYZING METAPHORICAL STORIES

Speakers: *Barbara Tomlinson*, University of California, San Diego
Peter Mortensen, University of California, San Diego
Introducer/ Recorder: *Anne O'Meara*, University of Minnesota

Barbara Tomlinson and Peter Mortensen gave conferees attending this session an opportunity to become students of their own writing processes. Much of the session was devoted to composing, sharing, and analyzing our own metaphorical stories about how we write. Tomlinson and Mortensen feel that using metaphorical stories in the classroom provides a means for students to take responsibility for their own writing, to balance personal with external assessment, and to center attention on the writing process rather than the product.

Tomlinson began by sharing some of her own metaphors for writing as well as some of those she found in her study of over 2000 professional writers. Handouts gave further examples from both professional and student writers. The metaphors were sometimes relevant to for the process of writing as a whole and sometimes symbols focusing on one aspect of writing. They ranged from clear analogies (e.g. building, giving birth, cooking, mining, gardening, hunting, getting the last bit of toothpaste) to metaphors that needed elaboration like a "gusset" (a small, irregular piece of material necessary for the construction of a garment, but hidden) and the "lost wax process" (a way of making a mold which then melts away when the product is finished). Tomlinson stressed that metaphors can reassure and guide her through composing problems as well as help her describe these problems.

The speakers then simulated their technique for using metaphorical stories in the classroom. As the participants began to compose their own metaphorical stories, Peter Mortensen asked some guiding questions to get us started, encouraging us to think of metaphors we might use for beginning writing, finishing writing, writing under pressure, writing badly, writing well, generating ideas, and so on. He suggested students could also use the guiding questions (distributed on the handout) in interviews or in collaboration to get started.

In the discussion that followed, Tomlinson and Mortensen stressed that metaphors should be accepted and explored, rather than judged. They may be original, adopted, or enforced; they may be idiosyncratic, contradictory, or even strike us as "bad." The important thing is that we and our students look at what the effects of writing metaphors are, what they imply about writing, and how they match or might amplify our experience. When they have students compare their metaphors to those of professional writers, Tomlinson and Mortensen minimize possible intimidation by emphasizing that the purpose is to find similarities and common problems.

Finally, the speakers summarized their reasons for using metaphorical stories in the classroom. In addition to taking authority for their own writing and balancing personal with external assessment, students also need to develop better self-monitoring processes because many do not have a language for thinking about their processes. (Tomlinson's survey of 23 secondary and college writing texts showed that there was very little figurative language in these texts). The speakers have found that by comparing metaphorical stories, students can gain confidence and learn that other writers (including professionals) may encounter similar problems. Students begin to talk like writers and develop a stronger interest in writing.

THE USES OF COMPUTERS IN THE ANALYSIS AND ASSESSMENT OF WRITING

Speakers:	*William Wresch*, University of Wisconsin-Stevens Point *Helen Schwartz*, Carnegie-Mellon University
Introducer/ Reporter:	*Marie Jean Lederman*, NTNW and Baruch College, CUNY

William Wresch discussed the current state of the field of computer analysis of student writing, dividing the software programs into six different categories, each of which has a different pedagogical orientation. The first category is error checkers. These programs focus on homonym confusions, sexist language, usage errors, and infelicitous phrases. Some examples are Writer's Helper (Conduit), Sensible Grammar (Sensible Software), RightWriter (RightSoft), Ghost Writer (MECC), and Writer's Workbench (AT&T).

The second category is reformatters which, rather than find errors, make it easier for writers to find their own errors. One of the first programs was Quill (DC Heath) which included a combination of prewriting, writing, and revising activities. For example, to help students revise their work, it displayed each sentence of their paper alone on the screen. Rather than make statements about or changes in the sentence, the program allowed students to look at each sentence in a new way. Other newer reformatters include Ghost Writer (MECC) and Writer's Helper (Conduit). The third category of programs is audience awareness programs. These programs include readability formulas and they pinpoint vague references and other problems.

The fourth category is student conference utilities. These computer programs try to help students develop editing skills as they read each other's papers and "send" comments to each other. Two examples are Quill and Alaska Writer (Yukon-Koyukuk School District). The fifth category is grading utilities, programs designed to help teachers in the clerical aspects of paper grading. Students turn in their work on disks, and the teacher uses the computer to help grade the work. By creating ten or twelve messages for major errors, teachers can respond with just a keystroke or two to most of the mistakes they are likely to see. Examples are the RSVP project (Miami-Dade Community College) and Writer's Network (Ideal Learning).

The last category is automatic graders. This is the logical "next step" after grading utilities. Ellis Page of the University of Wisconsin proved twenty years ago that a computer could grade papers quite well based on a formula of paper length, sentence length, level of subordination, and word length. However, merely assigning a grade isn't enough in a classroom situation in which students expect not only a grade but a range of responses from teachers. It might be possible, however, to use such computer graders in large-scale assessment programs. Wresch concluded that there are many decisions to be made about how computers will be used in writing analysis, but it is certain that there are already many opportunities and, surely, many more to come.

Helen Schwartz began by discussing several purposes of assessment: diagnosis and revision as well as improved self-evaluation. The range of writing behaviors which can be assessed are ideas, organization, rhetorical presentation (purpose and audience assessment) and grammatical correctness. In answer to the question, "How can computer programs assess these behaviors for these purposes?" she first gave a short answer, "No computer program alone is now accurate or helpful enough" and most of the existing programs may overwhelm the student with too much information at once. Style checkers can draw attention to problems, but the student must make the decisions. And sometimes readability formulas can lead students to vary sentence length by creating run-on sentences and fragments. Schwartz pointed out that "Computer programs are useful as delivery systems for teacher, peer and self-assessment. They help students become aware of problems in their writing and help them to solve these problems." She gave four examples:

1) Prewriting programs such as "ORGANIZE" (Helen Schwartz, Wadsworth Publishing) can be used not only to help students see the shape of their papers but also to desensitize peer review.

2) Templates, such as the self-evaluation form given in "Interactive Writing," help students assess strengths and weaknesses.

3) "SEEN" (Schwartz, Conduit) includes a built-in bulletin board where peer review can take place.

4) Programs for teacher and peer response to paper drafts, including (a) "Ch~~ and Comments," developed by Christine Neuwirth at Carnegie Mellon which facilitates discussion and peer review; (b) "PROSE" (Prompted Revision of Student Essays by Davis, Kaplan, Martin, McGraw Hill) which allows summary comments; comments embedded in the paper; revision notes; and handbook-like responses with an overview of the error, further explanation, and then interactive tutorials on each of 18 features; and (c) "Prentice Hall College Writer" which is a word processor that allows access to an on-line handbook and allows the insertion of comments that can include excerpts from the on-line handbook.

The discussion that followed centered on examples of software described and demonstrated by the speakers.

LEGAL RAMIFICATIONS OF WRITING ASSESSMENT

Speaker:	*William Lutz*, Rutgers University, Camden
Introducer/ Recorder:	*Chris Anson*, University of Minnesota

William Lutz, who holds a law degree and is a member of the Pennsylvania Bar, addressed the importance of considering the legal constraints under which testing must operated. Lutz began by distinguishing the different kinds of testing programs: those conducted within an institution and those conducted outside the institution. External testing programs, such as those conducted by a school district or by a state agency, are governed by a series of laws and court decisions. Internal testing programs, such as course placement and proficiency testing, come under fewer legal constraints and exist, at present, in a legal nether world. However, there is enough legal precedent to warrant caution by anyone involved in any testing program.

According to Lutz, testing programs may be attacked from a variety of legal approaches. Title VI of the Civil Rights Act prohibits any practice that would have the effect of restricting an individual, on the grounds of race, color, or national origin, "in the enjoyment of any advantage or privilege enjoyed by others receiving any service, financial aid, or other benefit." It is important to note that this law would judge a testing program by its effect, not its purpose. Moreover, the burden of proof in any legal action would fall on those conducting the test. Thus, under this law, testing programs with disproportionate effects on minority students are subject to close judicial scrutiny. If a state has a law guaranteeing an education to all its citizens, then all citizens have a property interest in an education. A testing program in that state can be attacked as a denial of a property right without due process. Such attacks have succeeded.

Lutz pointed out that a testing program can be attacked as a denial of a liberty interest. Due process guarantees a right to liberty, and this liberty interest is infringed where a stigma attaches to the student as a result of the test. The 14th Amendment to the Constitution states that "No person shall . . . deny to any person within its jurisdiction the equal protection of the laws." While state laws may treat differently for various purposes by classification persons who are similarly situated with respect to the purpose of the law," they must be accorded equal treatment. In hearing cases brought under this Amendment, the court will ask two questions: (1) has the state acted with an unconstitutional purpose? (2) has the state classified together all and only those persons who are similarly situated? For example, if someone wanted to attack a placement test there are two possible arguments under the 14th Amendment which might be used. First, the test itself can be attacked by arguing that while testing may be a legitimate means of classification, this particular test is so inadequate that one cannot possibly tell whether a particular student is ready for or has the ability to do college level work. A second approach is to attack the tests results by arguing that while the means used to classify a student may be legitimate, these means are so imprecise that one cannot possibly tell whether the student has been classified correctly.

There are some vague areas here, or the legal nether world as Lutz calls it. Before the due process requirements of the 14th Amendment can apply to a cause of action, two questions must be answered: (1) do the concepts of liberty or property encompass the asserted interest? and (2) if due process does apply, what formal procedures does due process require to protect the interest adequately? In other words, an individual must have a legitimate claim of interest before due process can apply. Thus far, a college education has not yet been found to be a benefit for which someone can assert a claim of entitlement. However, a claim of liberty could apply because testing may affect an individual's opportunity to choose his or her own employment. This issue is still open for litigation.

Based upon a review of federal court decisions, Lutz offered the following Guidelines for Testing:

1. The purpose of the test must be clearly delineated. The test must be matched with

specific skills and/or specific curriculum objectives.

2. Mere correlation between the test and the curriculum is not sufficient. There must be evidence, obtained from a regular process, that classroom activities are related to curriculum goals and test specifications.

3. All test items must be carefully developed and evaluated to ensure that they conform to curriculum and instructional practices. Moreover, there must be evidence that any bias related to racial, ethnic, or national origin minority status has been eliminated.

4. If possible, other measures of performance and ability should be used in conjunction with test results.

5. Cut-off scores should be the result of a well-documented process of deliberation that conforms to state and federal statutory requirements. There should be no suggestions of arbitrariness or capriciousness in setting cut off scores.

6. Students should be informed well in advance of what it is they need to know to perform well on the test. Students should also be informed in advance as to the nature of the test.

7. Options should be available for those students who fail th test. These should include, at the very least, the option to re-take the test, and institutional help to prepare and/or correct deficiencies.

8. Students should have access to their test scores and a full explanation of those scores.

Finally, Lutz suggested that anyone conducting a testing progra should do the following immediately:

1. Conduct a full, impartial review of the testing program, and document this review.

2. Examine all the documentation in the program, and write any necessary additional documentation.

3. Correct all the deficiencies identified in the program, and then document the process by which the deficiencies were identified and corrected.

4. Institute two procedures as a permanent part of the testing program:

   (1) a formal process for administering and conducting the testing program, including full documentation;

   (2) a formal review of the program conducted at regular intervals by an outside, impartial, objective reviewer.

Lutz concluded by saying that we live in a litigious age, and prudence suggests that those involved in testing be professional and institute the guidelines and take the steps he outlined in his talk.

SOME NOT SO RANDOM THOUGHTS ON THE ASSESSMENT OF WRITING

Speaker: *Alan C. Purves*, The State University of New York, Albany

As I near the end of a seven-year long comparative study of student performance in Written Composition sponsored by the International Association for the Evaluation of Educational Achievement, I should like to set forth some conclusions I have reached about writing assessment.

1. Written Composition is an ill-defined domain. There have been a few recent efforts at mapping the domain through an examination of writing tasks and through an examination of perceived criteria, but in general these have been ignored in most assessments of student performance. Most assessments tend to rely on a single assignment selected at random.

2. Written composition is a domain in which products are clearly the most important manifestation; the texts that students produce form the basis for judgments concerning those students. Teachers and assessors know that and so do students.

3. These products are culturally embedded, and written composition is a culturally embedded activity. The culture may be fairly broad or it may be relatively narrow such as the culture of a Lee Odell or an Andrea Lunsford, but students inhabit and produce compositions that reflect those cultures.

4. When a student writes something in a large scale assessment in the United States, what is usually written is a first-draft on an unknown assignment that is then rated by a group of people who make a judgment as to its quality. The result is an

index of "PDQ," Perceived Drafting Quality. Whether PDQ has any relation to writing performance or ability is unclear, although it is probably a fair index.

5. Given the fact that what is assessed is PDQ, it is little wonder that students see writing performance as comprising adequacy of content, handwriting, spelling, grammar, and neatness. Such is the case of the reports of secondary school students as to the most important features of the textual products of a school culture.

NATIONAL AND INTERNATIONAL WRITING ASSESSMENT: RESEARCH ISSUES

Speakers: *Alan C. Purves*, State University of New York at Albany
Thomas Gorman, National Foundation for Educational Research, Great Britain
Rainer Lehmann, Institute for Educational Research, Federal Republic of Germany

Introducer/ Recorder: *Wayne Fenner*, University of Minnesota

This session was the first of several sessions on research on international writing assessment. Alan Purves began with an overview of the background of the fourteen-nation Written Composition Study. Begun in 1980, this project is the most recent undertaken by the International Association for Educational Achievement (IEA). Previous studies have examined the teaching and testing of science, math, reading, foreign language, and civic education. Unlike earlier subjects, the domain of written composition is a cloudy one: it is both an act of communication and an act of cognitive processing. Researchers, then, had first to define this domain, both empirically and theoretically. After this phase of domain specification, researchers designed a series of specific writing tasks and writing purposes to be included in the study. Third, a five-point scoring scheme was devis'd that would be valid and reliable across languages and cultures. Finally, raters were chosen and trained.

Thomas Gorman discussed the results from a recent writing assessment program in England in order to clarify what can be learned from international studies and cannot be learned from separate, national writing assessment projects. The problem of domain specification seems to be culturally relative. The purpose of writing varies in its relation to general educational aims, and specific tasks may or may not reflect the kind of writing that is generally required of students in specific schools in particular cultures. There is, however, remarkable unanimity of assessment criteria and standards of performance across languages and cultures. Content, for example, as well as form, style, and tone appear to be rating factors utilized internationally. As a result of the IEA Study, we have learned more about the relative difficulty of various writing tasks, and we have gathered a great deal of information about background variable relative to writing performance. These variables include students' interest and involvement in life at school, plans for future education, amount of daily and weekly homework, and involvement of parents in the educational process.

Rainer Lehmann discussed the methodology of comparative writing assessment, specifically the application of multitrait-multimethod analysis to the problem of validating the analytical scoring scheme used by all countries in the IEA Study. Although his discussion was limited to results from the Hamburg data, Lehmann provided information from a non-English language context that appeared to confirm the IEA student's methods and findings.

TEACHING STRATEGIES AND RATING CRITERIA: AN INTERNATIONAL PERSPECTIVE

Speakers: *Sauli Takala*, University of Jvaskyla, Finland
R. Elaine Degenhart, University of Jvaskyla, Finland

Introducer/ Recorder: *Robin Murie*, University of Minnesota

This session reported on data gathered in the IEA (International Association for the Evaluation of Educational Achievement) study of Written Composition. The IEA study, now in its eighth year, is a large-scale examination of student writing in 14 countries (Chile, England, Finland, Hungary, Indonesia, Italy, the Netherlands, Nigeria, New Zealand, Sweden, Thailand, the USA, Wales, W. Germany). An internationally developed scoring system was used to rate the writing tasks in terms of organization, content, style, tone, mechanics, and handwriting. In addition, students, teachers, and schools filled out questionnaires. These data are now being examined in a number of ways.

Sauli Takala, one of the coordinators of this study, described patterns of agreement and disagreement among raters application of a five-point rating scale (which included the criterion "off the topic"). He found that raters behaved in a uniform manner. Most of the time, two readers were within one point of being in full agreement with each other. Beyond a one-point discrepancy on the rating scale, there was a significant drop in frequency (2 points off: 5-12%; "off the topic": 2.5-7.5%, 3 points off:

2.5-5%). He then discussed where on the scale these discrepancies were occurring. Agreement was greatest at the high end of the scale and least likely in the low-middle range of scores.

Takala then discussed where the rating of "off topic" appeared. In early discussions with colleagues, it was anticipated that this rating would pair up with ratings at the high end of the scale (an essay would be so creative as to elicit either "very good" or "off topic".) In fact, just the opposite was true: "off topic." appeared at the low end of the scale with "poor." Surprisingly, it also occurred in the middle range. Takala noted that perhaps some raters were unsure of how to score such essays and so chose a middle ground. In general, similarities between raters outweighed differences, lending credibility to further comparisons.

Elaine Degenhart, another coordinator of the IEA Study of Written Communication, looked at relationships between writing instruction and student performance, using data from the teacher questionnaires, and questionnaires on the background and curriculum of the schools involved in the IEA study. The purpose of her work was to identify some patterns in instructional approaches and to determine how well the variable that show these approaches discriminate between low, middle, and high achieving classes. The four main approaches that emerged were product, process, reading-literature, and a less well defined skills-oriented approach with emphasis on product. Based on mean scores on the writing tasks, classes were divided into achievement levels: 25% high, 50% middle, 25% low. The top two instructional approaches for each country were then examined in terms of how well they discriminate for the three levels of classes. Degenhart reported on findings from four of the countries: Chile, Finland, New Zealand, and the U.S.

The top two teaching strategies found for Chile were (1) a strongly student-centered approach with a process orientation and (2) a stronger product orientation. Here it appeared that low-achieving students had more process-centered teaching, whereas the product-centered approach distinguished well for the middle group. In Finland, the top two teaching strategies were (1) a reading-literature approach and (2) a process approach. The process approach did not distinguish between the top and bottom groups; the reading-literature approach was positive for low-achieving students. In New Zealand, the top two were (1) a teacher centered reading/literature approach and (2) a less clearly defined approach leaning toward process. Both discriminated between all three levels. In the United States, the top two approaches were (1) a structured reading/literature approach and (2) a strong student-centered product orientation. The product orientation was high for the low-level students.

Questions centered around possible interpretations of these findings. Degenhart was careful not to draw premature conclusions or make quick generalizations. From the discussion it became clear that a greater understanding of the background situation in each country would help with the interpretation of why classes were receiving a particular type of writing instruction.

EFFECTS OF ESSAY TOPIC VARIATION ON STUDENT WRITING

Speaker: *Gorden Brossell*, Florida State University
Jim Hoetker, Florida State University

Introducer/ Recorder: *Laura Brady*, George Mason University, VA

Gorden Brossell and Jim Hoetker presented the results of a study designed to analyze the ways in which systematic variations in essay topics affected the writing of college students under controlled conditions. To explore the question of whether a change in topic makes a difference in the quality of student response, Brossell and Hoetker chose extremes of topic and student population. The population consisted of remedial students and honors students writing in response to a regular course assignment. The year-long study (May 1987-April 1988) was based on 557 essays collected from four Florida sites: the University of Florida, Miami-Dade Community College, Valencia Community College, and Tallahassee Community College.

The general essay topic for this project, "The most harmful educational experience," was written according to procedures developed by Brossell and Hoetker in their previous research on content-fair essay examination topics for large scale writing assessments (CCC, October 1986). Brossell and Hoetker then varied this topic in two ways: (1) they controlled the degree of rhetorical specification and (2) they changed the wording to invite subjective and objective responses. These variations yielded four versions of the topic:

- Minimal rhetorical specification requesting an impersonal discussion
- Minimal rhetorical specification requesting a report of personal experience
- Full rhetorical specification requesting an impersonal account
- Full rhetorical specification requesting a report of personal experience

The essays written in response to these topic variations were scored holistically on a 7-point scale by experienced graders; the scale included operational descriptions for four levels of quality (1,3,5,7) and left the other three variables

(2,4,6) unspecified in order to give the raters greater flexibility. The essays were also scored analytically according to ten items in three categories: (1) development, (2) voice/speaker/persona, and (2) readability.

Although the original plan had been gathe samples from extreme student populations (high- and low-ability), differences between institutions ir the average quality of student writing were noticeable: many "low-ability" students wrote as well as or better than students ranked as "high-ability." As a result, the sample fell into a bell-curve distribution. The research concluded that there is no evidence from either the holistic-scale scores or the analytic-scale scores that even gross variations in phrasing affect either the quality of student responses or the nature of student-topic interaction. Other conclusions: the appearance of first-person voice is significantly higher in essays written in response to topics calling for accounts of personal experience, but it is unaffected by the degree of rhetorical specification.

In a discussion following the presentation of the research, Brossell and Hoetker mentioned plans for future work that include a study to evaluate the effect of content variation in essay topics when wording and rhetorical specification are held constant. They also plan to develop their analytic score further, based on additional essays written at greater leisure and revised, and representing average and high-ability students as well as low-ability students. With revision and development to make the scale reliable and "transportable," the analytic scale might, according to Brossell and Hoetker, have the potential to become an alternative to the single-digit holistic score.

WHAT SHOULD BE A TOPIC?

Speakers:	*Sandra Murphy*, San Francisco State University, *Leo Ruth*, University of California, Berkeley
Introducer/ Recorder:	*Robert L. Brown*, Jr., University of Minnesota

Taking a cue from the Bay Area Writing Project's collective spirit, Sandra Murphy and Leo Ruth rejected the usual panel format by opening the session to audience discussion of issues influencing subject-selection for holistic scoring. They directed the session with six questions (treated at greater length in their recent ABLEX book Designing Writing Tasks for the Assessment of Writing). Their questions examined the dual problem facing assessment designers: naming a subject and providing the writers with instructions about what to do with it. In part, the session provided a forum for a critique of both the entire agenda of holistic scoring and of the specifics of assessment design. But it also allowed Murphy and Ruth format in which to report some of the findings from their work.

The six questions treat variously the syntactico-semantic structure of the items, the discourse structures suggested, the power relationships established between test(er) and writer, and the cultural knowledge presupposed. The six questions and comments from the presenters and audience are as follows:

1. How much information should be provided about the subject?

   Murphy and Ruth's findings suggest that a simple referring phrase (NP) elicited less rich responses than a full proposition. When a predicate was provided, writer responses were more "reasonable and responsible."

2. How does specification of a subject constrain response?

   Discussion demonstrated the range of possible constraints: discourse type, qualification, quantification, text structure, style, and--always--ideology, explicit and implied.

3. How does knowledge of the subject affect performance?

   The session members soon raised the meta-question of whether any topic could not require "specialized knowledge," and therefore whether holistic essay testing could be free from political bias. Generally, Murphy and Ruth and the session members agreed that knowing a lot about the topic was a great advantage, and the "'knowledge'" extended well beyond simple propositional knowledge to familiarity with cultural discourses.

4. Should students be given options in selecting topics?

   Generally, options invite confusion. Items may not be equally difficult. Students may not be wise in selecting, picking complex topics and writing complex, bad essays. Confusion over the selection process may penalize.

5. How do rhetorical specifications affect performance?

   Students did not seem to be helped by suggestions of rhetorical type. Typically, they ignored them or found that the problem of executing the rhetorical command interfered with their writing in general.

6. To what extent should admonitions about the writing task be mentioned? Time limits, pitfalls, and so on?

   Again, the political demands of the writing assessment as an institution overwhelms the testers' attempts to help: students write the essay they have in mind, ignoring the instructions or finding themselves confounded by them.

The session eloquently expressed reservations about the ideology of holistic scoring and mass assessment in general. The conferees reacted to the inherent artificiality of pretending to write authentic prose while authentically demonstrating familiarity with academic conventions. They agreed that students who know the conventions of testing will, predictably, do best.

CLASSROOM RESEARCH AND WRITING ASSESSMENT

Speaker:	*Myles Meyers*, California Federation of Teachers
Introducer/ Recorder:	*Deborah Appleman*, Carleton College, Minnesota

Myles Meyers addressed the issue of large scale assessment from the perspectives of the K-12 administrator and classroom teacher. From these perspectives he finds large scale assessment to be problematic and often ill-advised. The enormous diversity of schools makes it difficult to capture the current "state of the art." Myers also contended that state assessments such as California's CTBS work against teaching as well as against the professionalization of teachers.

Meyers discussed at length the seemingly reductionist quality of large scale assessment. Although recent research on writing maintains that writing is a multiple construct, time and financial constraints limit the constructs that can be examined. The construct that is employed to define writing thus becomes the primary focus for a particular grade (for example, autobiography in grade 10). In our effort to handle the assessment task by limiting constructs, our definition of writing, as well as its instruction, therefore becomes uni-dimensional. Moreover, because of the inevitable prescriptive quality of the interpretation of assessment results as well as teachers' lack of involvement and consequently lack of ownership in the entire assessment process, Meyers claimed that statewide assessments can destroy teaching-as-inquiry and harm student learning.

Meyers then presented several suggestions for involving teachers in the assessment process. He emphasized the importance of having teachers participate significantly through summer institutes at university settings. He also underscored the importance of viewing assessment as a process of inquiry, one in which disagreement is as important as agreement. To illustrate the value of assessment as inquiry, Meyers handed out three sample student papers and asked the audience to rank them as high, middle, and low. The resulting scoring was quite discrepant, as were the reasons offered for the rankings. Meyers then discussed the value of discrepancy in our aim to improve literacy for all children. Rather than considering agreement as the ultimate goal in assessment, discrepancy can lead to a fruitful dialogue about our underlying assumptions about teaching good writing as well as about its evaluation.

Meyers pointed out that dialogues or debates such as those generated by the conferees when they were asked to rank the papers were a critical aspect of the assessment process. He stressed the importance of having classroom teachers as active participants in an on-going debate on assessment, rather than as recipients of an administrative decision to employ a particular large scale assessment instrument. He then handed out six additional student papers, and asked conferees to rank them and then to discuss the rankings in pairs. As with the first exercise, the rankings were widely discrepant. Meyers illustrated how this kind of exercise can be used to encourage teachers to think explicitly about their pedagogy and also described several ways in which the ranking of student writing can be employed to generate discussion among teachers. For example, he has asked teachers to devise sample lessons for students whose papers they have ranked.

Meyers ended his provocative discussion by suggesting several ways in which writing can be viewed as a speech act and as a collaborative social event. He discussed the differences and similarities between conversation and written presentation. Meyers concluded his talk with the following thought: "When you teach people how to write, you teach them a new definition of themselves."

COMPUTERS AND THE TEACHING OF WRITING

Speakers: *Michael Ribaudo*, The City University of New York,
Linda Meeker, Ball State University
Introducer/ Recorder: *Donald Ross*, University of Minneapolis

Both speakers discussed the National Project on Computers and College Writing, a three-year project supported by the Fund for the Improvement of Postsecondary Education and The City University of New York. This project is coordinated by three of the NTNW directors: Michael Ribaudo, Harvey Wiener, and Karen Greenberg.

Michael Ribaudo explained the goals of the project: it will (1) identify outstanding college programs that have incorporated computers in freshman-level composition courses, (2) conduct research on the impact of computers on students' writing abilities, (3) develop and disseminate reports on this research and on instructional philosophy and methodology, and (4) host a national conference showcasing the programs and the research.

Ribaudo noted that, at this point in time, fifteen colleges and universities from across the country are involved in the project. They are developing research designs that will pair three "computer" sections and three traditional sections and three traditional sections at each site. Some of the research instruments to assess students include essay tests (scored holistically and analytically), multiple-choice tests, and questionnaires on writing anxiety and writing attitudes.

Linda Meeker discussed her university's participation in the project, and summarized the efforts that Ball State has already made in evaluating the effects of computers on the teaching and learning of writing.

She described three of her recent studies. The first study assessed student attitudes toward using computer-assisted instruction (CAI) in basic writing classes. She found that CAI proved effective in terms of students' time management and that basic writing students developed positive attitude toward CAI. Her second study focused on using "invention" software to assist the composing processes of basic writing students. Results indicated highly positive student attitudes and a noticeable improvement in students' ability to focus on their topics. Meeker's third study examined the revising strategies of basic writing students. This study revealed that students spent significant amounts of time in a variety of prewriting and revising activities, but it was unclear whether the text manipulations were clearly related to a greater flexibility provided by CAI. However, Meeker did find that the computer enabled students to do more frequent --and more productive--pre-editing.

Next, Meeker described some of the studies that will be conducted by the National Project on Computers and College Writing. She noted that data collected from these large scale assessments will either confirm or call into question the results of her studies. Students attitudes toward CAI and the effectiveness of word-processing as a tool for inventing, composing, revising and editing will be evaluated. Moreover, each of the project sites will examine the comparative effectiveness of different hardware and software configurations available at their institutions.

For further information on this project, or the conference which is scheduled for Spring 1990, write to Dean Ribaudo at CUNY, 535 East 80th Street, New York, NY 10021.

ANNOUNCING: New works on writing assessment by NTNW members:

THE IEA STUDY OF WRITTEN COMPOSITION: THE INTERNATIONAL WRITING TASKS AND SCORING SCALES
Edited by T.P. Gorman, A.C. Purves, and R. E. Degenhart
Pergamon, Oxford, England, 1988

THE EVALUATION OF COMPOSITION INSTRUCTION, ***Second Edition***
by Barbara Gross, Michael Scriven, and Susan Thomas
Teachers College Press, NY, 1987

DESIGNING WRITING TASKS FOR THE ASSESSMENT OF WRITING
by Leo Ruth and Sandra Murphy
Ablex, Norwood, NJ 1987

How You May Participate

NTNW needs the active participation of those who have a concern with writing skills assessment, whether as specialists, administrators, or classroom teachers. If you wish to become a member of the network or to learn more about who we are, what we plan to be doing, and how our plans could involve you, just complete the coupon below and return it to us along with materials describing yourself and your professional interests in writing instruction and assessment.

Name____________________________________

Position__________________________________

Institution________________________________

Address___________________________________

____

____

____I would like to be on NTNW's mailing list.

____I would be willing to share information about my writing assessment program with NTNW.

Please return to:

Karen Greenberg, Director
National Testing Network in Writing
Office of Academic Affairs
The City University of New York
535 East 80th Street
New York, New York 10021
(212) 794-5446