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 disagreements 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 Auccher, GED Testing Service
Introducer/Recorder: Anne Aronson, University of Minnesota

Wayne Patience and Joan Auccher 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

14
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 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