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.

White 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. White 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 White’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, NTINW 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 himself/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
Introducer/Recorder: Anne Aronson, University of Minnesota

Wayne Patience and Joan Achtzer 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
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.

White 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. White 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 Bernardo
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 Auclter, GED Testing Service
Introducer/Recorder: Anne Aronson, University of Minnesota

Wayne Patience and Joan Auclter 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