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.