fewer conflicts emerged. Recent history of the Minnesota program suggests that assessment and writing across the curriculum projects with a local focus have an excellent chance, where massive programs mandated from the top down are deeply problematic.

ASPECTS OF HOLISTIC SCORING VALIDITY

Victor Froese, University of British Columbia

This session focused on a study of the validity of holistic scoring conducted by researchers at the University of British Columbia. Three related observations prompted this study: 1) insufficient research has been done as to whether readers trained in holistic rating base their judgments on substantive or superficial characteristics; 2) compositions can be scored more quickly by computer than by trained markers; and, 3) the validity of holistic scoring has not been satisfactorily demonstrated.

This study compared holistic scores with composition length, number of spelling errors, and sentence length—the first two being distractors frequently associated with holistic scoring. For each of grades 3, 6, and 9, teachers scored forty randomly selected narrative and forty randomly selected explanatory passages holistically and via the Writer's Workbench software (which scores many other features as well). A graduate research assistant entered the same compositions for computer analysis, being careful to encode the text and spellings exactly as in the originals.

Findings from statistical analyses indicated that grade-by-type interactions existed for holistic scores, sentence length, and passage length but not for spelling errors. Subsequent analyses indicated that holistic scores predicted sentence length for Grade 3 narrative and explanatory text; passage length for narrative text in Grades 3, 6, and 9; and explanatory text in Grades 3 and 9. Spelling errors could not be predicted from holistic scores. When sentence length, passage length, and spelling errors were combined, narrative text scores could be predicted at Grades 6 and 9 but explanatory text only at the Grade 9 level. This implies that some of the mechanically counted features (sentence length, passage length) of the Writer's Workbench predict holistic scores in a statistically significant manner (except for expository text at the Grade 6 level).

The researcher concluded that holistic scoring is apparently sensitive to the sometimes irrelevant factors such as sentence length and passage length, but that these operate somewhat differentially at different grade levels and for different types of writing. Also, the parsimony of computer-based composition scoring should not be overlooked since it is accurate, less costly, and faster than human-based holistic scoring. Finally, because the trained observers' scores were not related to misspellings but to sentence length and passage length, the validity of holistic scoring still needs further investigation in order to explain this relation.

COLLECTING EVIDENCE IN SUPPORT OF CONSTRUCT-RELATED VALIDITY

Sybil B. Carlson, Consultant, Educational Assessment and Research

One category of evidence for validity, that of construct-related evidence, is particularly critical to tests of writing skills in which readers assign scores to papers written by examinees. Because the direct assessment of writing relies on human judgments, its quality greatly depends on the validity of the perceptions invoked by individuals when making these judgments. Thus it is essential to focus on the process evaluation—a creative act of interpretation that is influenced not only by the perceptions of readers but also by the perceptions of the various audiences involved in the assessment.

The construct of writing competence must be described and defined within the context of a specific assessment. Furthermore, since readers are expected to apply that construct objectively and systematically in a testing situation, evidence must be collected to support inferences that the intended construct actually is being applied—by readers as well as persons who interpret and use test results.

Attention must be directed, not only to the mechanisms necessary for the implementation of the program, but also to the construct that is being
measured, and to how effectively it is actualized throughout all components of the program. This definition should be articulated at the inception of the program, in interaction with designers, writers, evaluators, instructors, and interpreters. Subsequently, they develop and refine this definition while the program is being designed, and examine it continuously during the evaluation of compositions, when they interpret and use the results, and when they realize the consequences.

Validity is not an issue that is reserved for investigation after the results are in, or only for experts in measurement and statistics. The demonstration of validity is a creative endeavor in which those who understand writing and the writing process are responsible for collecting evidence from as many different sources as possible to support inferences that a particular writing assessment is valid.

HOW DO TOPICS MEAN?

Sandra Murphy, San Francisco State University
Leo Ruth, University of California, Berkeley

This session addressed several questions that have long been largely ignored in writing assessment. We have usually assumed that each writer receives the same message to direct the writing performance. But contemporary reader response theory and constructivist theories of reading cast doubt on this assumption, requiring us to ask the question: What determines the meanings and effects of given writing topics? We called upon transactional theories of literary response as well as other recent constructivist approaches to the study of reading comprehension to develop a model of the participants’ constructions of meanings in a writing assessment episode. Our model suggests that readers of the topic (both the student writers and the teacher raters) choose among cues embedded in the text of the task, both honoring and ignoring elements which may enable them, with varying degrees of success, to match the test-makers’ intentions and expectations. We reported results of investigations confirming that the “meaning potential” of any given task for a student-writer or a teacher-rater is relative to the linguistic, cognitive, and social reverberations set off in the respondents.

Next, we questioned an assumption underlying much of current writing assessment practice, the assumption that there is only one objective reading of a topic text. We also argued that in order to make adequate evaluative judgments, it is necessary to consider the question: How does the topic text interact with the social context to affect the interpretations of test-makers and test-takers? Drawing on topics that have been used in actual writing assessments, we provided several examples of the way the informational content of a topic can interact with the social context and the testing event to provide unexpected variations in interpretation among respondents.

We then questioned the fairness of the practice of arbitrarily declaring a student “off topic,” drawing attention to a distinction between “selective reading” and “misreading,” suggesting that a more fruitful approach to the evaluator’s reading allows for a range of variation in student response.

Our materials called into question certain commonplace practices in writing assessment. We argued that we need to move toward a model of writing assessment that respects the complexity of individual responses to writing topics. That model must accommodate a range of reasonable interpretations of the task that may fall outside the narrow pre-ordained structures of primary trait rubrics. Contemporary reading theory provides insight that we can no longer ignore into ways that personal meanings are treated in the minds of the writers not only as they read the text of the prompt, but also as they write, read, and re-read the text that they are creating.