This presentation categorized the primary approaches people use in computer assessment of writing.

The first group is the "Usage Police." Programs such as Right Writer and Grammatik review a text and identify "errors" such as homonym confusion, passive voice, and minor grammar mistakes. I pointed out that accuracy is still a problem with such programs, with students being told something is wrong even when it may be correct. A bigger problem is triviality. While confused homonyms may be a nuisance, they are hardly central to the task of writing.

The second group was "Stylists." A primary program here is HOMER, which looks for "bureaucratic" writing by checking for such traits as long sentences, heavy preposition use, passive voice, and nominalization. With this and similar programs, the computer has a model of "good" writing and marks elements that vary from this norm. The problem here is that students may take the computer too seriously and come to believe that nominalization or passive voice are never acceptable.

I named the third group "Visionaries." These programs reposition text so writers can more easily see for themselves what revisions are necessary. Example programs are Quill, Writer's Helper, and Writer's Workbench. An example activity is Writer's Helper's Outline module that automatically prints the first sentence of each paragraph so writers can more easily see if they moved logically from one idea to the next, or jumped around erratically. Other modules print sentences individually, print the first and last sentence of each paragraph, count word frequencies, etc. In every case the writer must take responsibility for deciding what revisions to make, if any.

I named the last group "Phone Operators," newer networks that allow writers to communicate with each other as they write. Such networks promote peer revision. This approach once again uses the computer as a facilitator, but leaves responsibility for changes to writers themselves.

ASSESSING THE COMPOSING PROCESS

Lee Odell, Texas Christian University

Over the past several years, TCU has developed a number of procedures to help students engage in the composing process. Students keep journals, discuss topics in collaborative thinking groups, and present drafts to peer response groups to stimulate their thinking processes and refine their ability to articulate to others their ideas, perceptions, and feelings.

There are strong theoretical reasons for asking students to do this sort of work. But in order to improve our basic understanding of what we are asking students to do, and to help them make full use of these procedures, we must have some way of describing and assessing the work students do when they write in journals or take part in small group discussions. Since one goal of these procedures is to help students through the topics they are writing about, we need some way to describe the thinking that is reflected in students' work. And we also need some way to describe the interpersonal strategies students use when they work in groups.

The means of describing thinking can be synthesized from current work on thinking (work that includes rhetorical theory, cognitive psychology, and critical thinking). Six principal concepts have proven useful: Selecting and encoding; creating and acknowledging dissonance; considering alternatives; seeing relationships; drawing on prior knowledge; and using metacognition.

For analysis of interaction, the theory of Carl Rogers and the interaction process analysis of Robert Bales provide useful definitions of listening, one of the key factors in successful group interaction. Specifically, it can be useful to determine whether students are doing such things as: avoiding interruptions, paraphrasing or reiterating others' comments; responding to others' questions, comments or requests; inviting information or opinions from others; developing
others' statements; and dealing effectively with disagreement.

These definitions of thinking and listening have proven useful in working with students in grades 2-12 in the Fort Worth Writing across the Curriculum project. They enable teachers of math, social studies, science, and English to understand ways students currently negotiate the composing process. They also help teachers determine what students need to do in order to improve their work, and then plan classroom activities to help bring about these improvements. These assessments of the writing process make it considerably easier to assess and to help improve students' written products.

MAINTAINING SCORING STANDARDS IN LARGE-SCALE ASSESSMENT: A HERACLITEAN PERSPECTIVE

Belita Gordon, The University of Georgia

This session opened with questions about a revered assessment tenet: namely, that scoring standards remain constant over time. While score stability is necessary to ensure fairness, the pragmatics of large-scale testing make consistency difficult to attain. Instruction improves the quality of student writing, different contractors rate the same state's papers from year to year, prompts of inherently different difficulty are administered, and scoring guides and training procedures are refined. We respond with archival data and statistical equating procedures that anchor us to our past, obligating us to repeat not only our successes, but also our errors. Stability may be paramount, but this does not negate the need for change. The Basic Skills Writing Test (BSWT) is a case study of the benefits of change.

The BSWT is administered to Georgia's tenth graders as one of the requirements for a high school diploma. In the development stages, teachers expressed the desire for a diagnostic test. Consequently, they rejected a single holistic score in favor of five domain scores. The domains are holistically scored on a four-point scale. Two are weighted (the Content/Organization score is multiplied by three and the Style score by two), while the scores in Sentence Formation, Usage, and Mechanics are taken at face value. Students receive a total score and diagnostic statements for each domain. Over the two pilot years, the metric was changed from a two- to a four-point scale, as the two-point scale did not reflect the range of writing produced, and the Content and Organization domains were combined, as raters were unable to differentiate between the two.

An evaluation of the ratings on 1,866 papers, following the first operational year, strongly suggested that the domain subscores were too closely related to provide the desired diagnostic information. For the second operational year, the training procedure was revised and a comparative study implemented. Scoring accuracy was monitored, by domain, on the basis of agreement with prescored papers embedded within the packet of 80,000 "live" papers. The 1,866 anchor papers were scored again.

The analysis of the anchor papers revealed greater discrimination between domains. The highest correlation (between Content/Organization and Style) dropped from .72 to .49 while the lowest (between Content/Organization and Usage) dropped from .49 to .40. An increase in the absolute mean difference for each comparison further reinforced evidence of greater differentiation. The frequency with which raters assign the same score in all domains has dropped from 42% to 25% on the first rating, from 37% to 22% on the second rating, and from 19% to 7% on both ratings. While all these analyses suggest greater discrimination, they do not answer important questions. Does the test measure different subskills? Do the subscores provide instructionally useful information? Is writing a "unitary trait," and if not, how much discrimination is necessary before different attributes (domains) can be reported? Participants discussed these and related questions.