specific skills and/or specific curriculum objectives.

2. Mere correlation between the test and the curriculum is not sufficient. There must be evidence, obtained from a regular process, that classroom activities are related to curriculum goals and test specifications.

3. All test items must be carefully developed and evaluated to ensure that they conform to curriculum and instructional practices. Moreover, there must be evidence that any bias related to racial, ethnic, or national origin minority status has been eliminated.

4. If possible, other measures of performance and ability should be used in conjunction with test results.

5. Cut-off scores should be the result of a well-documented process of deliberation that conforms to state and federal statutory requirements. There should be no suggestions of arbitrariness or capriciousness in setting cut-off scores.

6. Students should be informed well in advance of what it is they need to know to perform well on the test. Students should also be informed in advance as to the nature of the test.

7. Options should be available for those students who fail the test. These should include, at the very least, the option to re-take the test, and institutional help to prepare and/or correct deficiencies.

8. Students should have access to their test scores and a full explanation of those scores.

Finally, Lutz suggested that anyone conducting a testing program should do the following immediately:

1. Conduct a full, impartial review of the testing program, and document this review.

2. Examine all the documentation in the program, and write any necessary additional documentation.

3. Correct all the deficiencies identified in the program, and then document the process by which the deficiencies were identified and corrected.

4. Institute two procedures as a permanent part of the testing program:

   (1) a formal process for administering and conducting the testing program, including full documentation;

   (2) a formal review of the program conducted at regular intervals by an outside, impartial, objective reviewer.

Lutz concluded by saying that we live in a litigious age, and prudence suggests that those involved in testing be professional and institute the guidelines and take the steps he outlined in his talk.

**SOME NOT SO RANDOM THOUGHTS ON THE ASSESSMENT OF WRITING**

**Speaker:** Alan C. Purves, The State University of New York, Albany

As I near the end of a seven-year long comparative study of student performance in Written Composition sponsored by the International Association for the Evaluation of Educational Achievement, I should like to set forth some conclusions I have reached about writing assessment.

1. **Written Composition is an ill-defined domain.** There have been a few recent efforts at mapping the domain through an examination of writing tasks and through an examination of perceived criteria, but in general these have been ignored in most assessments of student performance. Most assessments tend to rely on a single assignment selected at random.

2. **Written composition is a domain in which products are clearly the most important manifestation; the texts that students produce form the basis for judgments concerning those students.** Teachers and assessors know that and so do students.

3. **These products are culturally embedded, and written composition is a culturally embedded activity.** The culture may be fairly broad or it may be relatively narrow such as the culture of a Lee Odell or an Andrea Lunsford, but students inhabit and produce compositions that reflect these cultures.

4. **When a student writes something in a large scale assessment in the United States, what is usually written is a first-draft on an unknown assignment that is then rated by a group of people who make a judgment as to its quality.** The result is an
index of "PDQ," Perceived Drafting Quality. Whether PDQ has any relation to writing performance or ability is unclear, although it is probably a fair index.

5. Given the fact that what is assessed is PDQ, it is little wonder that students see writing performance as comprising adequacy of content, handwriting, spelling, grammar, and neatness. Such is the case of the reports of secondary school students as to the most important features of the textual products of a school culture.

NATIONAL AND INTERNATIONAL WRITING ASSESSMENT: RESEARCH ISSUES

Speakers: Alan C. Purves, State University of New York at Albany
           Thomas Gorman, National Foundation for Educational Research, Great Britain
           Rainer Lehmann, Institute for Educational Research, Federal Republic of Germany

Introducer/Recorder: Wayne Fenner, University of Minnesota

This session was the first of several sessions on research on international writing assessment. Alan Purves began with an overview of the background of the fourteen-nation Written Composition Study. Begun in 1980, this project is the most recent undertaken by the International Association for Educational Achievement (IEA). Previous studies have examined the teaching and testing of science, math, reading, foreign language, and civic education. Unlike earlier subjects, the domain of written composition is a cloudy one: it is both an act of communication and an act of cognitive processing. Researchers, then, had first to define this domain, both empirically and theoretically. After this phase of domain specification, researchers designed a series of specific writing tasks and writing purposes to be included in the study. Third, a five-point scoring scheme was devised that would be valid and reliable across languages and cultures. Finally, raters were chosen and trained.

Thomas Gorman discussed the results from a recent writing assessment program in England in order to clarify what can be learned from international studies and cannot be learned from separate, national writing assessment projects. The problem of domain specification seems to be culturally relative. The purpose of writing varies in its relation to general educational aims, and specific tasks may or may not reflect the kind of writing that is generally required of students in specific schools in particular cultures. There is, however, remarkable unanimity of assessment criteria and standards of performance across languages and cultures. Content, for example, as well as form, style, and tone appear to be rating factors utilized internationally. As a result of the IEA Study, we have learned more about the relative difficulty of various writing tasks, and we have gathered a great deal of information about background variable relative to writing performance. These variables include students' interest and involvement in life at school, plans for future education, amount of daily and weekly homework, and involvement of parents in the educational process.

Rainer Lehmann discussed the methodology of comparative writing assessment, specifically the application of multitrait-multimethod analysis to the problem of validating the analytical scoring scheme used by all countries in the IEA Study. Although his discussion was limited to results from the Hamburg data, Lehmann provided information from a non-English language context that appeared to confirm the IEA student's methods and findings.

TEACHING STRATEGIES AND RATING CRITERIA: AN INTERNATIONAL PERSPECTIVE

Speakers: Sauli Takala, University of Jyväskyla, Finland
           R. Elaine Degenhart, University of Jyväskyla, Finland

Introducer/Recorder: Robin Marie, University of Minnesota

This session reported on data gathered in the IEA (International Association for the Evaluation of Educational Achievement) study of Written Composition. The IEA study, now in its eighth year, is a large-scale examination of student writing in 14 countries (Chile, England, Finland, Hungary, Indonesia, Italy, the Netherlands, Nigeria, New Zealand, Sweden, Thailand, the USA, Wales, W. Germany). An internationally developed scoring system was used to rate the writing tasks in terms of organization, content, style, tone, mechanics, and handwriting. In addition, students, teachers, and schools filled out questionnaires. These data are now being examined in a number of ways.

Sauli Takala, one of the coordinators of this study, described patterns of agreement and disagreement among raters application of a five-point rating scale (which included the criterion "off the topic"). He found that raters behaved in a uniform manner. Most of the time, two readers were within one point of being in full agreement with each other. Beyond a one-point discrepancy on the rating scale, there was a significant drop in frequency (2 points off: 5.12%; "off the topic": 2.5-7.5%, 3 points off: 3.2%)