Miles Myers

PROTOTYPE THEORY AND HOLISTIC SCORING

The issues in the assessment of writing keep changing. At one time, the primary issue was the multiple-choice test versus the writing sample, but when scores of holistic assessments were shown to have excellent reliability, writing samples were generally accepted as the most appropriate data for writing assessments. The primary issue is now the extent to which writing is a combination of what should be collected (what topics or prompts, how many samples), and how the assessment should be anchored. It is the second question which I want to discuss.

There are those who argue that the best way to begin anchoring a group of readers is to give them a list of features—coherence, clear introduction, and so forth. Others argue that the best way to begin anchoring a group of readers is to have them score unmarked samples of the various score categories—1, 2, 3, and so forth. The argument is entirely a matter of emphasis because those who emphasize features use samples as supplementary information, and those who emphasize samples usually recommend that feature analysis be a follow-up activity to holistic scoring. At bottom, the issue is how do people form categories, with a list of features or with a sample.

A number of linguists have recently begun to challenge the traditional use of a checklist of semantic features to determine the necessary and sufficient conditions for set-membership in the category designated by a word. Key and McDaniels (1979), for instance, have shown that there are no necessary and sufficient conditions of redness, but rather red is a gradient quality whose category boundaries are fuzzy. George Lakoff (1972), Charles Fillmore (1976), Key, and others argue that checklist semantics should be replaced by prototype semantics in which word meaning is determined by a central prototype or typical use of the word, not the yes/no category boundaries of a list of features. One example of the problem is Fillmore’s question, “How old is a bachelor? Sixteen? Twenty-one? Thirty?”

The work of Eleanor Rosch has provided additional empirical evidence that people use prototypes, not a list of features, to define categories. Rosch (1977) had people compose sentences with category names like bird, fruit, and vehicle. She then replaced the category name (for instance, bird) with a member of the category (for instance, robin, eagle, ostrich) and asked people to rate how sensible the resulting sentences were. The prototypical bird is something like a robin, producing sensible results more often than chicken or ostrich or numerous other choices.

Rosch and Mervis (1975) also found that when people are asked to list the important features of a given member of a category, the list turned out not to match exactly the important features of another member of the same category. Rosch and Mervis asked people to list the characteristic attributes of twenty different kinds of fruit, and for these fruits, from apples to olives, there was no feature list common to all.

In prototype theory, then, a category is defined by a prototype or central tendency member which contains the maximal number of features in common with members of its own set and a minimal number of features in common

(Continued on page 20)

Lawrence Biemiller

MICROCOMPUTERS EXPECTED TO BRING RADICAL CHANGES IN TESTING

Microcomputers will alter standardized admissions and placement tests radically in the next few years, speakers at the Educational Testing Service’s annual invitational conference said.

They predicted that within 10 years:

- Students would take the Scholastic Aptitude Test and other examinations individually, using microcomputers at their high schools or colleges, rather than in large groups at times set by the E.T.S.
- Computerized tests would be “interactive” with the student—asking ever more challenging questions until he begins to give incorrect answers.
- Computer programs would not only rank a student in relation to others in his class but would analyze his strengths and weaknesses and, if necessary, begin remedial instruction on the spot.

Interactive tests, psychometricians here said, are considerably more efficient than paper-and-pencil versions now offered by the E.T.S. and other test manufacturers.

"After every item, the computer estimates the candidate’s test score, based on the information so far, and selects as the next item one that will give the most additional information about the person’s score," explained Bert F. Green, Jr., a professor of psychology at the Johns Hopkins University.

Because interactive tests do not waste time with items that are either too simple or too difficult for the student, as a conventional version must do to accommodate all those being tested, they can be completed in about half the time.

"We don’t learn anything when a student answers correctly an item that was too easy for him, or answers incorrectly an item that was too hard," said Ernest J. Anastasio, vice-president for research management at the testing service, in an interview. “The point is to converge as rapidly as possible on the student’s range of ability.”

The Educational Testing Service, the country’s most prominent test manufacturer, already offers a computerized career-counseling service called System for Interactive Guidance Information, or SIGI. Interactive versions of various E.T.S. tests are now being developed and an interactive basic-skills test for use by college placement officials—the first large-scale E.T.S. test to be computerized—is expected to be available in the fall of 1984.

OTHER ADVANTAGES

A number of standardized tests used in fields outside of education would also be affected by advances in microcomputer technology.

For example, Mr. Anastasio said, the Armed Services Vocational Aptitude Battery, a standardized test administered to military recruits, now requires about eight hours—often over two days—to administer to each of the 1.2 million people who take it each year. Cutting the time

(Continued on page 20)
MYERS (continued)

with members of other sets. In writing assessment, this means that the anchor papers for the one category are like the robin of the bird category. There are papers which are like chickens. That is, they are not exactly like the anchor prototype but they are close enough. Defining score categories by a list of features can lead readers to expect scoring to be a matter of yes or no, not more or less.

The more-or-less quality of paper scoring does not mean that features are ignored. Prototypes can be used to describe cases in which some features of the prototype for the category are more heavily weighted than others. But the more-or-less quality of paper scoring means the scoring is an approximation, not an adding up of features. Zadeh (1973), the father of the fuzzy set theory, has argued that in humanistic systems, people reason by approximations of fuzzy sets, not by precise, quantitative terms. Says Zadeh, "...as the complexity of the system increases, our ability to make precise yet significant statements about its behavior diminishes until a threshold is reached beyond which precision and significance (or relevance) become mutually exclusive characteristics." (Zadeh, 1973, p. 26).

In summary, then, from linguistics (Fillmore, Lakoff, Kay), from psychology (Rochs, Mervis), and from philosophy (Zadeh), there is good evidence that the categories people use in their daily problem solving do not have precise boundaries and are defined first by matching to prototypes, not a list of features. Holistic scoring which emphasizes the anchor papers, not a list of features, is following the same set of principles.


BiMillier (continued)

In half would save the Defense Department a considerable amount of money, he said.

Computerized testing, Mr. Green said, would offer several other advantages:

- The student would work at his own pace.
- Answer sheets—and the possibility that a student might mistakenly fill in the wrong blank—would be eliminated.
- Incorporating videodisc players into microcomputer systems would make it possible to include action sequences in tests. This would make them less dependent on narrative passages that place an emphasis on reading ability even when other skills are being tested—as is the case, for example, with examinations for fire fighters or police officers.
- It would be possible to score tests instantly.
- It would also be possible—as the sophistication of computer programs for tests increased—to eliminate multiple-choice questions in favor of items requiring "constructed responses"—answers consisting of numbers or words typed into the computer. This is important, said Mr. Anastasio, because "selecting the right response from among five is a very different behavior from actually working out and writing down an answer."

The psychometric methodology on which interactive testing is based has been understood for some time, Mr. Anastasio said, but test manufacturers have been obliged to wait until microcomputers became available at a reasonable price.

Mr. Green said it would be necessary to create computer programs that would be able, in choosing which questions to ask the student, to select items on the basis of both difficulty and variety of content. Programs must be created, he said, to assemble tests that are as well balanced as those produced by an experienced test editor.

Mr. Green also warned that "large amounts of data" would be needed to ensure that items are properly "calibrated" or given weights in the scoring process that correspond exactly to their difficulty. Furthermore, interactive tests require a much larger bank of items on which to draw than do conventional examinations.

After such technical problems are overcome, Mr. Green and others here predicted, it will be possible to integrate standardized testing in high school and college curricula, making it a far more useful tool than it has been to date.

W. James Popham, a professor in the Graduate School of Education at the University of California at Los Angeles, told conference participants to "take advantage of the new situation" and create tests that will "serve as instructional catalysts." Microcomputers, said psychometricians here, will allow such tests to be administered as often as necessary—rather than once or twice in a student's career, as is now generally the case.

Lawrence Biemiller is an Assistant Editor of The Chronicle of Higher Education, this article, which appeared November 10, 1982, is reprinted in its entirety with the permission of the author.