The NH Assessment System (NH SAS) will employ AutoScore to machine score student essay questions.

How it Works

AutoScore uses a statistical model to predict the scores of human raters based on a set of features found in the response. The features have been identified in essays previously scored by expert raters using an established scoring rubric. AutoScore does not attempt to read a student essay response.

Like Netflix, Amazon, Facebook, etc. predict your consumer choices based on a matrix of indicators in your online behavior; AutoScore predicts the scores based on a matrix of identified features in each response.

The Process

Learning:
1. AutoScore learns the scoring rubric by identifying scored features of sample essays.
2. AutoScore associates identified features with the scores assigned by expert raters.
3. Response features that do the best job of “explaining” the raters’ scores are combined into a model.
4. The model is validated using a sample of more than 500 responses not included during the model learning phase.
5. Disagreement between rater scores is resolved so there is a high degree of confidence in the scores.

Scoring:
6. AutoScore identifies features in response essays.
7. AutoScore associates identified features in response essays with the scores assigned by expert raters.
8. AutoScore uses matched relations to predict each individual response.

Component Features

- Grammatical and syntactical aspects: spelling errors, sentence length, patterns of communication.
  - Hundreds of linguistic features are used to describe responses in mathematical terms.
- Content: occurrence and frequency of terms collected in a term-by-document matrix.
  1. Latent semantic analysis, a data reduction technique similar to principal component analysis, is used to reduce the large number of terms to a smaller number of dimensions (typically 50 or less).
  2. Latent semantic analysis relies on the singular value decomposition of the term-by-document matrix. The dimensions that are derived by the latent semantic analysis constitute the second set of features.
Concerns

1. Students might trick automated-scoring technologies and receive undeserved high scores.
   a. Student produces an essay with complex syntax and vocabulary but inaccurate content.
   - Researchers simulated the effects of gaming strategies (i.e., lexical substitutions), the student would have to be a sophisticated writer to be capable of producing a syntactically complex but factually inaccurate essay.
   - Autoscore does not use simple linear relations among elements such as features and scores that a student might attempt to manipulate. A complex process
2. Automated-scoring technologies emphasize surface features of text such as length.
   - Autoscore models the decisions of professional raters, so the extent to which surface features are emphasized is a function of the value people assign these features.
3. Automated-scoring lacks transparency in weight text features.
   - AutoScore rater scoring is subject to the same limitations.
   - Even with rubrics, human raters apply scoring criteria individually with variability between raters, and application.

Benefits

• Time: Responses can be scored in seconds with rapid feedback to educators. AutoScore will allow the immediately delivery of student test scores.
• Accuracy: More accurate than multiple professional raters.
• Reliable: Will assign the same score to a response every time.
• Consistency: The same matrix is used to score every essay.

Additional Information in NH Proposal

Machine scoring is covered in the NH Statewide Assessment System RFP and prevailing American Institutes for Research proposal in Topic 16 Machine Scored Items.