Beyond Multiple Choice: Using Automated Analysis to Evaluate Student Writing About Biology

Kevin C. Haudek[1], Rosa A. Moscarella[1], Mark Urban-Lurain[1], John Merrill[2]

[1]Division of Science and Math Education; [2]Biological Sciences Program Michigan State University, East Lansing, MI 48823

Introduction

- Constructed-response assessments reveal student thinking and conceptual barriers.
- Automated analysis allows constructed-response items for JiTT in large classes.
- Our approach to automated analysis is shown in Figure 1.

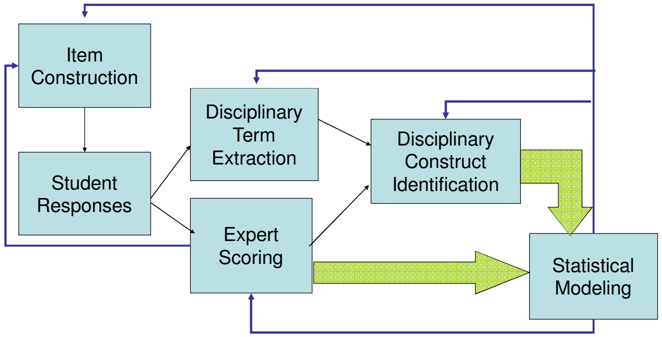


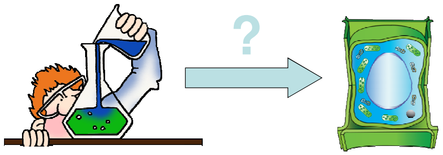
Figure 1. Workflow of item construction, analysis and statistical classification.

Sample assessment: Functional group question

Consider two small organic molecules in the cytoplasm of a cell, one with a hydroxyl group (-OH) and the other with an amino group (-NH2). Which of these small molecules (neither or both) is most likely to have an impact on the cytoplasmic pH?

A. Amino *B. Hydroxyl* *C. Both* *D. Neither*

Please explain your answer.



•Two independent human scorers rated all correct selections using 3-level rubric (see Table 1); agreement on 113 out of 129 (Cronbach alpha = .92)

Level	Number	Rubric	Example
1	41	Totally correct explanation	*Amino groups act as a base and pick up a hydrogen from its surrounding solution.*
2	14	Partially correct explanation	*The amino group acts as a base. It will lower the pH of the cytoplasm toward base (8+).*
3	58	Totally incorrect or irrelevant explanation	*Amine has two H atoms it may give up, but hydroxyl has only one OH molecule it may give up.*

Table 1. Scoring rubric used to rate student explanations. Number of correct multiple choice responses at each scoring level are indicated, along with an example student response at each level.

Word count applications fail to reveal complex concepts in responses

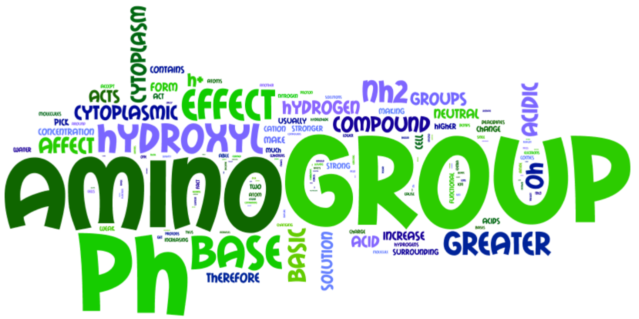


Figure 2. A word cloud of text in student responses.

Lexical analysis can categorize large number of student responses easily.

- Expert input required to customize libraries and develop categories.
- Categories contain multiple terms and rules
- Responses can be included in multiple categories.
- Output includes a variety of visualizations of responses (e.g. Figure 3).

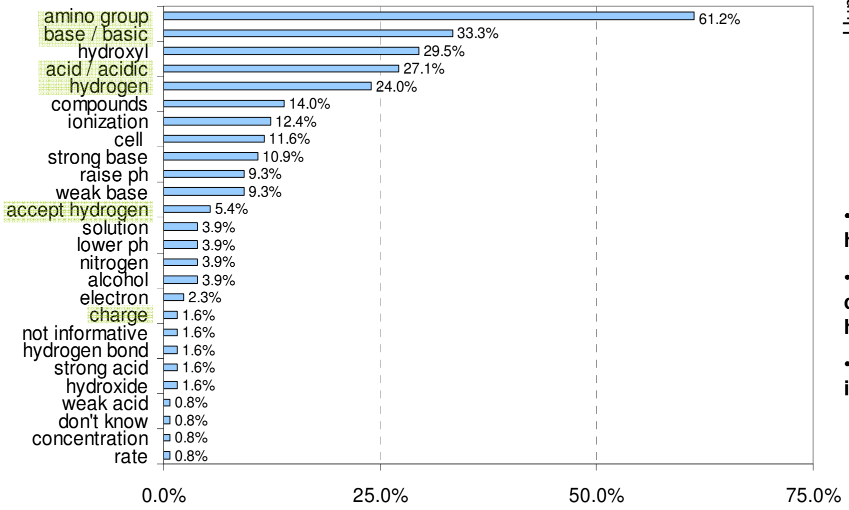


Figure 3. Distribution of responses in each category. Categories identified as significant in the scoring prediction function are highlighted in green.

Acknowledgements

We thank Ryan Sweeder for help in scoring student responses.
This work was supported by the Center for Research on College Science Teaching and Learning (CRCSTL) at Michigan State University and the National Science Foundation (NSF; awards 0736952 and 1022653). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

Discriminant analysis can create classification functions.

- Identified the most important 6 categories for predicting the human rating (see highlighted categories in Figure 3).
- Functions predict human score of student response with 77% accuracy (Table 2)
- Computer – Human Inter-rater Reliability = 0.835

		Computer Predicted Score		
		1	2	3
Human Score	1	**82.9**	12.2	4.9
	2	21.4	**42.9**	35.7
	3	6.9	12.1	**81.0**

Table 2. Classification percentages of cross-validated student responses for functional group classified at each level.

Conclusions

- **Lexical and discriminant analyses predict human scoring with 77% accuracy.**
- **Can provide formative feedback from constructed response assessments, which can help an instructor address conceptual barriers .**
- **Lexical analysis can capture complex student ideas**

For more information, people involved and other projects please visit the *Automated Analysis of Constructed Response Research Group* at:

aacr.crcstl.msu.edu

MICHIGAN STATE UNIVERSITY