Computer Scoring of Essay Questions

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

[1]Biological Sciences Program; [2]Division of Science and Math Education; Michigan State University

Biological Sciences Program
Michigan State University

Introduction

Time and resources constrain instructors from using constructed-response (essay) assessments in large undergraduate science courses, resulting in multiple choice assessments predominating. However, open-response assessments are often better at revealing student thinking and conceptual barriers. Computerized tools can analyze open-responses with a level of accuracy that gives instructors the richness of information they need to understand the learning status of their students. Using assessments about weak and strong acids, we show how combined Lexical and Discriminant Analyses are capable of predicting, with 86% accuracy, how an expert human scorer would classify student responses. This can provide a powerful tool for instructors in large courses to follow the learning status of their students at a level of detail previously not available.

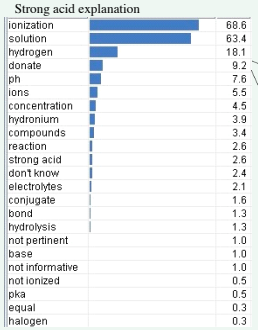
Sample Assessment: Weak and Strong Acids

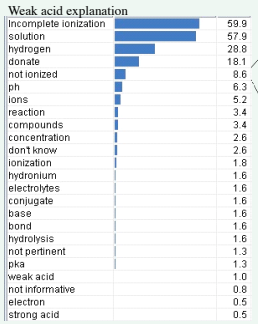
Give an explanation of a strong acid.
Give an explanation of a weak acid.

- Introductory biology, 382 student responses collected.
- Two human scorers rated 150 of 382 responses using 3-level rubric
- Scorers agreed on 140/150 responses (Cronbach Alpha for the 3 level classification = .967, p < .000)

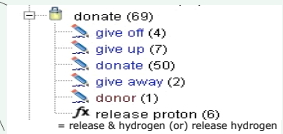
Level	#	Rubric	Examples
1	91	Correct definitions of strong and weak acids.	*"Strong acids ionize completely in solution. Weak acids only partially ionize in solution."*
2	31	Correct definition for one acid, incorrect for the other OR partially correct in both definitions.	*"Strong acids ionize completely. Weak acids do not ionize."* *"A strong acid are the acids that are more to the right of the table being mainly hydrogen ions. Weak are to the left with very few ions in solution."*
3	18	Totally incorrect / irrelevant response for both acids.	*"Strong acids have the lower pH level (1-3), the more acidic it is. Weak acids are closer to neutral but still low enough in the acidic numbers (3-6)."*

Lexical Analysis categorizes large numbers of student responses.



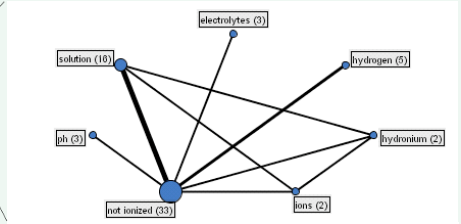


Terms are extracted and grouped in Categories.



Categories are built from related terms and can include functions.

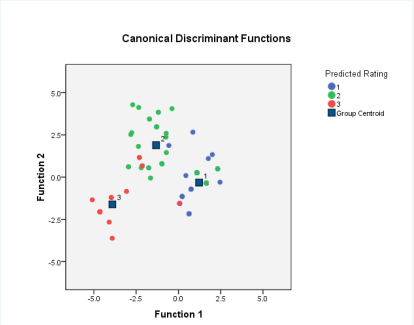
Category web shows students' conceptual connections.



A category web of all weak acid responses categorized as "not ionized". Categories that share responses with "not ionized" are depicted as nodes and shared responses between categories are indicated by connecting lines. Relative numbers of responses are shown by the size of the nodes and thickness of the lines.

Discriminant Analysis creates classification functions.

- Analyzes categories from lexical analysis.
- Only responses for which the human raters agreed were used for discriminant analysis (a total of 140 responses) .
- Identified the most important 14 categories (Table 1) for response categorization
- Predict human score of student response with 86% accuracy (Table 2)



Plot of response values determined from classification functions. Note: function 1 accounts for 73.5% of the variance.

Functions Predict Score

$$f(x) = b_1x_1 + b_2x_2 + \ldots.+ b_nx_n$$

where: b = coefficient (see Table 1) and x = term/category presence (1 or 0)

Table 1. Standardized canonical discriminant function coefficients.

	Category	Function 1	Function 2
Strong acid:	Hydrolysis	0.265	0.659
	Donate	0.444	0.454
	Solution	0.398	0.709
	Hydronium	0.143	0.538
	Hydrogen	0.247	0.311
	Conjugate	0.246	0.461
	Ionization	1.077	-0.098
	Electrolyte	0.339	0.330
Weak acid:	Incomplete ionization	0.471	0.261
	Not ionized	-0.255	0.933
	Solution	0.237	-0.263
	Ionization	-0.530	0.143
	Bond	-0.150	0.420
	Hydronium	0.056	-0.586

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

		Computer Predicted Score			
		1	2	3	Total
Human Score	1	**93.4 (85)**	6.6 (6)	0 (0)	100 (91)
	2	25.8 (8)	**67.7 (21)**	6.5 (2)	100 (31)
	3	5.6 (1)	11.1 (2)	**83.3 (15)**	100 (18)

Conclusions

- **Combined Lexical and Discriminant Analyses can predict human scoring with 86% accuracy.**
- **Computer scoring can provide rich formative feedback from extended response assessments.**

Lexical analysis can return an accurate picture of the words, phrases and ideas used by students in their responses, while discriminant analysis estimates the significance of these terms in predicting how an expert scorer would evaluate each response. When provided with such an analysis, an instructor can better direct instructional intervention to prevalent and important conceptual barriers.

References

Moscarella, R. A., Urban-Lurain, M., Merrill, J. E., Richmond, G., Patterson, R.,Parker, J. M., Merritt, B. W., Wilson, C. D., Long, T. M. (2008, March 30 - April 2). Understanding undergraduate students' conceptions in science: Using lexical analysis software to analyze students' constructed responses in biology. Proceedings, NARST 2008 Annual International Conference, Baltimore, MD.

Richmond, G., Urban-Lurain, M., Parker, J., Merrill, J., & Merritt, B. (2008, March 30 - April 2). Assessment-informed instructional design to support model-based reasoning in college-level biology. Proceedings, NARST 2008 Annual International Conference, Baltimore, MD

Acknowledgements: This work is supported by the Center for Research on College Science Teaching and Learning (CRCSTL) at Michigan State University and the National Science Foundation (award 0736952). 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 National Science Foundation.