Exploring Computerized Lexical Analysis to Predict Calibrated Peer Review Ratings of Student Writing in Chemistry

Kevin C. Haudek
Mark Urban-Lurain
Michigan State University

Arlene A. Russell
University of California – Los Angeles

ceer Center for Engineering Education Research @ Michigan State University

MICHIGAN STATE UNIVERSITY

Acknowledgements

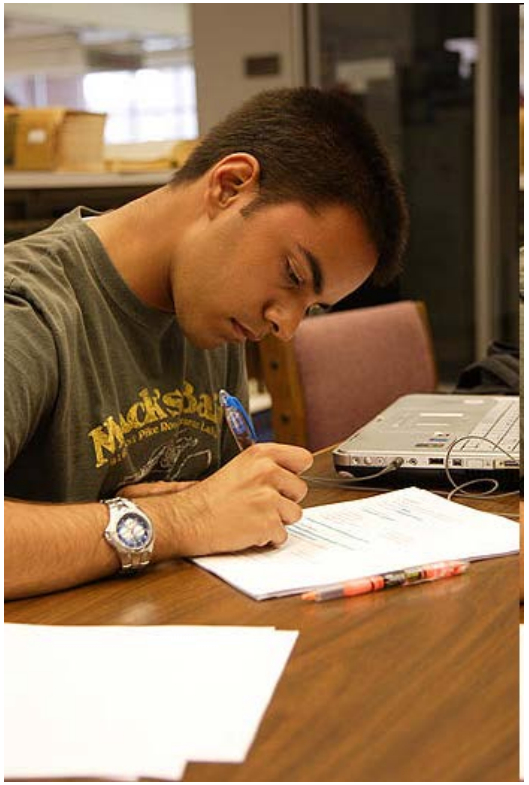
- Automated Analysis of Constructed Response Research Group (AACR)
  - Michigan State University
  - University of California – LA
  - SUNY – Stony Brook
  - University of Colorado – Boulder
  - University of Maine
  - University of Georgia
  - Western Michigan University

This work was supported by the National Science Foundation (Grants DUE 1022653, DUE 1143642 ,DUE 04-42828, DUE 08-37229 and DUE 08-16660). 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.

Overview

- Paper on CD
- Background
  - Computerized text analysis
  - Calibrated Peer Review
- Research Questions
- Methods
- Results
- Future directions

Assessment to Reveal Student Thinking



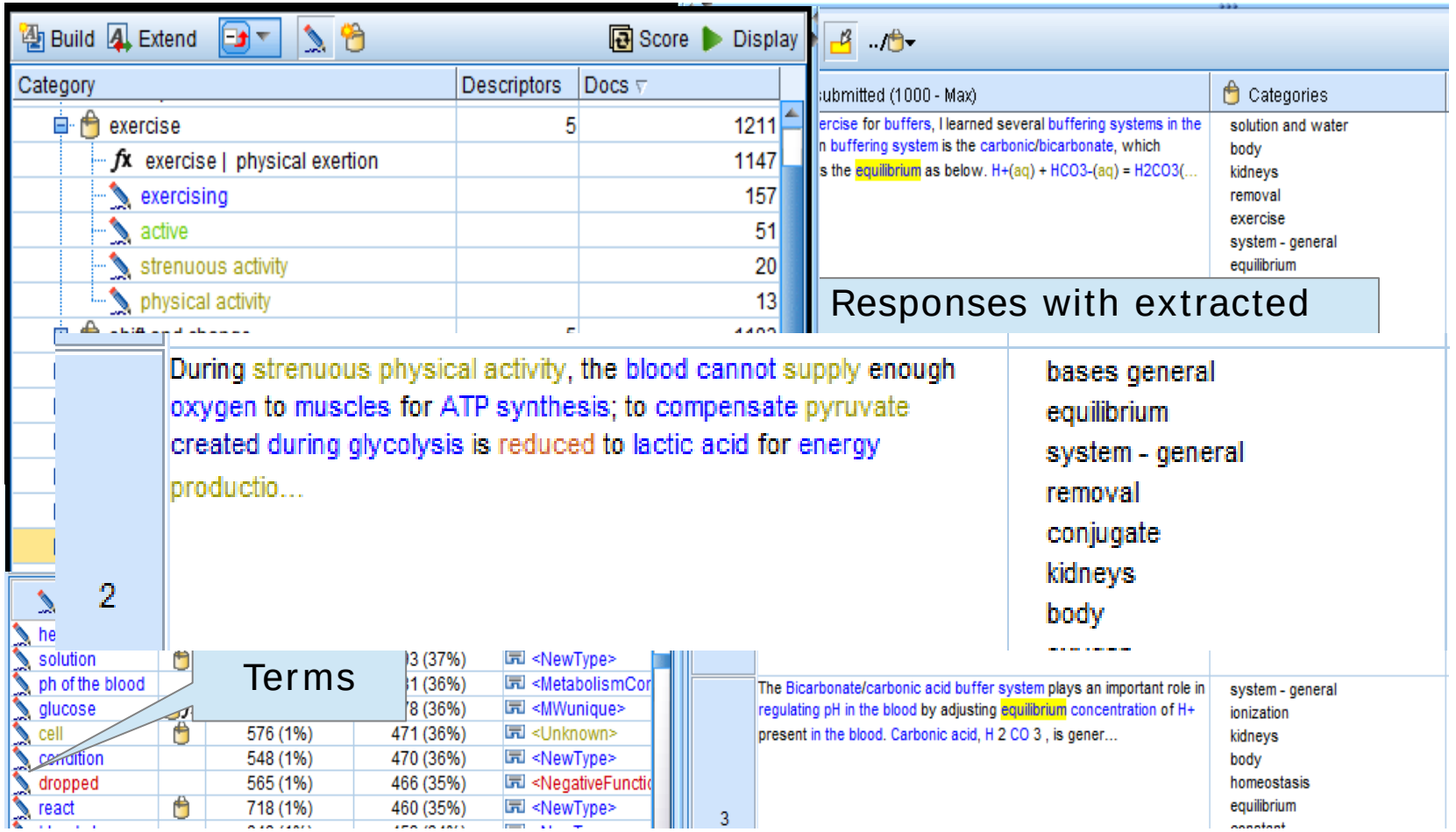
Background

- Encouraging writing in STEM courses
  - Authentic task for scientists
  - Students must actively construct explanations
  - Better reveals student understanding
- Major barrier: evaluating large number of student responses
- Two approaches:
  - Computerized text analysis
  - Calibrated peer review (CPR

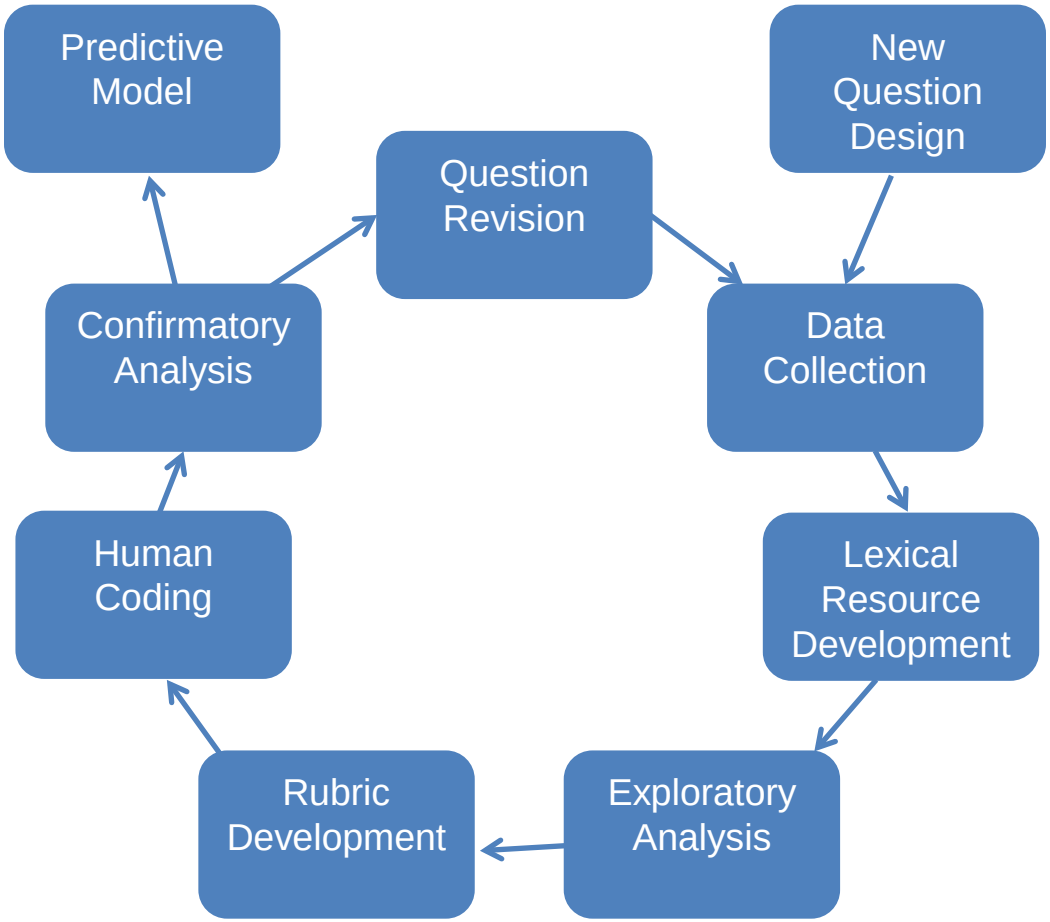
Computerized Text Analysis

- Software extracts and categorizes terms and phrases in electronic text
  - *Categories* can be pre-defined, based on expert answer, or emergent, through student writing
- Lexical categories used as independent variables to predict expert scoring
  - Scoring models approach human-human IRR
  - Requires large numbers of human scored responses
- More information:
  - www.msu.edu/~aacr/
  - Haudek et al. (2012) What are they thinking? Automated analysis of student writing about acid/base chemistry in introductory biology. *CBE-Life Sci Educ*, 11.

IBM SPSS Text Analytics



AACR Question Development Cycle



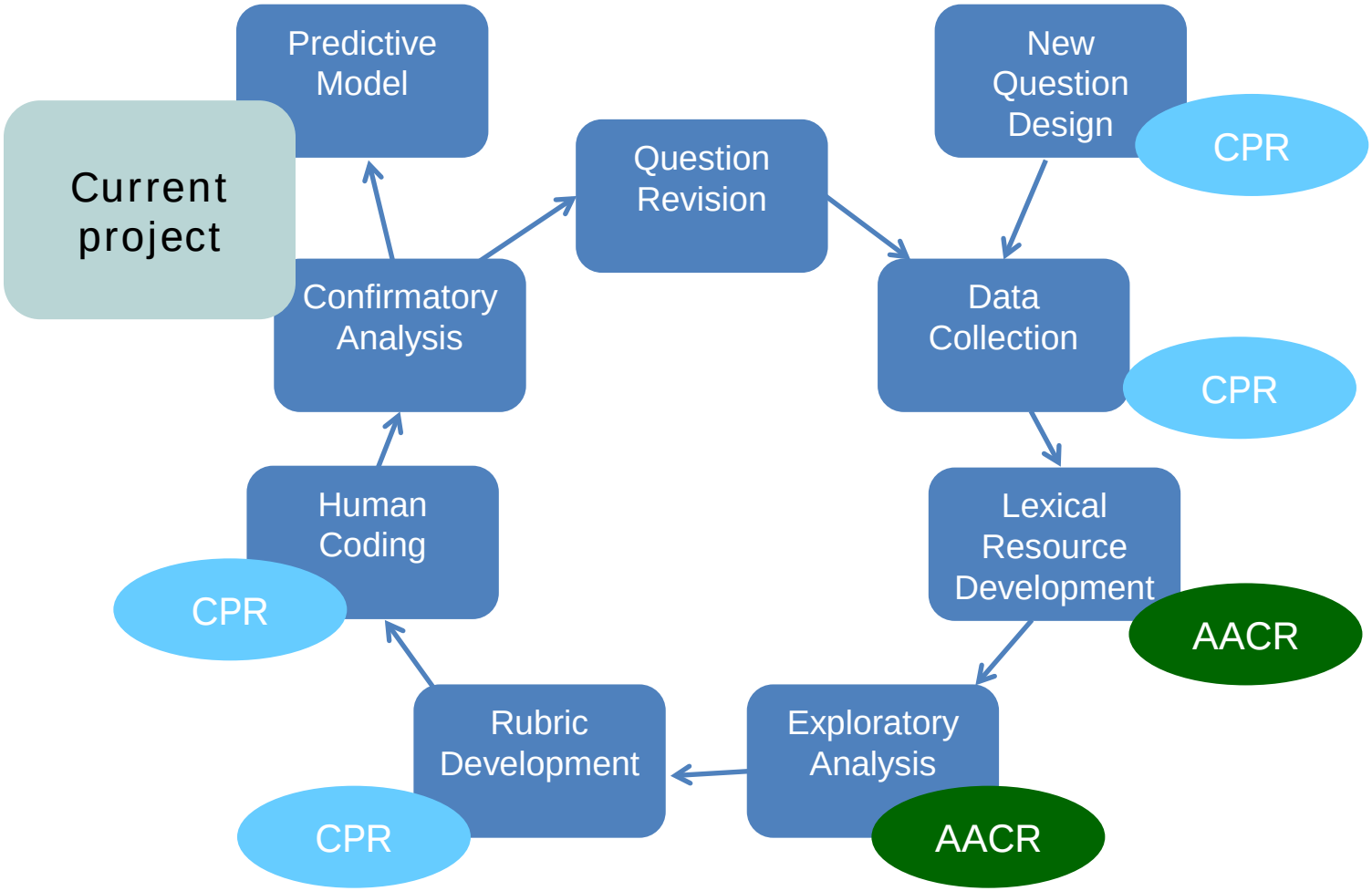
Calibrated Peer Review (CPR)

- Each student
  - Submits an essay
  - Peer Reviews 3 other essays
- Reviewer Calibration Index (RCI)
  - Each student uses rubric to score 3 known essays
  - Overall score weighted-average of peer scores
- More information:
  - http://cpr.molsci.ucla.edu
  - Russell et al. (1998) Molecular Science: Network-Deliverable Curricula. *J. Chem. Ed.* 75

Research Questions

- How well do previously developed automated analysis techniques transfer and predict trained peer scoring of longer essays?
  - Previous resources developed for shorter responses about general chemistry
- What improvements can be made in scoring models by including additional information about the text?
  - Limited semantic information and writing-quality evaluation

AACR Question Development Cycle



Methods: CPR Assignment

- Assignment
  - Write about bicarbonate buffering capabilities in the bloodstream
    - Given learning goals, source materials and seven “guiding prompts” to aid writing
      - *What other pathways does the body use to remove excess protons, carbon dioxide or bicarbonate from the blood?*
- Reviewing
  - Rubric
    - 2 Writing quality items: topic sentence and free of errors
    - 9 Content items:
      - *Is the removal of bicarbonate ion through the kidneys discussed?*
  - Assign overall score: 1 - 10

Methods: CPR Writing Prompt

Write a paragraph of the required length on how the bicarbonate system maintains homeostasis in the blood system. In your paragraph be sure to address the issues raised in the "Guidance for Writing Your Text". Remember that you are writing a paragraph, not just answering a list of questions, so you should include a topic sentence and a summary statement.

Your text should also include both a topic sentence and a summary sentence. A topic sentence should introduce the concepts and ideas relevant to the rest of the essay. A summary sentence should summarize the ideas and concepts discussed in the essay.

Methods: Text Analysis

- Applied *term libraries* developed for general chemistry items
- *Word proximity* program to search for common word pairs
- Modeling peer scoring
  - Lexical analysis and word proximity variables in step-wise regression

Results: Response Characteristics

- 2nd semester general chemistry students
  - N = 388
- Mean response character length
  - 1681 ± 91 (std. dev)
- Mean CPR weighted score
  - 6.2 ± 1.6
- Total number of lexical categories = 119
  - 4 of which were used by every response
  - Some categories used infrequently (e.g. ATP, temperature)
- Mean categories / response
  - 43 (range 31-55)

Results: Predictive Scoring Model Using Lexical Categories

- 18 lexical categories
- $R^2 = 0.465$

Variable	Beta	Variable	Beta
Character count	0.134	energy	0.106
increase pH	0.268	glucose	0.103
ratio	0.267	LeChatelier's principle	0.086
bases - general	0.236	strong acid	-0.094
strong	0.166	Environment	-0.107
constant	0.160	Help	-0.124
water	0.154	acids – general	-0.139
glucose breakdown	0.131	ions	-0.152
kidneys	0.124	extracellular	-0.153

Results: Predictive Scoring Model Using Lexical Categories

- 18 lexical categories
- $R^2 = 0.465$

Variable	Beta	Variable	Beta
Character count	0.134	energy	0.106
increase pH	0.268	glucose	0.103
ratio	0.267	LeChatelier's principle	0.086
bases - general	0.236	strong acid	-0.094
strong	0.166	Environment	-0.107
constant	0.160	Help	-0.124
water	0.154	acids – general	-0.139
glucose breakdown	0.131	ions	-0.152
kidneys	0.124	extracellular	-0.153

Results: Predictive Scoring Model Using Lexical Categories

- 18 lexical categories
- $R^2 = 0.465$

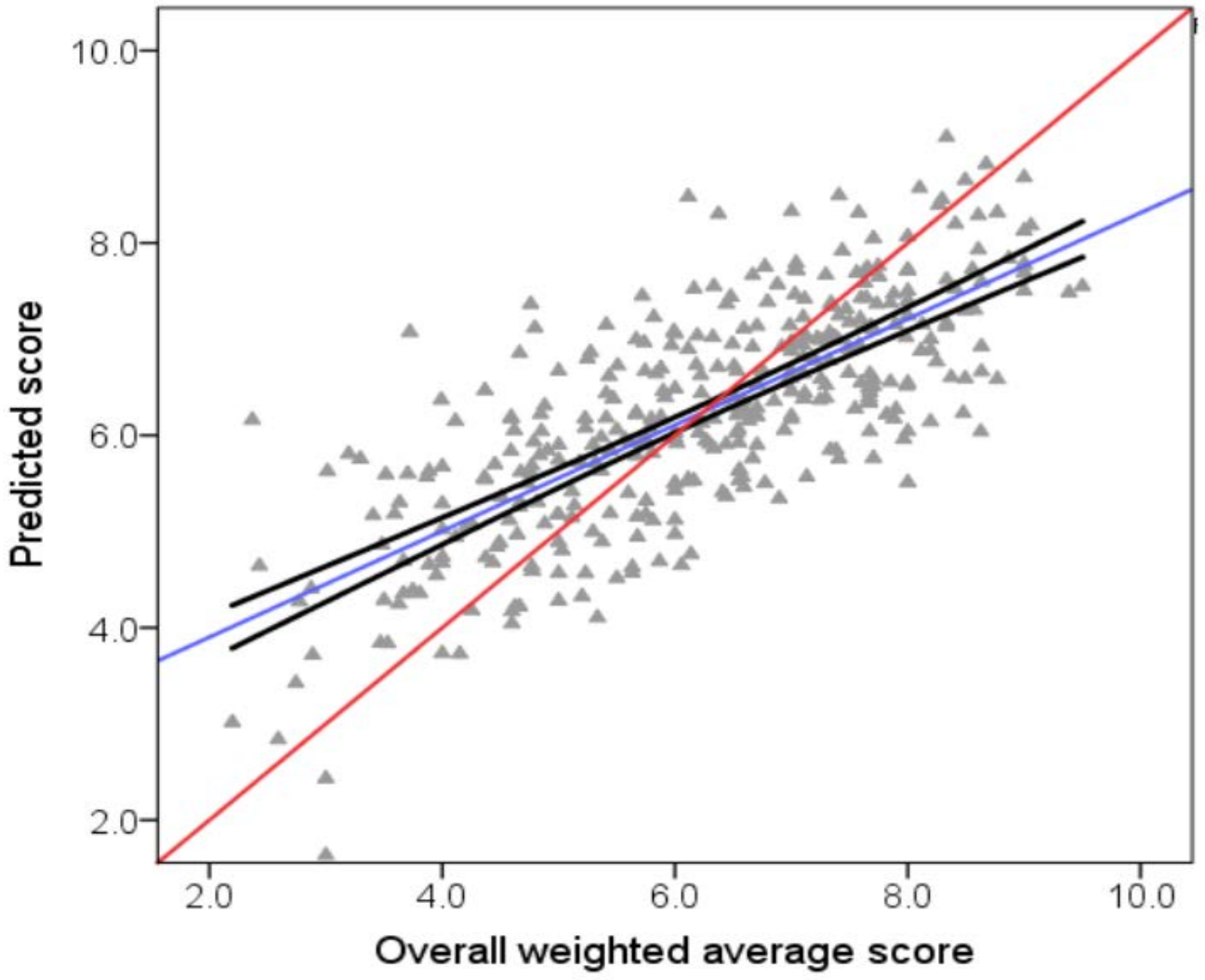
Variable	Beta	Variable	Beta
Character count	0.134	energy	0.106
increase pH	0.268	glucose	0.103
Ratio	0.267	LeChatelier's principle	0.086
bases – general	0.236	strong acid	-0.094
Strong	0.166	Environment	-0.107
Constant	0.160	Help	-0.124
water	0.154	acids – general	-0.139
glucose breakdown	0.131	ions	-0.152
kidneys	0.124	extracellular	-0.153

Results: Improving Scoring Using Writing Quality Scores

- Added writing quality rubric items
- 24 variables chosen
  - 17 shared with initial model
- $R^2 = 0.551$

Variable	Beta	Variable	Beta
Rubric – Free of errors	0.233	exchange	0.082
Rubric – Topic sentence	0.123	carbonate ion	0.075
Character count	0.154	named chemical compound	-0.081
increase pH	0.254	function	-0.082
ratio	0.236	phosphate buffer	-0.084
bases - general	0.202	strong acid	-0.097
water	0.169	lungs	-0.101
glucose	0.132	transport	-0.113
strong	0.128	help	-0.123
glucose breakdown	0.127	Ions	-0.124
constant	0.120	acids - general	-0.130
kidneys	0.109	extracellular	-0.144

Results: Improving Scoring Using Writing Quality Scores



Conclusions

- Existing lexical resources applied in new context
  - Additional semantic information improves model
- Successfully analyzed long, complex, scientific essays
- Writing quality important consideration in peer review
- Model best at predicting “average responses”
  - Need additional data to improve accuracy at ends

Future Directions

- Balancing generalizable scoring models with instructor calibration
  - Unique calibration and evaluation rubrics limits predictive model's transferability
- Exploring how students use rubrics
  - Which factors effect how student assign overall scores?

Questions

Center for Engineering Education Research @ Michigan State University

- Mark Urban-Lurain
  - urban@msu.edu
- Automated Analysis of Constructed Response research group (AACR)
  - www.msu.edu/~aacr
  - Related talk: Tuesday, 8:30 – 10:00, Strand 5
    - Michele Weston: *Comparing Formative Feedback Reports: Human and Machine Analysis of Constructed Response Questions in Biology*
- Calibrated Peer Review (CPR)
  - cpr.molsci.ucla.edu