

# Developmental Approach to Computer-assisted Scoring of Student Essays

John Kremer, Instructor

Paul Hancock, Programmer

Heather Richards, Assistant

# Goal of Project is to ...

gradually develop and enhance a cost-efficient computer essay grading program that will work jointly with human essay graders & course coordinator to reliably assess student essays and provide students with feedback on the quality of their essays

# Context

- Want to use essays to evaluate critical thinking skills (Bloom) :
  - Comprehension
  - Application
  - Analysis
- Large gateway course
- Introductory psychology, B104

# Grading Demands

- Five exams (5)
- Each exam may be taken twice (5x2)
- Two essays per exam (5x2x2)
- 3000 students (3000x5x2x2)
- Potentially 60,000 essays per year
- Actually 30,000

# Solution One

## Professors Grade Exam

- Advantages
  - Great feedback for students
  - Great feedback for professors
- Disadvantages
  - Time consuming task
  - Repetitive, often boring
  - Low professional payoff
  - High cost

# Solution 2

## Student Raters

- Advantages
  - Moderate cost, less than professors
  - Raters can be trained if essays are constant
- Disadvantages
  - Moderate cost (.1 hrs x 30,000 x \$10/hr)
  - Reliability not consistent
  - High turnover
  - Giving feedback raises the cost
  - Raters taken from aiding student learning

# Solution 3

## Computer Graded

- Advantages
  - Low cost after development
  - Reliable scoring
- Disadvantages
  - High development cost
  - Not valid (from our experience) for all essays
  - Negative perception by some faculty

# Solution 4

## Computer-Assisted Grading

- **Computer** scores appropriate essays
- **Students raters** score others
- Raters and computer grade some essays to check reliability & validity
- Students can appeal any grading with a potential penalty
- All essays scored within 12 hours
- **Professor** writes essay & criteria

# Computer Solution 1: Purchase Existing Program

- Some not available to the university
- Others are expensive but
  - Low input into program functioning
  - Low control over implementation
- A couple are not expensive
  - Do permit input into program functioning but
  - Require technological sophistication

# Computer Solution 2: Developmental Approach

- Start simple
- Stay cost-efficient
- Develop program as time and resources permit

# Computer Solution 2: Developmental Approach

- Requirements
  - One interested faculty member
  - One experienced, part-time programmer
    - Database management – SQL
    - Information retrieval
    - Interactive web-based experience
    - Algorithm intensive
  - Student assistant

# Stage 1

## Simplify Computer's Task

- Focus on content, not style
- Professor develops scoring criteria
- Computer only scores essay
- Using key words
- Three Phases
  - Software development of the program,
  - Specification of essay criteria, and
  - Validation

# Phase 1

## Software Development

- Input of criteria – Auto Essay Grader
  - Rapid development tool from Ironspeed.com
- Running the “search” engine
  - Stop words: SIL International (Summer Institute of Linguistics): [www.sil.org](http://www.sil.org)
  - Stemming (Porter Stemmer, 1980)
    - [http://www.dcs.gla.ac.uk/idom/ir\\_resources/linguistic\\_utils/](http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/)
  - (Students required to spell correctly)
- Output of the results

# Phase 2

## Essay Criteria

- Professor writes essay and criteria used by computer and human raters
- Criteria: Single word, joined word group (and), synonym word group (or)
- Assistant changes criteria after
  - Comparing computer scores with raters
  - Reading essays to find holes
  - Examining the frequency distribution of words used by all students in the essay

# Phase 3 Validation

- Sample
  - 44 different essays
  - 200 student responses per essay
  - Essays scored (4 point scale)  
by computer (**C**) & raters (**R**) (2-3)
  - 30,000 ratings across all essays
- Descriptive Statistics
  - Ave Score: 2.04 (C) vs. 1.91 (R)
  - Computer higher on 2 out of 3 essays

# Validity

## All Essays

- Comparison of Cs & ave. of Rs' scores
  - 25% identical,
  - 68% < one point difference
- Interrater (Rs only) reliability = .74
- Concurrent validity (C to ave Rs) = .70
- Using data from 1<sup>st</sup> round, assistant increased validity for 80% of the essays

# Validity

## Individual Essays

- Concurrent validity: (C to ave Rs)
  - 25% above .8,
  - 55% above .7,
  - 70% above .6, and only
  - 10% below .5

# Validity

## Individual Essays

- Difference b/w concurrent validity & reliability
  - 65% both high & similar ( $< .10$ ) requiring no change in criteria nor essays
  - 6% validity  $>$  reliability indicating that the human graders were not using the same criteria.
  - 29% reliability  $>$  validity coefficient indicating that the criteria needed to be changed or the computer was not capable of grading that particular essay.
  - 2% both low ( $< .5$ ) and similar indicating an essay that probably needed to be rewritten.

# Subsequent Stages

- Stage 2: Add feedback from textbook to essays
- Stage 3: Add spellchecker (server based) & thesaurus
- Stage 4: Add criteria development function to computer (Bayesian approach)
- Stage 5: Add another development function (Latent Semantic Analysis)

# Summary

- A simple computer program using key words can reliably and validly
  - Score most essays (at least 70%)
  - Given a narrow range of content with
  - Iterative input from professor & assistant
- And when combined with student raters, provides a valid & cost efficient approach to scoring essays
- And provides opportunities for further development.

# Resources

- Description

- Overview -

- <http://pareonline.net/getvn.asp?v=7&n=26>

- <http://jite.org/documents/Vol2/v2p319-330-30.pdf>

- Peg & IEA (LSA) – George Chung, UCLA

- <http://www.cse.ucla.edu/Reports/TECH461.pdf>

- Betsy – Rudner

- [http://edres.org/betsy/bayesian\\_ov.htm](http://edres.org/betsy/bayesian_ov.htm)