Application of the Many-facets Rasch Measurement (MFRM) to Supporting Self-directed Learning in Writing Assessment of Undergraduates

Kinnie Kin Yee CHAN
The Open University of Hong Kong

Purpose of the research

- **Investigate the effectiveness of an Automated Essay Scoring (AES) system, in measuring essay writing ability as scored by human raters**

SO:

1. Implement the MFRM to detect and adjust for rater effects in human essay scoring

2. Apply Rasch model to co-calibrate the scales of an AES system and human raters for scoring essay writing

Research Questions

Does the AES system perform well in assessing students' English essay writing?

the extent to which the Rasch model can be used to develop:

i. **a scale for human essay scoring on the students' essays?**
ii. **a comparative scale for the AES system calibrated against the students' MFRM human ratings?**

Background

- **Automated Essay Scoring (AES) systems adopt computer technology to evaluate and score written prose in the place of the usual human scoring.**

- **Most developed to score written work in English**

Background

- **In the early 1960s, Ellis Page developed the first AES, Project Essay Grade (PEG), to make large scale essay scoring practical and effective (Page, 2003; Rudner & Gagne, 2001).**

- **GMAT**

  http://www.mba.com/global/store/store-catalog/gmat-preparation/gmat-write.aspx

  http://www.vantageonlinestore.com/home.php?cat=299

  http://www.cengage.com/writeexperience/

Background

- **Automated Essay Scoring (AES) systems**

  - numerous research revealed strong evidence

  - demonstrates well correlation with human rater behaviour (Shermis & Burstein, 2003)

Background

- **The scores of PEG showed a correlation with human raters' scores of .71, but human raters scores correlated with each other's at .62 (Page, Poggio, and Keith, 1997).**

- **Foltz, Laham, and Landauer (1999) reported a study of essay writing by American university students which revealed a correlation of .8 between IEA and human raters.**

- **The Pearson *r* correlations of agreement between human raters and the IntelliMetric averaged .83, with a range of .67 to .94 (Rudner, Garcia and Welch, 2006)**

Background

- **Some debate on automated essay scoring:**
- **"Computers can't tell the difference between reasonable prose and nonsense idea."**

  ~ David Ravitch, Research Professor at NYU
- **"It can't tell you if you've made a good argument, or if you've made a good conclusion."**

  ~ Mark Shermis, University of Akron

- **Professionals against machine scoring of student essays in high-stakes assessment:**

http://humanreaders.org/petition/index.php

Methodology

Participants

- **More than 100 undergraduates**
- **Lee Shau Kee School of Business & Administration at the Open University of Hong Kong**
- **Students responded to 4 prompts**

Data collection

- **Part of the undergraduates' assessment in the course**

Methodology

Raters

- **4 English tutors who teach the course as human raters**
- **AES system**

Scoring

- **Each essay will be scored by 2 raters**
- **Holistic rubrics follow the testing design**
- **An AES system measure**

Methodology

Analysis

- **Winsteps: The Rasch Model**
- **FACETS: The Many-facets Rasch Measurement (MFRM)**


THANK
YOU