

Improving machine-learning scoring models of written student response assessments based on predictive accuracy scores

Nicholas Yoshida, Megan Shiroda, Kevin Haudek

Automated Analysis of Constructed Response (AACR) Research Group, Michigan State University



CREATE for STEM Institute

WHY IS THIS IMPORTANT?

- Constructed response questions reveals student thinking better than traditional multiple-choice
- Professors are limited by large class-size and high time requirement for grading
- Computer Scoring Models (CSMs) automate scoring, making constructed response more accessible.

RESEARCH PROBLEM

- Creating CSMs is time consuming.
- Require Input of human assigned scores and each model has unique issues
- Little is known about how to improve CSMs for text assessments
- Currently new responses are added until the model works.

RESEARCH QUESTION

Can predictive accuracy scores be used to efficiently improve CSM performance? We hypothesized that using subsets of responses with different predictive accuracies would lead to different CSM performance.

CONCLUSIONS

- CSMs improve more when trained with lower probability responses and higher number of CRs.
- Train CSMs by focusing on responses the computer is less confident about.

FUTURE DIRECTIONS

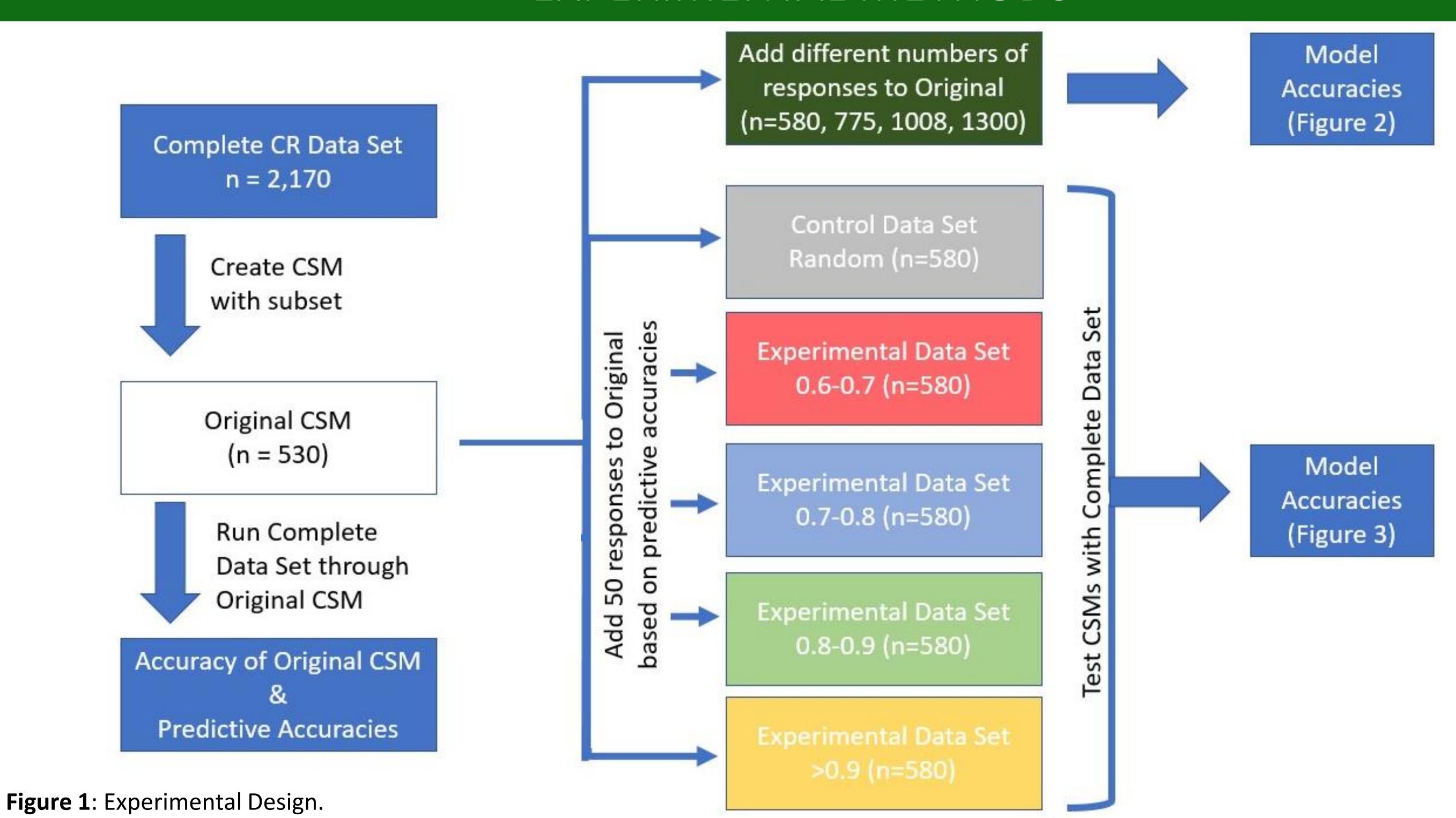
- Test these conclusion on other data sets and CSMs
- Since models are unique these finding may not always apply.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation (DUE grants 1323162 & 1660643). 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 supporting agencies.



EXPERIMENTAL METHODS



PRELIMINARY FINDINGS

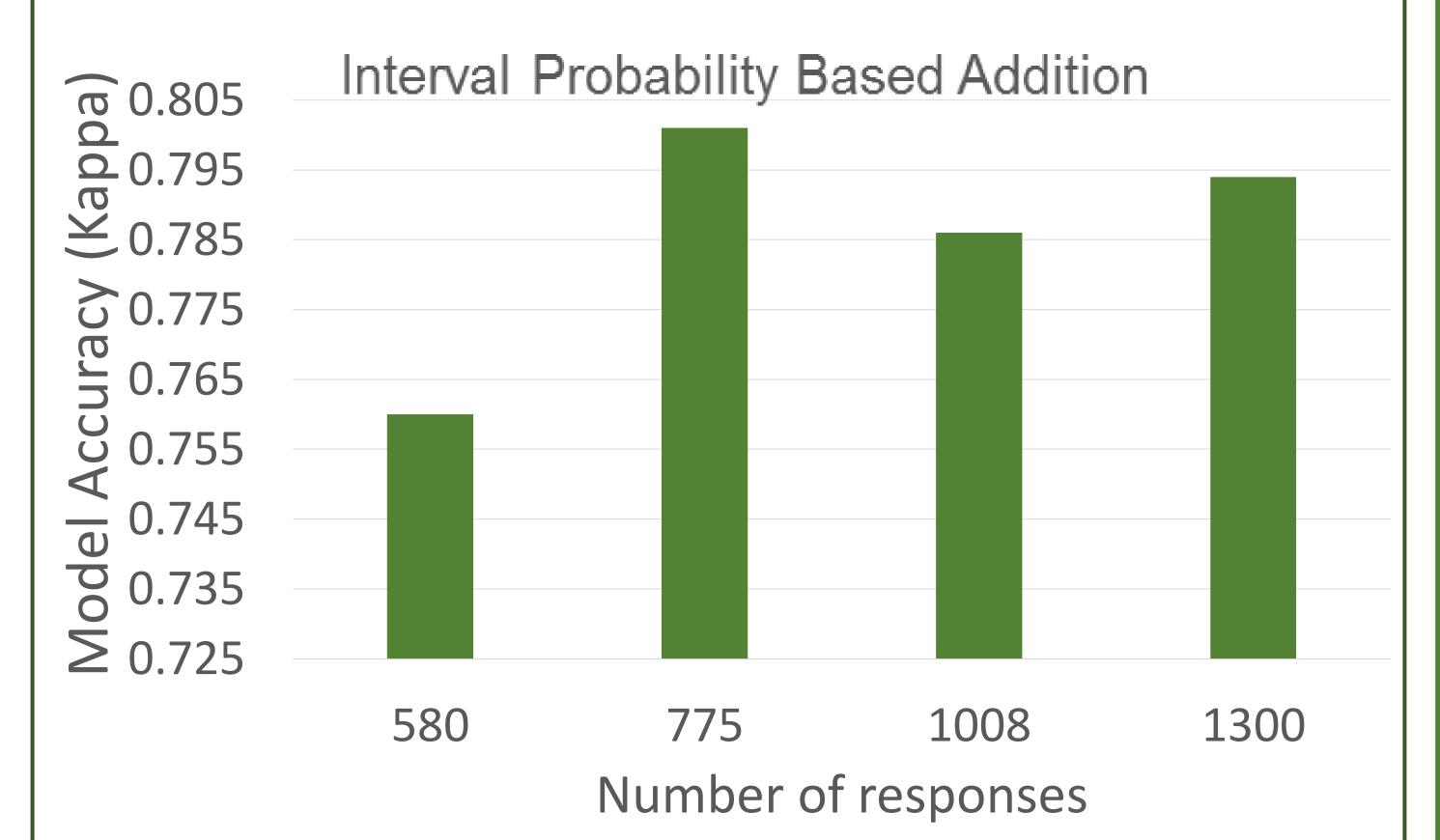


Figure 2: Lower probability responses showed the most improvement. As response probability grew the model improvement decreased.

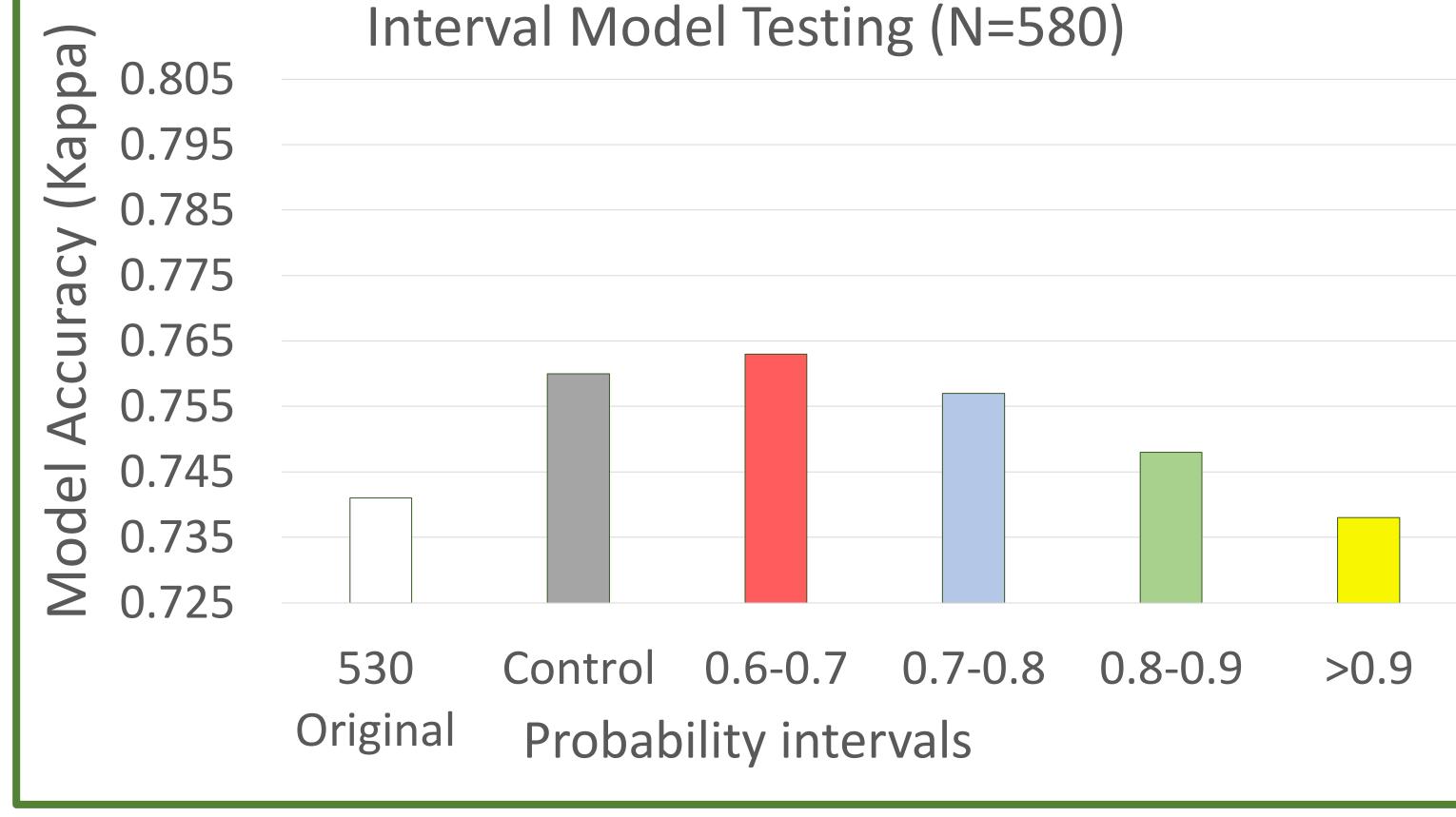


Figure 3: Higher number of responses had the biggest improvement. Different interval probabilities with varying number of responses added.





Landis, J., & Koch, G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159-174. doi:10.2307/2529310