# FOUNDATIONS FOR AI-ASSISTED FORMATIVE ASSESSMENT FEEDBACK FOR SHORT-ANSWER TASKS IN LARGE-ENROLLMENT CLASSES

**Authors:** Susan E Lloyd, Matthew D Beckman, Dennis K Pearl, Rebecca J Passoneau, Zhaohui Li, Zekun Wang

**Institution:** The Pennsylvania State University

# TAKE-HOME MESSAGE ABOUT THE STUDY

Motivation

- "Write-to-learn" tasks improve learning outcomes
- Feasibility of effective formative assessment in large classes
- Algorithms can assist this aim

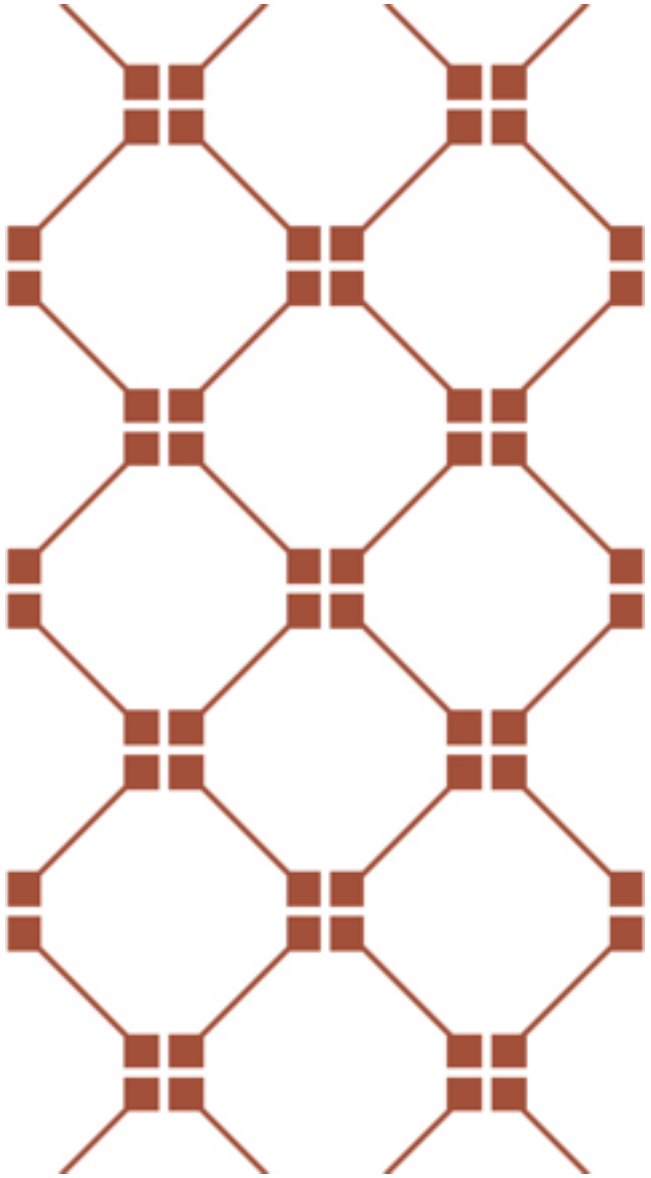
Methods & Results

- 6 short-answer tasks completed by 1,935 students
- Scored by several human raters and an algorithm
- Substantial inter-rater agreement (QWK > 0.74 for each pair; Fleiss' kappa = 0.68 for group)
- High intra-rater agreement (QWK = 0.88)

Implications

- Pilot cluster analysis for scalable formative assessment
- Instructors of all class sizes would benefit

Lloyd, S.E., Beckman, M.D., Pearl, D.K., Passonneau, R.J., Li, Z., & Wang, Z. (in review). Foundations of NLP-assisted formative assessment feedback for short-answer tasks in large enrollment statistics classes. Preprint URL: http://arxiv.org/abs/2205.02829



**RQ1: What level of agreement is achieved among trained human raters labeling (i.e., scoring) short-answer tasks?**

**RQ2: What level of agreement is achieved between human raters and an NLP algorithm?**

RQ3: What sort of NLP representation leads to good clustering performance, and how does that interact with the classification algorithm?

# RESEARCH QUESTIONS OF INTEREST

# METHODS

- Data: 1,935 students completed 6 short-answer tasks about statistical inference as part of a prior study (see Beckman, 2015)
- Responses were divided among 4 human raters with sufficient intersection to evaluate inter-rater agreement
- A subset of student responses scored in 2015 by one of the raters were again evaluated by this same rater to evaluate intra-rater agreement
- An algorithm scored a subset of student responses for correctness

# RESULTS

- Substantial inter-rater agreement among human raters A, C, D
- Almost perfect intra-rater agreement for rater A
- Similar calculations performed with algorithm as an additional rater
- Substantial inter-rater agreement among algorithm and human raters

| Rater Comparison | Measure of Reliability |
|---|---|
| Rater A & Rater C | QWK = 0.83 |
| Rater A & Rater D | QWK = 0.80 |
| Rater C & Rater D | QWK = 0.79 |
| Rater A (2015) & Rater A | QWK = 0.88 |
| Rater A & Rater C & Rater D | Fleiss' Kappa = 0.698 |
| Rater A & SFRN | QWK = 0.79 |
| Rater C & SFRN | QWK = 0.82 |
| Rater D & SFRN | QWK = 0.74 |
| Rater A & Rater C & Rater D & SFRN | Fleiss' Kappa = 0.678 |

Table 1: Reliability comparisons among human raters (A, C, D) and an NLP algorithm (SFRN).

# LIMITATIONS & FUTURE WORK

Limitations

- Students come from classes of varying sizes (not a single large class)

Future Work

- Test complete prototype on large enrollment class data
- Manage tradeoff between classification of correctness and density of clusters
- Investigate semantic meaning manually to derive a process for algorithm clustering
- Continued research toward large class formative assessment that approaches small class quality and instructor burden