

Foundations for AI-assisted formative assessment feedback for short-answer tasks in large-enrollment classes

Matthew Beckman
Penn State University

11th International Conference on Teaching Statistics (ICOTS)

Rosario, Argentina
September 13, 2022

Motivation

- Formative assessment benefits both students & instructors (GAISE, 2016; Pearl, et al., 2012)
- “Write-to-learn” tasks improve learning outcomes (Graham, et al., 2020)
- Critical for citizen-statisticians to communicate statistical ideas effectively (Gould, 2010)
- Continual practice with communicating improves statistical literacy and promotes retention (Basu, et al., 2013)
- Human-machine collaboration is a promising mechanism to assist rapid, individualized feedback at scale (Basu, 2013)
- NLP-assisted feedback has previously been studied for essays or long-answer tasks (see e.g., Attali, et al., 2008; Page, 1994)

Research Questions

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

Preprint

Susan Lloyd, Matthew Beckman, Dennis Pearl, Rebecca Passonneau, Zhaohui Li, & Zekun Wang (accepted). Foundations of NLP-assisted formative assessment feedback for short-answer tasks in large enrollment statistics classes. Preprint URL: <http://arxiv.org/abs/2205.02829>

Spoilers?!

- **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?

Spoilers?!

- RQ1: substantial inter-rater & intra-rater agreement
- RQ2: substantial agreement among human & NLP labeling
- RQ3: work in progress, but promising

Methods (Sample)


Study utilized de-identified extant data & scoring rubrics from an unrelated previous study (Beckman, 2015)

- 6 short-answer tasks
- 1,935 students total
- 29 class sections 15 distinct institutions

Methods (Short-answer task)

4. Walleye is a popular type of freshwater fish native to Canada and the Northern United States. Walleye fishing takes much more than luck; better fishermen consistently catch larger fish using knowledge about proper bait, water currents, geographic features, feeding patterns of the fish, and more. Mark and his brother Dan went on a two-week fishing trip together to determine who the better Walleye fisherman is. Each brother had his own boat and similar equipment so they could each fish in different locations and move freely throughout the area. They recorded the length of each fish that was caught during the trip, in order to find out which one of them catches larger Walleye on average.

a. Should statistical inference be used to determine whether Mark or Dan is a better Walleye fisherman? Explain why statistical inference should or should not be used in this scenario.



b. Next, explain how you would determine whether Mark or Dan is a better Walleye fisherman using the data from the fishing trip. *(Be sure to give enough detail that a classmate could easily understand your approach, and how he or she would interpret the result in the context of the problem.)*

Figure 1: Sample task including a stem and two short-answer prompts.

Methods (Humans)

- 3 human raters typical of large-enrollment instruction team
- responses allocated such that 63 student responses in common for each combination of raters to quantify agreement
- only constraint: sufficient data for intra-rater analysis for person that had labeled a previous set of 178 responses 6 years prior

Methods (NLP)

The set of task-responses were randomly split four ways:

- 90% were split into the typical division of training (72%), development (9%) and test (9%)
- 10% held in reserve for more rigorous testing

Two NLP algorithms were compared for accuracy using a subset of student responses (Li et al., 2021).

- LSTM: a logistic regression combined with a Long Short-Term Memory for learning vector representations
- SFRN: Semantic Feature-Wise Transformation Relation Network

Results (RQ1)

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

Comparison	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 & 2021	QWK = 0.88
Raters A, C, & D	FK = 0.70

Reliability interpretation¹: 0.6 < substantial < 0.8 < near perfect < 1.0

¹Viera & Garrett (2005)

Results (RQ2)

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

The SFRN algorithm achieved much higher classification accuracy than LSTM (83% vs. 72%)². Human & SFRN agreement:

Comparison	Reliability
Rater A & SFRN	QWK = 0.79
Rater C & SFRN	QWK = 0.82
Rater D & SFRN	QWK = 0.74
Raters: A, C, D, & SFRN	FK = 0.68

Reliability interpretation³: $0.6 < \text{substantial} < 0.8 < \text{near perfect} < 1.0$

²SFRN & LSTM comparison excludes instances when human labels disagree

³Viera & Garrett (2005)

Results (RQ3)

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

- SFRN learns a high-dimension ($D = 512$) vector representation on training data.
- Experiments with K-means and K-medoids clustering showed SFRN produce more consistent clusters when retrained (0.62), in comparison to other classifiers.⁴
- Highest consistency (0.88; $D = 50$), however, was achieved using a matrix factorization method that produces static representations (WTMF; Guo & Diab, 2011)

⁴Consistency is measured as the ratio of all pairs of responses in a given class per question that are clustered the same way on two runs (in the same cluster, or not in the same cluster).

Discussion

- **RQ1:** Substantial agreement achieved among trained human raters provides context for further comparisons
- **RQ2:** NLP algorithm produced agreement reasonably aligned to results achieved by pairs/groups of trained human raters
- **RQ3:** Classification and clustering have competing incentives for dimensionality; Low D is better for cluster stability, High D better for classification reliability.

Future work:

Currently working to evaluate human-generated feedback provided to the short answer tasks being studied. Early indications reveal promising economy of scale for feedback provided when conditioned on consensus labels applied to task-responses.

References (1/2)

- 1 Attali, Y., Powers, D., Freedman, M., Harrison, M., & Obetz, S. (2008). Automated Scoring of Short-Answer Open-Ended Gre® Subject Test Items. *ETS Research Report Series, 2008(1)*, i–22.
- 2 Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: a Clustering Approach to Amplify Human Effort for Short Answer Grading. *Transactions of the Association for Computational Linguistics, 1*, 391–402. https://doi.org/10.1162/tacl_a_00236
- 3 Beckman, M. (2015). Assessment Of Cognitive Transfer Outcomes For Students Of Introductory Statistics.
<http://conservancy.umn.edu/handle/11299/175709>
- 4 GAISE College Report ASA Revision Committee (2016). Guidelines for Assessment and Instruction in Statistics Education College Report 2016. URL: <http://www.amstat.org/education/gaise>
- 5 Gould, R. (2010). Statistics and the Modern Student. *International Statistical Review / Revue Internationale de Statistique, 78(2)*, 297–315. <https://www.jstor.org/stable/27919839>
- 6 Guo, W., Diab, M. (2012) Modeling Sentences in the Latent Space. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 864–872. Association for Computational Linguistics.

References (2/2)

- 7 Graham, S., Kiuahara, S. A., & MacKay, M. (2020). The Effects of Writing on Learning in Science, Social Studies, and Mathematics: A Meta-Analysis. *Review of Educational Research*, 90(2), 179–226. <https://doi.org/10.3102/0034654320914744>
- 8 Li, Z., Tomar, Y., & Passonneau, R. J. (2021). A Semantic Feature-Wise Transformation Relation Network for Automatic Short Answer Grading. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6030–6040. Association for Computational Linguistics. <https://aclanthology.org/2021.emnlp-main.487>
- 9 Page, E. B. (1994). Computer Grading of Student Prose, Using Modern Concepts and Software. *The Journal of Experimental Education*, 62(2), 127–142.
- 10 Pearl, D. K., Garfield, J. B., delMas, R., Groth, R. E., Kaplan, J. J., McGowan, H., & Lee, H. S. (2012). Connecting Research to Practice in a Culture of Assessment for Introductory College-level Statistics. URL: http://www.causeweb.org/research/guidelines/ResearchReport_Dec_2012.pdf
- 11 Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine*, 37(5), 360–363.

Thank You

Foundations for AI-assisted formative assessment
feedback for short-answer tasks in
large-enrollment classes

Matthew Beckman
Penn State University

11th International Conference on Teaching Statistics (ICOTS)

Rosario, Argentina
September 13, 2022

Resource Page URL: <https://mdbeckman.github.io/ICSA2022/>