Research evaluating NLP tools designed to assist instructors with formative assessment for students in large-enrollment STEM education classes

Matthew Beckman
Penn State University

November 20, 2024

Research evaluating NLP tools designed to assist instructors with formative assessment for students in large-enrollment STEM education classes

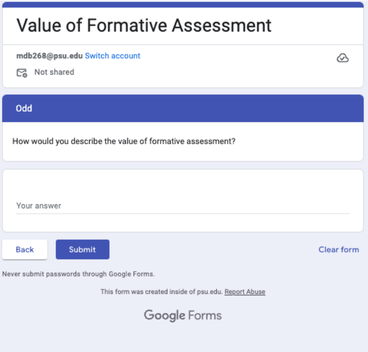
Matthew Beckman
Penn State University

November 20, 2024

Two question survey before seminar

Figure 1: (QR Code) https://forms.gle/

Responses to our survey?



Value of Formative Assessment

mdb268@psu.edu Switch account

Not shared

Odd

How would you describe the value of formative assessment?

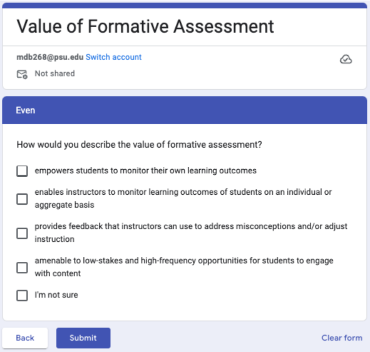
Your answer

Back | Submit | Clear form

Never submit passwords through Google Forms.

This form was created inside of psu.edu. Report Abuse

Google Forms



Value of Formative Assessment

mdb268@psu.edu Switch account

Not shared

Even

How would you describe the value of formative assessment?

- [ ] empowers students to monitor their own learning outcomes
- [ ] enables instructors to monitor learning outcomes of students on an individual or aggregate basis
- [ ] provides feedback that instructors can use to address misconceptions and/or adjust instruction
- [ ] amenable to low-stakes and high-frequency opportunities for students to engage with content
- [ ] I'm not sure

Back | Submit | Clear form

Motivation

- "Write-to-learn" tasks improve learning outcomes (Graham, et al., 2020)
- Critical for citizen-statisticians to communicate effectively (Gould, 2010)
- Frequent practice w/ communicating improves statistical literacy and promotes retention (Basu, et al., 2013)
- Formative assessment benefits both students & instructors (Black & Wiliam, 2009; GAISE, 2016; Pearl, et al., 2012)
- A majority of U.S. undergraduates at public institutions take at least one large-enrollment STEM course (Supiano, 2022)
- **Logistics of constructed response tasks jeopardize use in large-enrollment classes** (Garfield & Ben-Zvi, 2008; Woodard & McGowan, 2012)

Easy!

Erm...

Goal

Develop technology that can assist instructors for large (STEM) classes with providing targeted formative assessment feedback to students, such that instructor burden is similar to small class (∼30 students)

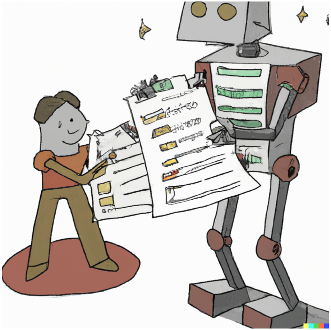
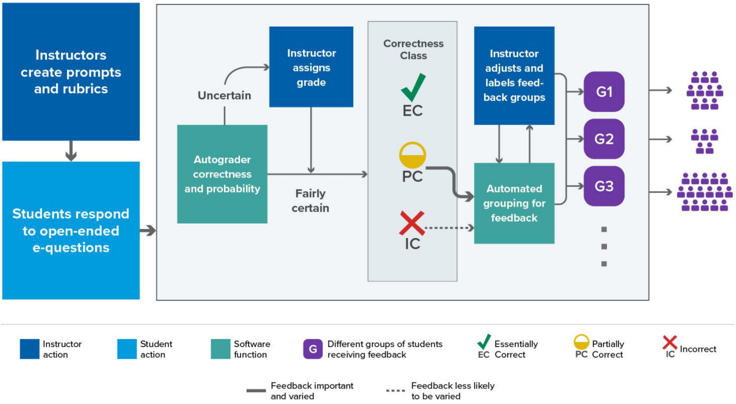


Figure 2: image created with assistance of DALL · E 2 by Open AI

Project Schematic



Goal: Computer-assisted formative assessment feedback for short-answer tasks in large-enrollment classes, such that instructor burden is similar to small class (~30 students)

Research Questions

- **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?

Pilot Study & Follow-Up Investigation

- Lloyd, S. E., Beckman, M., Pearl, D., Passonneau, R., Li, Z., & Wang, Z. (2022). Foundations for AI-Assisted Formative Assessment Feedback for Short-Answer Tasks in Large-Enrollment Classes. In *Proceedings of the eleventh international conference on teaching statistics*. Rosario, Argentina.
- Beckman, M., Burke, S., Fiochetta, J., Fry, B., Lloyd, S. E., Patterson, L., & Tang, E. (2024). Developing Consistency Among Undergraduate Graders Scoring Open-Ended Statistics Tasks. Preprint URL: https://arxiv.org/abs/2410.18062

Collaborators

Figure 3: Lloyd et al., (2022) Pilot Project Team

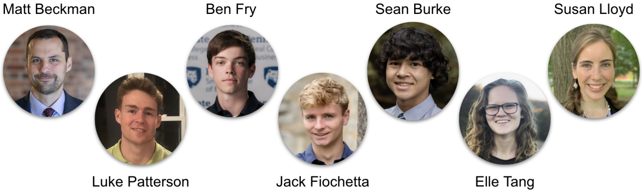


Figure 4: Beckman et al., (2024) Project Team

Some results before the methods

- "short-answer" tasks are good for students, but hard to scale
- Can NLP tools help instructors give students feedback?
  - Evaluate & group student responses
  - Compare agreement between NLP & humans
  - What might scalable, personalized feedback look like anyway?

(Spoiler Alert)

- Results suggest. . .

Some results before the methods

- “short-answer” tasks are good for students, but hard to scale
- Can NLP tools help instructors give students feedback?
  - Evaluate & group student responses
  - Compare agreement between NLP & humans
  - What might scalable, personalized feedback look like anyway?

(Spoiler Alert)

- Results suggest...
- Instructor agreement (QWK ≈ 0.7 to $0.8+$)

Some results before the methods

- "short-answer" tasks are good for students, but hard to scale
- Can NLP tools help instructors give students feedback?
  - Evaluate & group student responses
  - Compare agreement between NLP & humans
  - What might scalable, personalized feedback look like anyway?

(Spoiler Alert)

- Results suggest. . .
- Instructor agreement (QWK ≈ 0.7 to $0.8+$)
- UTA agreement (QWK ≈ 0.6 to $0.7+$)

Some results before the methods

- "short-answer" tasks are good for students, but hard to scale
- Can NLP tools help instructors give students feedback?
  - Evaluate & group student responses
  - Compare agreement between NLP & humans
  - What might scalable, personalized feedback look like anyway?

(Spoiler Alert)

- Results suggest...
- Instructor agreement (QWK ≈ 0.7 to $0.8+$)
- UTA agreement (QWK ≈ 0.6 to $0.7+$)
- NLP algorithm agreement with instructors (QWK $\approx 0.7+$)

Some results before the methods

- "short-answer" tasks are good for students, but hard to scale
- Can NLP tools help instructors give students feedback?
  - Evaluate & group student responses
  - Compare agreement between NLP & humans
  - What might scalable, personalized feedback look like anyway?

(Spoiler Alert)

- Results suggest...
- Instructor agreement (QWK ≈ 0.7 to $0.8+$)
- UTA agreement (QWK ≈ 0.6 to $0.7+$)
- NLP algorithm agreement with instructors (QWK $\approx 0.7+$)
- Human-Algorithm partnership may be even better? ($\approx 0.85+$)

Some results before the methods

- "short-answer" tasks are good for students, but hard to scale
- Can NLP tools help instructors give students feedback?
  - Evaluate & group student responses
  - Compare agreement between NLP & humans
  - What might scalable, personalized feedback look like anyway?

(Spoiler Alert)

- Results suggest...
- Instructor agreement (QWK ≈ 0.7 to $0.8+$)
- UTA agreement (QWK ≈ 0.6 to $0.7+$)
- NLP algorithm agreement with instructors (QWK $\approx 0.7+$)
- Human-Algorithm partnership may be even better? ($\approx 0.85+$)
- Pursuing multiple avenues for grouping & feedback

Methods (Data Set)

Study utilized de-identified extant data & scoring rubrics (Beckman, 2015)

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

Note: this sample was *not* a single large class at one institution; these data include introductory statistics students from many class sections at many institutions–some large, others small.

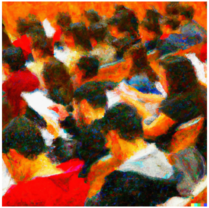


Figure 5: image created with assistance of DALL · E 2 by Open AI

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

Methods (Rater Agreement)

- Lloyd et al., (2022)
  - 3 raters typical of large-enrollment instruction team
  - (6 tasks) x (1,935 students) distributed among the team
  - sufficient intersection to assess inter-rater agreement
  - responses judged Correct / Partial / Incorrect against rubric
- Beckman at al., (2024)
  - 4 Undergraduate Teaching Assistants (UTAs) and 1 instructor
  - (4 tasks) x (63 students) scored by each UTA + Instructor
  - 5 sequential exercises associated with progression of scoring development

Results: Instructors as Graders

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

Figure 7: Interrater agreement among three instructors; intra-rater agreement for Rater A with several years delay

Reliability intuition[1]: moderate $< 0.6 <$ substantial $< 0.8 <$ near perfect < 1.0

[1] Viera & Garret (2005)

Results: Instructor and UTA Graders

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

Raters	Day 1	Day 5	Week 10
A & E	0.46 (0.35, 0.58)	0.57 (0.47, 0.67)	0.58 (0.49, 0.67)
A & F	0.61 (0.50, 0.71)	0.72 (0.64, 0.79)	0.78 (0.71, 0.85)
A & G	0.63 (0.55, 0.72)	0.73 (0.66, 0.80)	0.73 (0.66, 0.81)
A & H	0.72 (0.65, 0.80)	0.71 (0.63, 0.78)	0.68 (0.59, 0.78)

Figure 8: Pairwise agreement between UTAs and an instructor (Rater A)

Reliability intuition: moderate $<$ 0.6 $<$ substantial $<$ 0.8 $<$ near perfect $<$ 1.0

Results: Instructor & UTA (cont'd)

Raters	QWK	95% CI
A	0.82	(0.76, 0.88)
E	0.57	(0.46, 0.68)
F	0.74	(0.67, 0.82)
G	0.66	(0.56, 0.76)
H	0.74	(0.67, 0.81)

Figure 9: Intra-rater agreement (self-consistency) for each participant as measured with Quadratic Weighted Kappa (QWK) while scoring the same set of student responses on two occasions approximately 10 weeks apart.

Date (Exercise)	Rubric Description	AC_2	95% CI
Day 1 (Ex 1)	Solution with Verbal Instructions	0.688	(0.63, 0.74)
Day 5 (Ex 4)	Expert Rubric, Part 1	0.784	(0.75, 0.82)
Week 10 (Ex 5)	Expert Rubric, Part 2	0.778	(0.74, 0.81)

Figure 10: Group agreement among four undergraduate TAs and one instructor, as measured with Gwet's (2014) AC2; 95% confidence intervals accompany each estimate.

Methods (RQ2)

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

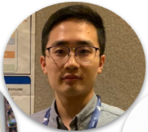
Susan Lloyd

Matt Beckman

Dennis Pearl

Becky Passonneau

Zhaohui Li



Semantic Feature-Wise Transformation Relation Network (SFRN)

Paper introducing SFRN

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

Meet the "machine": NLP for Assessment

- Natural language processing (NLP) involves how computers can be programmed to analyze language elements
- NLP-assisted feedback for educational use:
  - automated short-answer grading (ASAG) from 2009
  - essays & long-answer tasks earlier
- Human-machine collaboration is a promising mechanism to assist rapid, individualized feedback at scale (Basu, 2013)
- Deep neural networks application since 2016
- Relational (neural) networks

Meet the "machine": Relational Networks

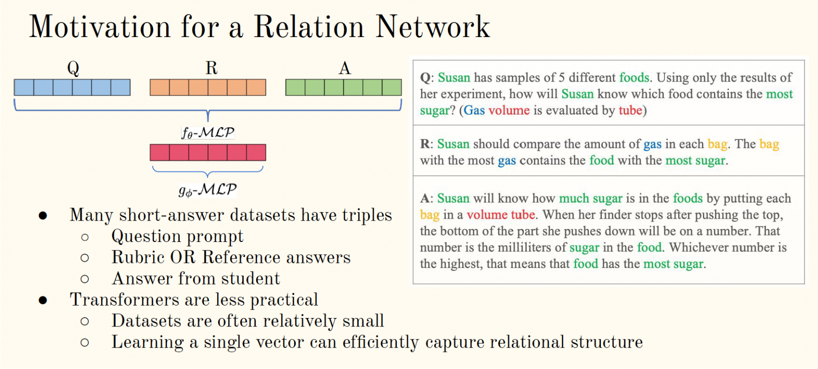


Figure 11: Image credit: Becky Passonneau

- much of the architecture inspired by work from computer vision
- more efficient than transformer networks (e.g., LLMs)

Meet the “machine”: SFRN Schematic

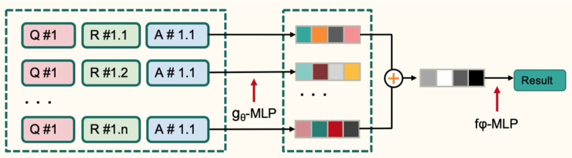


Figure 12: encoder (Left); fusion function (Middle); classifier (Right).

Semantic Feature-Wide Transformation Relation Network (SFRN):

- end-to-end model with three components:
  - ($g_\theta MLP$) pretrained BERT encoder (LLM) » vector representations
  - ($+$) learned feature-wise transformation function fuses multiple representations, if necessary (e.g., multiple reference answers)
  - ($f_\phi MLP$) is a classifier algorithm, i.e., neural network
- relation networks designed to learn generalizations that infer meaning in a data-efficient way
- data augmentation during training step

Results (RQ2)

RQ2: What level of agreement is achieved between instructors and the machine (an NLP algorithm)?

Comparison	Reliability
Rater A & SFRN	QWK = 0.79
Rater C & SFRN	QWK = 0.82
Rater D & SFRN	QWK = 0.74

Figure 13: Pairwise agreement with SFRN algorithm

Reliability intuition: moderate $< 0.6 <$ substantial $< 0.8 <$ near perfect < 1.0

Human-Machine Partnership?

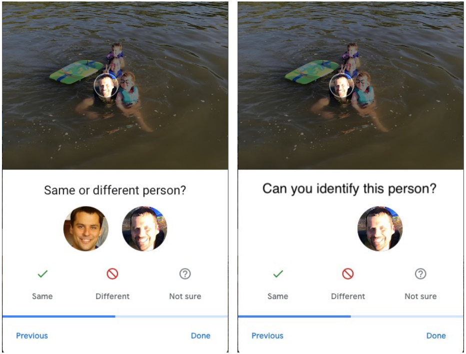


Figure 14: Illustration adapted from Google Photos

Our approach to human-in-the-loop (HIL) did not make a recommendation (e.g., Left), it just shows examples to the human when it needs help (e.g., Right).

Human-Machine Partnership Method

Want to evaluate accuracy of marking algorithm when designed to "defer" to human judgment

- algorithm evaluates a probability for each label (EC, PC, IC)
  - if a label has high probability, use algorithm label
  - if no label has sufficiently high probability, defer to human
- interests
  - estimate how frequently the algorithm defers
  - estimate accuracy of the combined process

Human-Machine Partnership Results

Our work is first that we know of to impelement controllable, selective prediction deferral policy for the classifier (i.e., scoring) step.

Threshold	Deferral Rate	Simulated HIL Accuracy
0.68	9.5%	0.855
0.75	13.2%	0.861
0.80	16.0%	0.871
0.85	20.2%	0.884
0.90	25.6%	0.899

Figure 15: Accuracy of Human-in-the-loop compared with expert label ground truth.

Methods (RQ3)

How similar is feedback provided by trained humans?

- Experiment #1: Humans
  - Two reviewers independently evaluated 100 “partial credit” responses
  - Each reviewer provided free-text feedback to each student
  - Verbatim feedback captured for each reviewer and cross-tabulated for analysis.
- Experiment #1: NLP Tools
  - retrain k-means & k-mediods clustering & evaluate stability
  - compare representations with higher & lower dimensionality
- Experiment #2
  - if feedback labels are pre-determined, how consistently are they applied?
  - (i.e., ~~clustering~~ => FB Classifier??)
  - Both Humans & NLP Tools attempt
  - New tool “AsRRN” (Li, Lloyd, Beckman, & Passonneau, 2023)

Results (RQ3 humans)

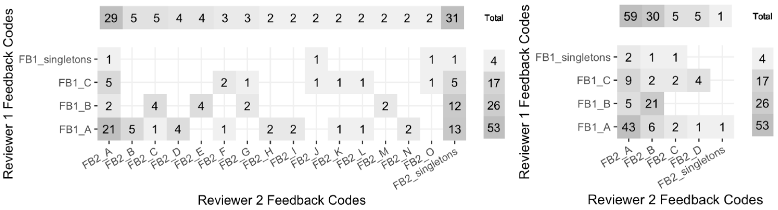


Figure 16: Cross-tabulation of feedback distribution for the two reviewers for the initial feedback (left) compared with the same analysis for the portion of feedback related to the statistical concept at issue (right).

- Reviewer 1 favored feedback on statistical concepts (only).
- Reviewer 2 provided same, plus a quote from the student
- Reviewer 2 parsed feedback to compare remarks related to the statistical concepts (only) with that of Reviewer 1.

Results (RQ3 humans)

Feedback Code	Feedback verbatim text suggested by the Reviewer
FB1_A (Reviewer 1)	What can we do to evaluate whether [the] result is better than we would expect for someone that is strictly guessing?
FB2_A (Reviewer 2)	Think about what inferential statistical method might we use to evaluate the percentage of correctly identified notes.
FB1_B (Reviewer 1)	Good idea to have a threshold for comparison, but it's very important that it be established carefully. For example, how might you establish a threshold that…
FB2_B (Reviewer 2)	Why this threshold? What inferential statistical method might we use to evaluate the percentage of correctly identified notes?

Figure 17: Verbatim feedback most frequently provided by each reviewer for responses to task 2B.

Results (RQ3 machines)

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

- SFRN (D = 512) produced reasonably consistent clusters when retrained (0.62)
- Highest consistency (0.88; D = 50) was achieved using a matrix factorization method that produces static representations (WTMF; Guo & Diab, 2011)
- AsRRN compared to humans (A & B) grouping students by pre-determined feedback categories:

Task	Sample Size	A & B	A & AsRRN	B & AsRRN
1	90	0.71	0.53	0.69
2	100	0.45	0.70	0.41

Discussion

OFFICE OF
Educational Technology

Artificial Intelligence and the Future of Teaching and Learning

Insights and Recommendations

May 2023

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
  - Human-in-the-Loop » Instructor / Algorithm partnership
- **RQ3**: Promising results based on “man-made clusters” but classification and clustering have competing incentives when it comes to dimensionality of NLP vector representations
  - Lower Dim is generally better for cluster stability
  - Higher Dim better for classification reliability
  - Exploring Topological Analysis as alternative to clustering
  - Feedback as a classifier (Li et al., 2023)

Current Events

- challenge system with diverse tasks, institutions, student populations;
  - partnering with ISU, MSU, PSU, UCSB, UF, & UTEP
  - two "consensus" tasks implemented by all
  - 2-3 local tasks at each institution
  - (so far) approx 22,300 responses from $\sim$ 13,000 students
- robustness toward languistic diversity (HSI & NZ data)
- accumulated data to be shared with broader NLP community
  - this will be among the largest *open* data sources of it's kind
  - addresses barriers imposed by proprietary data sources on NLP research
- shiny new algorithm: AsRRN
  - contrastive loss function
  - correct answers are similar; there are a few distinct ways to earn partial credit; there are many diffuse ways to be incorrect
  - accommodates more complex task structure

Acknowledgment

- We're grateful to the US National Science Foundation for funding this research; Project CLASSIFIES (NSF DUE-2236150).
- Thanks to students and faculty at partner institutions that have assisted us with data collection.

References (1/3)

1. 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
2. Beckman, M. (2015). Assessment Of Cognitive Transfer Outcomes For Students Of Introductory Statistics. http://conservancy.umn.edu/handle/11299/175709
3. Beckman, M., Burke, S., Fiochetta, J., Fry, B., Lloyd, S. E., Patterson, L., & Tang, E. (2024). Developing Consistency Among Undergraduate Graders Scoring Open-Ended Statistics Tasks. Preprint URL: https://arxiv.org/abs/2410.18062
4. Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21,* pp 5-31. https://doi.org/10.1007/s11092-008-9068-5
5. 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

References (2/3)

6 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

7 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.

8 Graham, S., Kiuhara, 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.

9 Gwet, K. (2014). Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters

10 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

References (3/3)

11 Li, Z., Lloyd, S., Beckman, M. D., & Passonneau, R. J. (2023). Answer-state Recurrent Relational Network (AsRRN) for Constructed Response Assessment and Feedback Grouping. *Findings of the Association for Computational Linguistics: EMNLP 2023*. https://doi.org/10.18653/v1/2023.findings-emnlp.254

12 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

13 U.S. Department of Education, Office of Educational Technology (2023). Artificial Intelligence and Future of Teaching and Learning: Insights and Recommendations, Washington, DC.

14 Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine, 37*(5), 360–363.

15 Woodard, R., & McGowan, H. (2012). Redesigning a large introductory course to incorporate the GAISE guidelines. *Journal of Statistics Education, 20*(3).

Thank You

Research evaluating NLP tools designed to assist instructors with formative assessment for students in large-enrollment STEM education classes

Matthew Beckman
Penn State University

November 20, 2024