Progress toward NLP-assisted formative assessment feedback

Matthew Beckman
Penn State University

November 15, 2023

Progress toward NLP-assisted formative assessment feedback

Matthew Beckman
Penn State University

November 15, 2023

Two question survey before seminar (scan with mobile phone)

Figure 1: (QR Code) https://forms.gle/hpW72fMYE1SsB19JA

One slide overview.

- *Pedagogy*: Writing is important for students

One slide overview.

- *Pedagogy*: Writing is important for students
- *Logistics*: Rapid feedback doesn't scale easily

One slide overview.

- *Pedagogy*: Writing is important for students
- *Logistics*: Rapid feedback doesn't scale easily
- *Theory*: Similar responses benefit from similar feedback

One slide overview.

- *Pedagogy*: Writing is important for students
- *Logistics*: Rapid feedback doesn't scale easily
- *Theory*: Similar responses benefit from similar feedback
- *Tools*: NLP algorithms can help classify & cluster responses

One slide overview.

- *Pedagogy*: Writing is important for students
- *Logistics*: Rapid feedback doesn't scale easily
- *Theory*: Similar responses benefit from similar feedback
- *Tools*: NLP algorithms can help classify & cluster responses
- *Solution*: Human-in-the-loop AI; complimentary strengths

One slide overview.

- *Pedagogy*: Writing is important for students
- *Logistics*: Rapid feedback doesn't scale easily
- *Theory*: Similar responses benefit from similar feedback
- *Tools*: NLP algorithms can help classify & cluster responses
- *Solution*: Human-in-the-loop AI; complimentary strengths
- *Evaluation*: Algorithm shouldn't be "worse" than humans

One slide overview.

- *Pedagogy*: Writing is important for students
- *Logistics*: Rapid feedback doesn't scale easily
- *Theory*: Similar responses benefit from similar feedback
- *Tools*: NLP algorithms can help classify & cluster responses
- *Solution*: Human-in-the-loop AI; complimentary strengths
- *Evaluation*: Algorithm shouldn't be "worse" than humans
- *Evaluation*: How well do humans agree?

One slide overview.

- *Pedagogy*: Writing is important for students
- *Logistics*: Rapid feedback doesn't scale easily
- *Theory*: Similar responses benefit from similar feedback
- *Tools*: NLP algorithms can help classify & cluster responses
- *Solution*: Human-in-the-loop AI; complimentary strengths
- *Evaluation*: Algorithm shouldn't be "worse" than humans
- *Evaluation*: How well do humans agree?
- *Challenge*: clusters vs. feedback

Responses to our survey?

1. Is your lucky/favorite number odd or even?
2. How did you describe the value of formative assessment?
   - [Odd] Free text response: *write anything you like*
   - [Even] Selected response: *endorse provided options*

(Recent) survey responses: “Odd” number

A formative assessment is one that can be used to evaluate someone understanding of formalized learning concepts.

I’m not sure what it is, but thinking that it probably is helpful for understanding pain points in the learning process as it is ongoing. Assuming it gives an oppty to provide iterative feedback at multiple points in time?

Assess the formative value of tasks

It's a very good tool to evaluate the performance of students

Great way to gauge student progress

Very helpful for both learners and instructor to have understanding

It allows students to reevaluate their thinking, which helps their learning by correcting misunderstandings. It can also encourage students who are doing well.

Figure 2: How would you describe the value of formative assessment? (as a constructed response task)

(Recent) survey responses: “Even” number

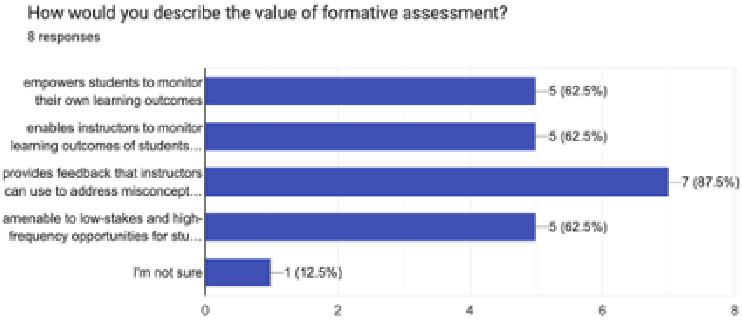


Figure 3: How would you describe the value of formative assessment? (as a selected response task)

Motivation

Pedagogy

- "Write-to-learn" improve learning outcomes (Graham, et al., 2020)
- Critical for citizen-statisticians to communicate well (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)

Logistics

- 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!

Figure 4: One of my classes; students did lots of writing (& computer programming) and everyone frequently received personalized feedback.

Erm...

Figure 5: One of my classrooms; students did NOT do much writing and personalized feedback was very rare.

Figure 6: Recent report on AI and the future of teaching and learning from US Dept of Education, Office of Educational Technology (May 2023)[1]

[1] 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*.

Recommendations from US Dept of Education[2]

[2] 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*.

Goal state

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

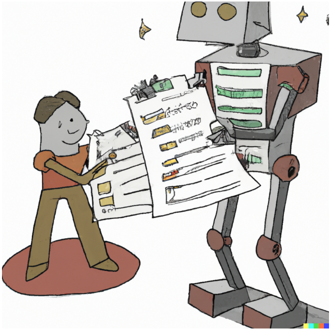


Figure 7: image created with assistance of DALL · E 2 by Open AI

Collaborators (humans)

Susan Lloyd Dennis Pearl Zhaohui Li

Matt Beckman Becky Passonneau

Tools (machines)

- 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
- Semantic Feature-Wise Transformation Relation Network (SFRN)[3]
  - back-translation data augmentation (French & Chinese)
  - can accommodate rubrics, expert solutions, or both

---

[3] 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*.

SFRN Detail (Li et al., 2021)

SFRN is an end-to-end model with 3 components:

1. encode QRA triples producing vector representations for question (Q), a possible reference (R), and student answer (A)
2. when relation network includes multiple QRA triples, a learned feature-wise transformation network merges all relation vectors for a student answer into a single relation vector by leveraging attentions calculated by a QRA triple;
3. the resulting vector representation is passed as an input to a classifier (i.e., neural network)

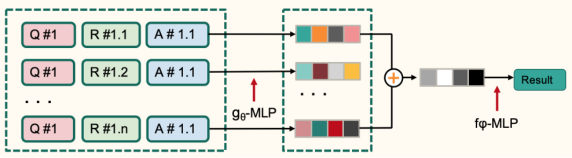
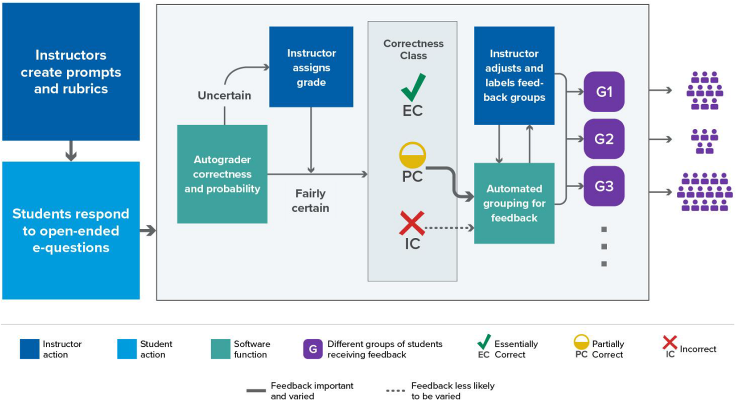


Figure 8: The $g_\theta MLP$ function computes the relation vector for each [Q,R,A] triple. A set of relation vectors is combined ($+$) using *SFT*. The $f_\phi MLP$ function is the assessment classifier.

Schematic of Partial Solution



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

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

ICOTS Paper

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.

Research Questions

- **RQ3**: What sort of NLP representation leads to good clustering performance, and how does that interact with the classification algorithm?
- **RQ2**: What level of agreement is achieved between human raters and an NLP algorithm?

ICOTS Paper

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.

Research Questions

- **RQ3**: What sort of NLP representation leads to good clustering performance, and how does that interact with the classification algorithm?
- **RQ2**: What level of agreement is achieved between human raters and an NLP algorithm?
- **RQ1**: What level of agreement is achieved among trained human raters labeling/scoring short-answer tasks (a few sentences)?

ICOTS Paper

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.

Research Questions

- **RQ1**: What level of agreement is achieved among trained human raters labeling/scoring short-answer tasks (a few sentences)?
- **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?

ICOTS Paper

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.

Spoilers?!

- **RQ1**: What level of agreement is achieved among trained human raters labeling/scoring short-answer tasks (a few sentences)?
- **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: evidence of productive clustering; more work to do

Methods (Sample)

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 is *not* a single large class at some institution; the available data includes introductory statistics students from many class sections at many institutions–some classes were quite small.

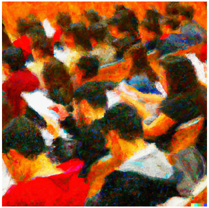


Figure 9: image created with assistance of DALL · E 2 by Open AI

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

Methods (RQ1)

- 3 human raters typical of large-enrollment instruction team
- entire sample (1,935 students) distributed among the team with sufficient intersection to assess rater agreement
- 63 student responses in common for each *combination* of raters to quantify agreement (e.g., pairwise, consensus, etc)
- constraint: sufficient data for *intra-rater* analysis for person that had labeled 178 responses 6-7 years prior

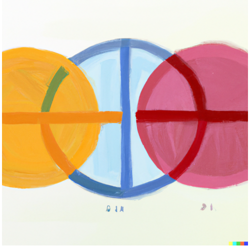


Figure 11: "Venn diagram of three intersecting sets" according to DALL · E 2 by Open AI. Kind of a swing and a miss. . .

Results (RQ1)

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

Inter-rater agreement:

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 interpretation[4]: $0.6 <$ substantial $< 0.8 <$ near perfect < 1.0

[4] Viera & Garrett (2005)

Methods (RQ2)

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

- 90% of data for development purposes (train); 8:1:1 partition
  - training (72%),
  - development (9%)
  - evaluation (9%)
- 10% of data being held in reserve (test)

SFRN was compared to other NLP algorithms for accuracy using a subset of student responses (Li et al., 2021).

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

Results (RQ2)

Prerequisite–comparing machines: The SFRN algorithm achieved much higher classification accuracy than LSTM (83% vs. 72%) when judged against human consensus ratings.[5]

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

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 interpretation: $0.6 <$ substantial $< 0.8 <$ near perfect < 1.0

[5] SFRN & LSTM comparison excludes instances when human labels disagree

Methods (RQ3)

Manual pilot of human-generated clustering

- Two reviewers independently evaluated 100 student responses that earned "partial credit" on inference tasks
- Each reviewer provided free-text feedback to each student
- Verbatim feedback captured for each reviewer and cross-tabulated for analysis.

Preliminary experiment with NLP representations

- retrain k-means & k-mediods clustering to evaluate cluster stability
- compare representations with higher & lower dimensionality

Results (RQ3 humans)

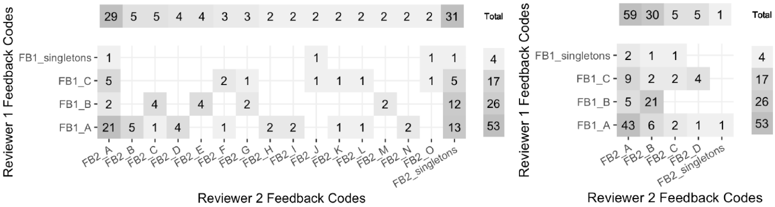


Figure 12: Cross-tabulation of feedback distribution for the two reviewers for the initial feedback (left) compared with the same analysis for the portion of feedback related to the statistical concept at issue (right).

- Reviewer 1 favored feedback on statistical concepts (only).
- Reviewer 2 provided same, plus a quote from the student
- Reviewer 2 parsed her feedback to compare her remarks related to the statistical concepts (only) with the feedback 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 13: 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 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 LSTM *despite 8X higher dimensionality* [6]
- Highest consistency (0.88; D = 50) was achieved using a matrix factorization method that produces static representations (WTMF; Guo & Diab, 2011)

[6] Consistency 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; Lower D is better for cluster stability, Higher D better for classification reliability. (SFRN clustering was respectable despite high D, though)

Limitations

- Study uses extant data from prior study collected from many classes of varying size
  - not a single large class
  - no covariates available to identify and mitigate bias labeling (human or machine)
  - Tasks & rubrics used for pilot were developed for research purposes; likely more polished than tasks developed “in the wild”
- Clustering performance vs semantic meaning
  - clustering is necessary, but not sufficient, for meaningful feedback
  - semantic meaning of NLP clusters not yet rigorously studied

Ongoing Data Collection (Fall 2023 +)

- challenge labeling algorithm with linguistic diversity;
  - approx 13,000 task-responses in Fall 2023
  - 2 of 5 institutions are HSI's
- self-reported demographic covariates
  - language(s) at home
  - race & ethnicity
  - gender
  - academic major
- diversify item and rubric input to challenge performance

Current Events: HIL deferral policy

Our work is first (that we know of) to implement controllable, selective prediction deferral policy

Threshold	Deferral Rate	Simulated HIL Accuracy
0.68	0.095	0.855
0.75	0.132	0.861
0.80	0.160	0.871
0.85	0.202	0.884
0.90	**0.256**	**0.899**
0.95	0.418	0.931

Current Events: Improving on SFRN

- Answer-state Recurrent Relational Network (AsRRN)
  - Breaks from reliance on linear architecture of SFRN
  - Allows flexibility to accommodate shared stem with multiple prompts
  - Better incorporates reference answers corresponding to rubric guidance
- Contrastive Loss Function
  - Correct answers are generally alike
  - Many PC results align with a few common archetypes
  - Diverse ways to be incorrect

EMNLP Paper

Li, Z., Lloyd, S. E., Beckman, M., & Passonneau, R. (accepted). Answer-state Recurrent Relational Network (AsRRN) for Constructed Response Assessment and Feedback Grouping. In *Findings of the 2023 Empirical Methods in Natural Language Processing.* Singapore.

Future work

- accommodation for bias (e.g., human; algorithm)
- iterative instructor input to group conceptual representations
- field test key aspects of project CLASSIFIES in large classes
- open questions for "what works" in formative assessment
- accumulated data made available to broader NLP community
    - would be among the largest open data sources of it's kind
    - addresses barriers imposed by proprietary data on NLP research

Acknowledgments & Disclaimers

This work was supported by the National Science Foundation under Grant No. DUE-2236150. 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 National Science Foundation.

References (1/2)

- 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.
- 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
- Beckman, M. (2015). Assessment Of Cognitive Transfer Outcomes For Students Of Introductory Statistics. http://conservancy.umn.edu/handle/11299/175709
- 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
- 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
- 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
- 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.
- 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.

References (2/2)

- 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

- Li, Z., Lloyd, S. E., Beckman, M., & Passonneau, R. (accepted). Answer-state Recurrent Relational Network (AsRRN) for Constructed Response Assessment and Feedback Grouping. In *Findings of the 2023 Empirical Methods in Natural Language Processing.* Singapore.

- 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.

- Page, E. B. (1994). Computer Grading of Student Prose, Using Modern Concepts and Software. *The Journal of Experimental Education, 62*(2), 127–142.

- 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

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

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

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

Thank You

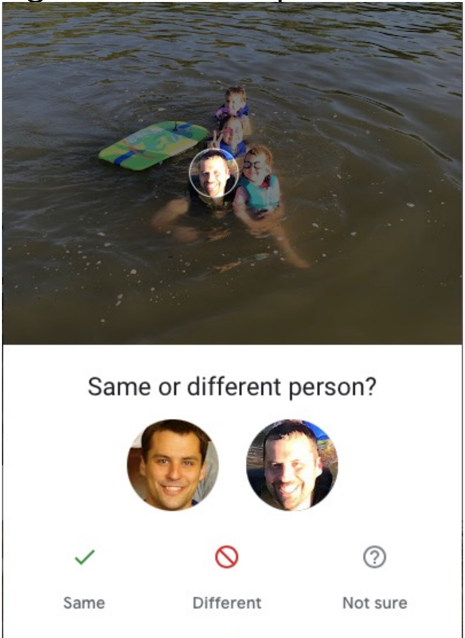
Progress toward NLP-assisted formative assessment feedback

Matthew Beckman
Penn State University

November 15, 2023

Resource Page URL: https://mdbeckman.github.io/QUT2023/

Google Photos Comparative Judgment



Google Photos "Deferral"