Using Machine Learning to Assess Question Performance

Hanyu Sun, Ting Yan, Anil Battalahalli, *Westat*

10/16/2023

Background

- The goal of questionnaire evaluation and testing is to reduce measurement error
- Behavior coding is one evaluation method built on coding of interviewer-respondent interactions during question-answer (Q-A) process
- Paradigmatic question-answering sequence (e.g., Schaeffer and Maynard, 1996):
  - I: How many days a week do you watch television?
  - R: Seven days
- Deviation or departure from this paradigmatic sequence indicates problems with the Q-A process

Behaviors Indicative of Poor Question Performance

- Interviewer behaviors:
  - Re-reading question
  - Probing
- Respondent behaviors:
  - Request for clarification/repeat/definition
  - Initial answer inadequate
  - Uncertainty/qualified answers

Using Machine Learning for Question Assessment (2)

- Automated processing of recordings
- Generated metrics that can be used for question assessment
  - Problematic respondent behaviors
    - Total number of respondent’s turns
      - >1 turn indicating respondent requesting for clarification/definition, inadequate initial answer
    - Duration of respondent’s 1st turn
      - Long turn indicating respondent having trouble understanding or answering the question

Using Machine Learning for Question Assessment

- Automated processing of recordings
- Generated metrics that can be used for question assessment
  - Problematic interview behavior
    - Total number of interviewer’s turns
      - >1 turn indicating interviewer re-reading question, probing
  - Q-A process deviating from paradigmatic sequence
    - Total duration
      - Longer time indicating problems with Q-A process

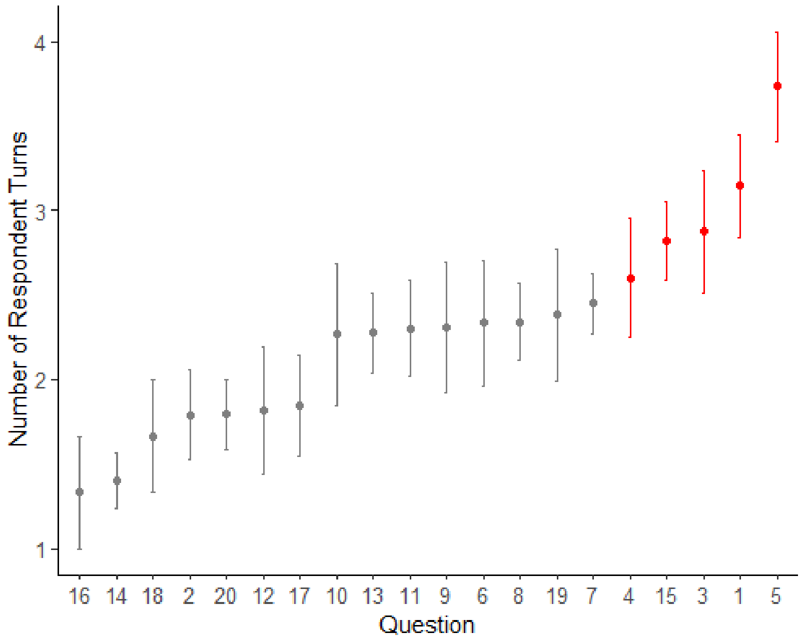
Data and Methods (1)

- 20 questions selected from a large-scale cross-sectional study of a nationally representative sample:
  - 479 question-answer recordings from 53 cases
    - 13 closed questions, 7 open-ended questions
    - 6 single choice questions, 7 multiple choice questions
    - 9 questions with showcards

Data and Methods (2)

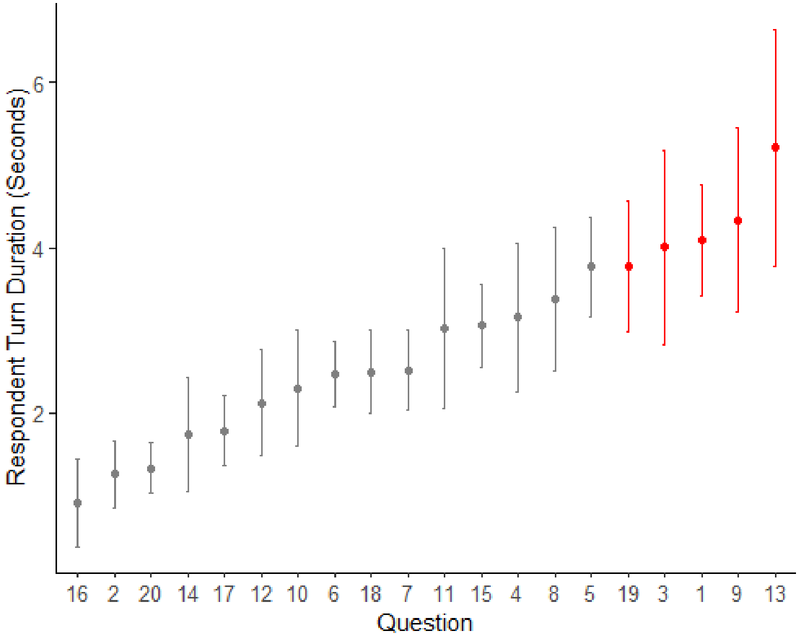
- Using metrics from the audio pipeline to identify questions with poor performance:
  - Number of interviewer's turns
  - Number of respondent's turns
  - Duration of respondent's 1st turn
  - Duration across all turns
- Expert review as validation:
  - 1 (not at all difficult) and 5 (the most difficult)
  - Produced a mean difficulty rating for each question

Results: Number of Respondent’s Turns



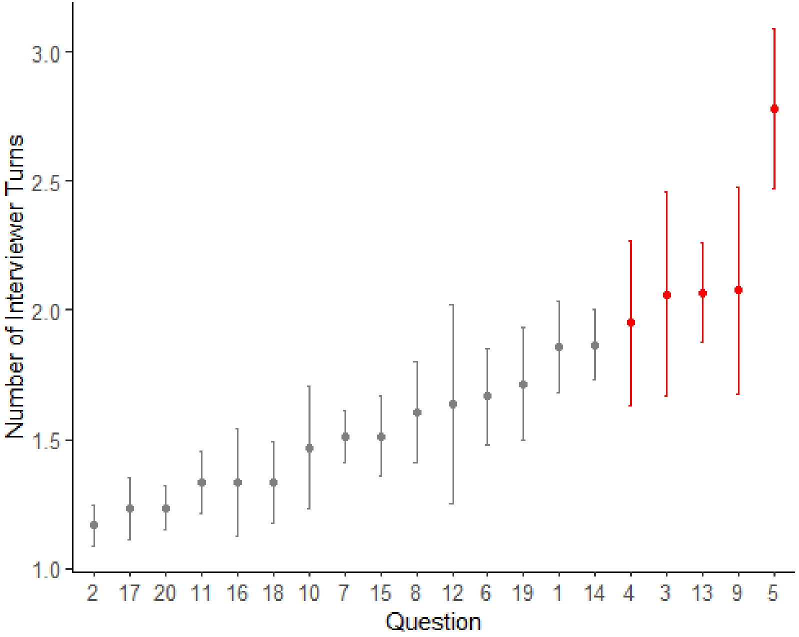
- >1 turn indicting respondent requesting for clarification/definition, inadequate initial answer
- According to the expert review:
  - The mean difficulty rating for Q1, Q3, and Q4 is 4.5
  - The mean difficulty rating for Q5 and Q15 is 3

Results: Duration of Respondent's 1st Turn



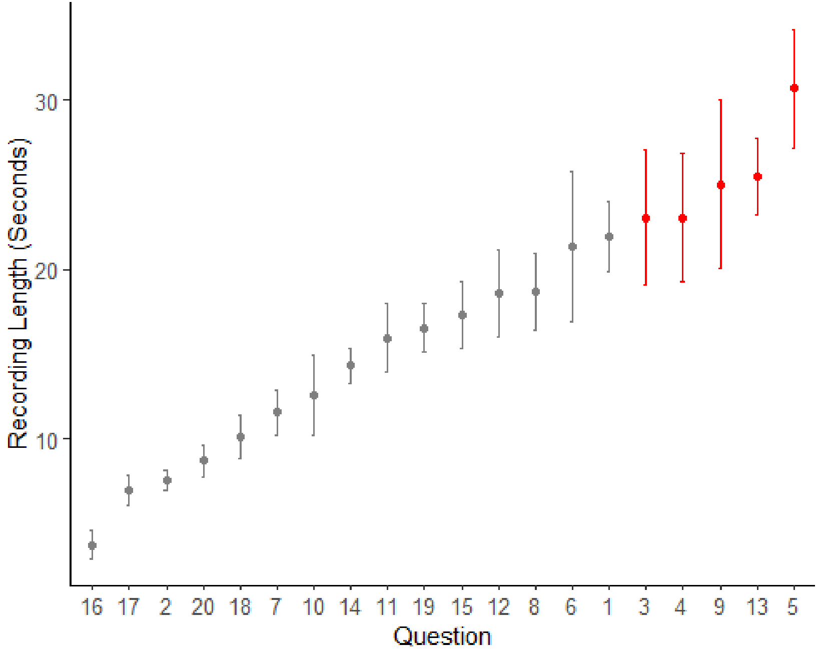
- Long turn indicating respondent having trouble understanding or answering the question
- According to the expert review:
  - The mean difficulty rating for Q1, Q3, Q9, and Q13 is 4.5
- Q19 is the last question of the interview asking for respondent's final comments, and its mean difficulty rating is 1.5

Results: Number of Interviewer's Turns



- >1 turn indicating interviewer re-reading question, probing
- According to the expert review:
  - The mean difficulty rating for Q3, Q4, Q9, and Q13 is 4.5
  - The mean difficulty rating for Q5 is 3

Results: Recording Length



- Longer time indicates problems with the Q-A process
- According to the expert review:
  - The mean difficulty rating for Q3, Q4, Q9, and Q13 is 4.5
  - The mean difficulty rating for Q5 is 3

Conclusions and Discussion (1)

- Advantages of machine learning
  - Real time automated processing, cost-efficient
  - Prioritize questions for human review
- The findings suggest that the metrics produced by the pipeline can be used for detecting problematic questions:
  - A common set of questions were identified as problematic by various metrics, e.g.,
    - Technical or unfamiliar terms
    - Not having the information in memory
    - Estimation difficulties

Conclusions and Discussion (2)

- Future work
  - Validate these metrics with conventional behavior coding
  - Improve the pipeline with results of conventional behavior coding
  - Understand relationship between metrics, question characteristics, question difficulty
  - Derive a composite score to rank questions on difficulty/issues

Thank You

Hanyu Sun (hanyusun@westat.com)

Ting Yan (tingyan@westat.com)

Anil Battalahalli (anilbattalahalli@westat.com)