Yan, Ting, Hanyu Sun, and Anil Battalahalli. 2025. "Using Machine Learning to Evaluate Questions in a Multilingual Survey." *Survey Practice* 19 Special Issue (March). https://doi.org/10.29115/SP-2024-0021.

ARTICLES

Using Machine Learning to Evaluate Questions in a Multilingual Survey

Ting Yan[1], Hanyu Sun[2], Anil Battalahalli[2]

[1] NORC at the University of Chicago, [2] Westat

Keywords: Computer-Assisted Recorded Interviewing (CARI), question assessment, machine learning, multilingual surveys

https://doi.org/10.29115/SP-2024-0021

Survey Practice

Vol. 19 Special Issue, 2025

Machine learning is becoming an increasingly popular tool in survey data collection. We recently conducted a successful application of machine learning to automatically detect survey questions at a higher risk of poor performance. The proof-of-concept study utilized items from an establishment survey and demonstrated that the machine learning methodology was able to identify the worst-performing items as evaluated by survey methodology experts. This paper extends ongoing research by applying the methodology to assess the performance of items in a multilingual household survey. In particular, we compared the performance of survey questions in both English and Spanish interviews. The results show that the use of machine learning identified a common set of items as difficult and prone to poor performance in both English and Spanish interviews. The implications of the findings are discussed.

Introduction

Question evaluation and testing are critical for multilingual surveys to uncover the effects of language, culture, and translation on survey answers and to ensure comparability across cultures and languages (e.g., Ji, Schwarz, and Nisbett 2000; Johnson et al. 1996; Lee, Schwarz, and Goldstein 2014; Yan and Hu 2019). A question assessment method that can evaluate questions in multiple languages simultaneously and effectively will greatly improve the efficiency and success of a multilingual survey.

Machine learning (ML) is becoming an increasingly popular tool in survey data collection. Sun and Yan (2023) described a pipeline that couples Computer-Assisted Recorded Interviewing (CARI) with ML and demonstrated that the pipeline could automatically and efficiently process 100% of recorded interviews to meet both objectives of behavior coding—identifying interviewers at a higher risk of falsification and *not* adhering to standardized interviewing protocol. They showed that, among 142 recordings validated as falsified cases, the ML pipeline was able to correctly detect 78% of them as falsified. Additionally, the measures generated by the ML algorithms to assess the extent to which interviewers read question verbatim and the extent to which interviewers maintained question meaning were in line with human coders' evaluations.

Yan, Sun, and Battalahalli (2024) described an extension of the pipeline aimed at improving the efficiency of using CARI for question evaluation and testing. Interested readers are referred to Figure 1 in Yan, Sun, and Battalahalli (2024) to understand how the pipeline works. In particular, the pipeline automatically generates seven outcome measures that can indicate poor question performance based on the literature of interaction analysis of question-answering sequences (see Table 1 in Yan, Sun, and Battalahalli [2024] for definitions of the measures). Six measures—number of respondent turns, duration of respondents' first turn, number of interviewer turns, total duration, presence of pauses, and presence of overlapping speech—flag interactional difficulties and breakdowns in the question-answering process. For question evaluation purposes, questions inducing more turns by respondents, longer first turns by respondents, more turns by interviewers, longer question-answering sequences, more pauses, and more overlapping speech are at a higher risk of poor performance. The seventh measure—positive emotions—enables researchers to examine challenges faced by respondents in the question-answering process from an emotional perspective. Specifically, questions that induce fewer instances of positive emotions are at a higher risk of poor performance. Yan, Sun, and Battalahalli (2024) used each of the seven outcome measures to rank 20 survey questions from an establishment survey and selected five of the worst performing items using each outcome as the criterion. They found that survey items judged to be the most difficult by survey experts were flagged as having worse performance by at least one measure, speaking to the promise of the pipeline in evaluating and testing survey questions.

As the advantages of the ML pipeline lie in the automatic processing of all recordings, Yan, Sun, and Battalahalli (2024) noted that the ML pipeline is an excellent diagnostic tool for screening items at a higher risk of poor performance and targeting them for further investigation. If the pipeline works equally well for recordings of interviews conducted in different languages, using it to screen and target items at a higher risk of poor performance in multiple languages simultaneously could lead to greater cost savings for multilingual surveys.

We applied the pipeline to interviews conducted in Spanish and compared its performance to evaluate survey questions administered to both English-speaking and Spanish-speaking respondents. Our focus is *not* on understanding the differences in interviewer-respondent interactions between English-speaking and Spanish-speaking interviews. In other words, we are *not* examining whether Spanish-speaking respondents have more turns or longer first turns than English-speaking counterparts. Instead, we are interested in evaluating whether the outcome measures generated by the pipeline (such as the number of respondent turns and duration of first respondent turn) can effectively identify problematic questions for Spanish-speaking respondents as successfully as they do for English-speaking respondents. If this approach

proves successful, researchers and practitioners could incorporate it into their toolkit to evaluate and monitor the performance of survey questions and interviewers before, during, and potentially after data collection, all in a cost-efficient manner.

Data

We randomly selected 2,926 recordings of question-answering sequences from a nationally representative survey of the U.S. civilian noninstitutionalized population. These recordings represent 17 survey questions and involve 1,123 respondents interviewed by 218 interviewers. The 17 survey items vary in question format. Fifteen items are closed-ended and two are open-ended. Among the 15 closed-end questions, 14 items require a single answer, and one is a check-all-that-apply item. Furthermore, a showcard is used for six out of 17 questions. In addition, 1,425 recordings are in Spanish and 1,501 in English.

We processed the recordings through the pipeline, which automatically generated the seven outcome measures described in the introduction. Each of the seven outcome measures was aggregated to the question level for assessment purposes for both English interviews and Spanish interviews. Questions were then ranked based on the aggregated question-level outcome measures and, for each measure, the five worst items were flagged, following the same procedures used by Yan, Sun, and Battalahalli (2024).

Results

We ranked the 17 questions by each of the aggregated outcome measures and plotted the results in Figures A1 to A7 in the Appendix. In each figure, the dots represent the averages, while the vertical lines indicate the spread of that outcome measure across the recordings. The five questions with the worst performance on each measure are marked in red. We have summarized the results in Table 1.

Table 1 demonstrates substantial agreement between outcome measures in identifying problematic survey items, as well as agreement between items identified in Spanish and English interviews. Among the 17 survey items examined, two items (Q14 and Q11) were *not* flagged as problematic by any outcome measure in either Spanish or English interviews. Eleven items were flagged as one of the top five worst-performing items in Spanish interviews by at least two outcome measures. These items were also flagged by at least one outcome measure in English interviews, suggesting that they were problematic for both Spanish and English respondents. We discuss these items in detail below.

Table 1. Results of Using MI for Question Assessment in English and Spanish Interviews

Question Number	Number of times flagged as top 5 worst items in Spanish interviews	Measure(s) that flagged item as worst performing in Spanish interviews	Number of times flagged as top 5 worst items in English interviews	Measure(s) that flagged item as worst performing in English interviews
13	6	**Respondent turn, R first turn duration, Interviewer turn, Total duration,** Long pause, overlapping speech,	5	**Respondent turn, R first turn duration, Interviewer turn, Total duration,** Positive emotion
17	4	Respondent turn, R first turn duration, Interviewer turn, Total duration	1	Overlapping speech,
16	3	**Interviewer turn, Total duration,** Positive emotion	4	Respondent turn, **Interviewer turn, Total duration,** Overlapping speech
4	3	**Respondent turn, R first turn duration, Long pause,**	4	**Respondent turn, R first turn duration, Long pause,** Overlapping speech,
6	3	Respondent turn, Interviewer turn, Total duration	3	R first turn duration, Long pause, Positive emotions
8	3	Respondent turn, Interviewer turn, Total duration	1	Long pause,
12	3	**R first turn duration**, Overlapping speech, Positive emotion	1	**R first turn duration**,
15	2	**Overlapping speech,** Positive emotion	5	Respondent turn, R first turn duration, Interviewer turn, Total duration, **Overlapping speech**
3	2	R first turn duration, **Long pause**	5	Respondent turn, Interviewer turn, Total duration, **Long pause**, Positive emotion
10	2	Long pause, Positive emotion	2	Interviewer turn, Total duration
5	2	**Long pause, Positive emotion**	2	**Long pause, Positive emotion**
1	1	Overlapping speech	0	
9	1	Positive emotion	0	
7	1	Overlapping speech	0	
2	0		2	Overlapping speech, Positive emotion
14	0		0	
11	0		0	

Q13 was flagged six times in Spanish interviews and five times in English interviews as one of the five worst items. Q13 is an open-ended question. It induced more respondent turns, more interviewer turns, and longer question-answering sequences in both Spanish and English interviews.

Q17 was flagged four times for inducing more respondent turns, longer first turns by respondents, more interviewer turns, and longer question-answering sequences in Spanish interviews. It was flagged once in English interviews for having more overlapping speech. Q17 pertains to consent to record linkage and interviewers recorded the consent outcome.

Both Q16 and Q4 were flagged three times for poor performance in Spanish interviews and four times in English interviews. Q16 is a single-choice item with a showcard listing two answer options. Two outcome measures (number of interviewer turns and total duration) identified this item as one of the worst-performing items. Q4 is also a single-choice item, using a showcard with 16 answer options. Three outcome measures (number of respondent turns, duration of respondent's first turn, and long pauses) flagged this item as one of the worst.

Q6 asks respondents to select one from 11 response options. It was flagged three times for poor performance in both Spanish and English interviews, but for different reasons. Q6 induced more respondent turns, more interviewer turns, and longer question-answering sequences in Spanish interviews. By contrast, in English interviews, it induced longer first turns by respondents, more pauses, and fewer instances of positive emotions.

Both Q8 and Q12 were flagged three times in Spanish interviews and once in English interviews. Q8 is a check-all-that-apply question, and a showcard with eight options was provided to respondents. Q8 led to more respondent turns, more interviewer turns, and longer question-answering sequences in the Spanish interviews, while in English interviews, it resulted in more pauses. Q12 is a single-choice item with two answer options, and no showcard was used for this item. This item led to longer first turns for respondents in both English and Spanish interviews. In addition, there were more overlapping speech and fewer instances of positive emotions in Spanish interviews.

Q15 exhibited more problems in English interviews; it was flagged five times for having more respondent turns, longer first turns by respondents, more interviewer turns, longer question-answering sequences, and more overlapping speech. In contrast, it was flagged twice in Spanish interviews for having more overlapping speech and fewer instances of positive emotions. Q15 is a single-choice item that uses a showcard with two options.

Q3 is a closed-ended question that provides a showcard with 21 options for the respondent. It was more problematic in the English interviews, where it resulted in more respondent turns, more interviewer turns, longer question-answering sequences, more pauses, and fewer instances of positive emotions. In contrast, in Spanish interviews, it incurred longer respondent turns and more pauses.

Q5 and Q10 were flagged twice in both English and Spanish interviews. Q5 is a single-choice item with two response options. It incurred more pauses and fewer instances of positive emotions in both languages. Q10 is a single-choice item with two response options. However, it resulted in more pauses and fewer instances of positive emotions in Spanish interviews, while in English interviews, it led to more interviewer turns and longer question-answering sequences.

Three questions (Q1, Q9, and Q7) were flagged by one outcome measure in the Spanish interviews; Q1 and Q8 incurred more pauses while Q9 had fewer instances of positive emotions. However, none of these questions were flagged in the English interviews. Q1 is an open-ended question asking for a numeric response from the respondent. Both Q9 and Q7 are single-choice items with two response options, and no showcard was used for either.

Q2 is a single-choice item with two response options. It was not flagged as a problematic survey item in the Spanish interviews but was flagged twice in the English interviews for more overlapping speech and fewer instances of positive emotions.

Conclusions and Discussion

Yan, Sun, and Battalahalli (2024) applied machine learning to recordings from CARI and used seven outcome measures to identify questions at higher risk of poor performance. They found that the outcome measures effectively screened items that might not perform well when administered in English. This paper applied the same procedures to recordings of Spanish interviews and used the same outcome measures to flag questions with poorer performance. We compared the results for English and Spanish recordings.

The results are promising. First, as shown in the appendix figures, all seven outcome measures exhibited meaningful distributions in the Spanish recordings, qualifying them as a feasible triage tool for identifying survey items with poor performance. In our application, we used the ML pipeline to process all recordings and selected the five worst-performing items flagged by the ML pipeline for behavior coding. However, researchers and organizations using this tool have the flexibility to choose cutoff values, the number of items to flag, and how to handle the flagged items based on their goals and budget.

Second, survey items flagged in the Spanish interviews by at least two outcome measures for poor performance were also flagged in the English interviews as problematic by at least one outcome measure. Three items were flagged once as problematic in the Spanish interviews but were not found to be problematic in the English interviews. One item was flagged twice for poor performance in the English interviews, but not in the Spanish interviews. Despite the discrepancies between the English and Spanish interviews for these four items, it is particularly reassuring that survey items difficult due to their question structure (e.g., open-ended questions, questions with a long list of response options) were flagged by the pipeline as problematic. The results are sufficient for project teams to use these outcome measures to pre-select, target, and prioritize items for further investigation.

Third, the cost of applying the pipeline to Spanish interviews, in addition to English interviews, is marginal, given that the processing of recordings is automated. This makes the pipeline an ideal tool for multilingual surveys testing questions in more than one language.

In this study, we relied on the mean of each outcome measure across recordings to screen for questions with poor performance. Future research should also explore the possibility of using the spread of outcome measures across recordings to flag either survey items or survey interviewers exhibiting excessive variation for further examination. In addition, this study focused solely on the capability of the ML pipeline to flag problematic survey items for both English and Spanish interviews, but future research can easily extend it to evaluate behaviors of interviewers administering interviews in Spanish, following the same procedures described by Sun and Yan (2023).

Overall, the results demonstrate the potential of using machine learning to triage survey items in a multilingual survey with items worded in and administered in different languages. As mentioned in Yan, Sun, and Battalahalli (2024), the strength of this ML pipeline is its ability to automatically, efficiently, and inexpensively process recordings of interviews in real time, regardless of whether the interviews are conducted in person, over the phone, or virtually via Zoom. We now demonstrate that the same procedures can be applied to Spanish interviews, making it a powerful tool for any multilingual surveys.

We recommend that researchers and practitioners add the use of machine learning to their toolkit for question evaluation, using it in combination with other question evaluation methods. This approach aligns with the recommendation from Tourangeau et al. (2020) that employing a combination of question evaluation methods is more effective than relying on a single method. We suggest utilizing machine learning as an initial triage step to identify questions and/or interviewers with a high risk of poor performance for further evaluation to improve efficiency. We also recommend that researchers and practitioners leverage the automation of this machine learning approach to evaluate and monitor the performance of survey questions and interviewers before, during, and potentially after data collection. Researchers and practitioners can additionally capitalize on its flexibility to determine the scope and nature of the follow-up activities based on their budget and timeline.

Leading Author's contact information

Ting Yan
55 East Monroe St, 30th Floor, Chicago, IL 60603, (312) 759-4000
yan-ting@norc.org

Submitted: September 17, 2024 EST. Accepted: November 02, 2024 EST. Published: March 02, 2025 EST.

REFERENCES

Ji, Li-Jun, Norbert Schwarz, and Richard E. Nisbett. 2000. "Culture, Autobiographical Memory, and Behavioral Frequency Reports: Measurement Issues in Cross-Cultural Studies." *Personality and Social Psychology Bulletin* 26 (5): 585–93. https://doi.org/10.1177/0146167200267006.

Johnson, Timothy P., Diane O'Rourke, Noel Chavez, Seymour Sudman, Richard B. Warnecke, Loretta Lacey, and John Horm. 1996. "Cultural Variations in the Interpretation of Health Survey Questions." In *Health Survey Research Methods Conference Proceedings*, 57–62. Hyattsville, MD: National Center for Health Statistics.

Lee, Sunghee, Norbert Schwarz, and Leanne Streja Goldstein. 2014. "Culture-Sensitive Question Order Effects of Self-Rated Health between Older Hispanic and Non-Hispanic Adults in the United States." *Journal of Aging and Health* 26 (5): 860–83. https://doi.org/10.1177/0898264314532688.

Sun, Hanyu, and Ting Yan. 2023. "Applying Machine Learning to the Evaluation of Interviewer Performance." *Survey Practice* 16 (1). https://doi.org/10.29115/SP-2023-0007.

Tourangeau, Roger, Aaron Maitland, Darby Steiger, and Ting Yan. 2020. "A Framework for Making Decisions about Question Evaluation Methods." In *Advances in Questionnaire Design, Development, Evaluation and Testing*, edited by P. Beatty, D. Collins, L. Kaye, J. Padilla, G. Willis, and A. Wilmot, 47–73. Hoboken, NJ: Wiley.

Yan, Ting, and Mengyao Hu. 2019. "Examining Translation and Respondents' Use of Response Scales in 3MC Surveys." In *Advances in Comparative Survey Methods: Mulitnational, Multiregional, and Multicultural Contexts (3MC)*, edited by T.P. Johnson, B. Pennell, I. Stoop, and B. Dorer, 501–18. Hoboken, NJ: Wiley.

Yan, Ting, Hanyu Sun, and Anil Battalahalli. 2024. "Applying Machine Learning to Survey Question Assessment." *Survey Practice*. https://doi.org/10.29115/SP-2024-0006.

SUPPLEMENTARY MATERIALS

Appendix

Download: https://www.surveypractice.org/article/125777-using-machine-learning-to-evaluate-questions-in-a-multilingual-survey/attachment/253088.docx