Comparison of Text-To-Speech with Human Voice Recordings on Comprehension of Survey Questions in Audio Computer-Assisted Self Interviewing

Emily McFarlane Geisen, RTI International
Gretchen McHenry, RTI International
Patty LeBaron, RTI International
Gil Rodriguez, RTI International
Grace Medley, SAMHSA
Peggy Barker, SAMHSA
Dicy Painter, SAMHSA
Joel Kennet, SAMHSA
Acknowledgements

The presentation is sponsored by RTI International’s Survey Research Division, with the research included in the presentation stemming from ongoing methodological work conducted under the contract for the National Survey on Drug Use and Health (NSDUH). The NSDUH is funded by the Substance Abuse and Mental Health Services Administration (SAMHSA), Center for Behavioral Health Statistics and Quality under contract no. 283-2010-00003C and project no. 0212800.

The views expressed in this presentation do not necessarily reflect the official position or policies of SAMHSA or the U.S. Department of Health and Human Services; nor does mention of trade names, commercial practices, or organizations imply endorsement by the U.S. Government.
About the NSDUH

- The NSDUH provides national, state and substate data on substance use and mental health in the civilian, noninstitutionalized population age 12 and older.
- Data are collected on a quarterly basis each year.
- Approximately 700 field interviewers (FIs) staffed.
- Approximately 140,000 household screenings and 67,500 interviews completed annually.
- Conducted by RTI under contract with SAMHSA.
- Conducted in person using a CAPI survey with ACASI components.
- Through 2014, the ACASI used human voice recordings in English and Spanish.
Purpose of the Study

- Determine effects of replacing human ACASI voice with text-to-speech (TTS) voice

- Will transitioning affect respondent preference or comprehension?

- Will respondents prefer voice at slow or moderate speed?
In 2013, NSDUH investigated the feasibility of replacing ACASI human audio with Text to Speech audio.

- NSDUH contains 40 modules and 3000 screens.
- English and Spanish administration
- Transition to TTS has a number of advantages
- Microsoft Speech Platform product was chosen
Advantages of a Text to Speech Approach

- Eliminate dependence on a single human “reader” and enable consistency of voice over time.
- Flexibility in changing the wording/ordering of survey questions and responses, since audio is generated dynamically for each screen.
- Pronunciations can be updated “on the fly” quickly and easily by modifying a pronunciation file (text file).
- Smoother delivery of audio, especially where human voice relies on fills.
- Cost and process efficiencies in ACASI development.
- Eliminate management of audio files, but must manage customized pronunciations.
Concerns about TTS Transition and Cognition

- Respondents to the NSDUH are diverse in age, education, SES, primary language, and reading ability.
- This diversity leads to variability on how much respondents rely on the audio components of the questionnaire administration.
- Comprehensibility of TTS in NSDUH has not been evaluated by individuals who fall into these demographic subgroups.
- We hypothesize that these individuals are likely to be the youngest and oldest respondents (i.e. those aged 12-17 and 65 or older), as well as respondents with low levels of literacy and non-native English speakers.)
Purpose of the Cognitive Interviews

- Evaluate whether the TTS delivery has any effect on comprehension of specific survey items or administration of those items compared with the human voice
- Determine the best TTS presentation speed for administering the survey questions,
- Evaluate participants’ preferences and opinions regarding the voices used for ACASI
- Identify any major issues in administration time or any unanticipated issues with the use of TTS.
Recruited a convenience sample of cognitive interview participants

- Participants age 12-17
- Participants age 65 and older
- Participants with low levels of literacy
- Non-native English speakers who would complete the interview in English

- Participants were recruited through word of mouth, Craigslist, and a literacy advocacy organization
- 36 interviews were conducted: 24 interviews in English and 12 in Spanish
- Spanish speakers had limited or no English speaking ability
- Interviews were conducted in Chicago, IL, Washington, DC and Research Triangle Park, NC
During the interview, participants listened to survey questions read in three voices. Participants relied only on voice. They could not see the question text.

- Human voice
- TTS moderate speed
- TTS slow speed

- Participants were assigned to listen to the voices in different orders
- Instructed to repeat the question or paraphrase the question to assess comprehension
- After listening to a particular voice, participants answered debriefing items
- Debriefing items measured preference for each voice
Methodology: Rating System

Example Debriefing Item

How would you rate the speed or pace of the voice that read the interview questions? Would you say the pace of the voice was much too slow, a little too slow, just right, a little too fast, or much too fast?

- Participants were asked to rate each voice (TTS moderate, TTS slow, and human) on a number of characteristics.
- Participants rated the cadence, the pace, the pronunciation, comprehension, pleasantness, and overall quality of each voice.
- Comfort of listening to the voice for 30 minutes was also rated.
- Coding was reversed when necessary so lower scores were more favorable.
Results: Preference in English

Preferences for English-Speaking Voice
*lower ratings are more favorable
Results: Preference in Spanish

Preferences for Spanish-Speaking Voice
*lower ratings are more favorable

![Bar chart showing preferences for Spanish-speaking voices.](chart.png)
Responses among participants in the Spanish cognitive interviews and the English cognitive interviews did not differ very much.

- Participants rated the human voice as slightly better than either of the TTS voices
- Participants would be more comfortable listening to the moderate speed voice for 30 minutes than the slow speed
- The TTS Slow English voice was rated the lowest on the pleasant scale
Results: Comprehension

Comprehension Ratings by Set and Language
1=poor, 2=fair, 3=good

- English: Order 1, Order 2, Order 3
- Spanish: Order 1, Order 2, Order 3

Human TTS Moderate

Results: Comprehension

- Interviewers rated the participants’ comprehension based upon their ability to summarize the survey question.
- Participant understanding did not differ by voice or language.
- Generally, participants understood question meaning or intent.

Average Comprehension

Scale: 1 (Poor) – 3 (Good)

2.5 ENGLISH

2.7 SPANISH
Summary and Conclusions

- Participants deemed the TTS voices *pleasant* and *understandable*
- There were no major differences between comprehension for the English items and the Spanish items
- Several participants noted that the moderate speed was more difficult to understand
- Prioritize comprehension over preference, as respondent understanding has largest potential for impact on measurement error.
- The results of the cognitive testing and field test suggest that TTS could be implemented in the 2015 NSDUH without serious impacts on data quality or overall instrument timing
- The TTS Slow voice was implemented on the 2015 NSDUH
Contact Information
Patty LeBaron
312.777.5204
plebaron@rti.org

rti.org/aapor