Michael Schober
Vice Provost for Research
Professor of Psychology

INTERACTING WITH INTERVIEWERS IN VOICE AND TEXT INTERVIEWS ON SMARTPHONES

Michael F. Schober
Frederick G. Conrad
Christopher Antoun
Alison W. Bowers
Andrew L. Hupp
H. Yanna Yan

UNIVERSITY OF MICHIGAN

THE NEW SCHOOL

Interviewers and Their Effects from a Total Survey Error Perspective Workshop
University of Nebraska-Lincoln
February 26-28, 2019

ACKNOWLEDGMENTS

- NSF grants SES-1026225 and SES-1025645 to Frederick Conrad and Michael Schober
- Collaborators at The New School: Stefanie Fail, Courtney Kellner, Kelly Nichols, Leif Percifield, Lucas Vickers
- Collaborators at University of Michigan: Monique Kelly, Mingnan Liu, Chan Zhang
- Collaborators (formerly) at AT&T Research Labs: Patrick Ehlen, Michael Johnston

HOW INTERVIEWERS INTERACT WITH RESPONDENTS IS EVOLVING

- Many more options for Rs beyond FTF and landline phone
- Phone Rs more and more likely to be mobile and multitasking
- Landscape of Rs' (non-survey) communicative habits transforming
  - People more and more likely to use and switch between multiple modes (text, voice, video, email) on same device
    - choosing mode appropriate to current setting, goals, needs, interlocutor
  - People more and more used to human-machine interactions
    - ATMs, ticket kiosks, self-check-out at grocery store
    - Automated phone agents who route and respond to calls for, e.g., travel reservations, tech support
    - Online help "chat" with bot
    - Etc.

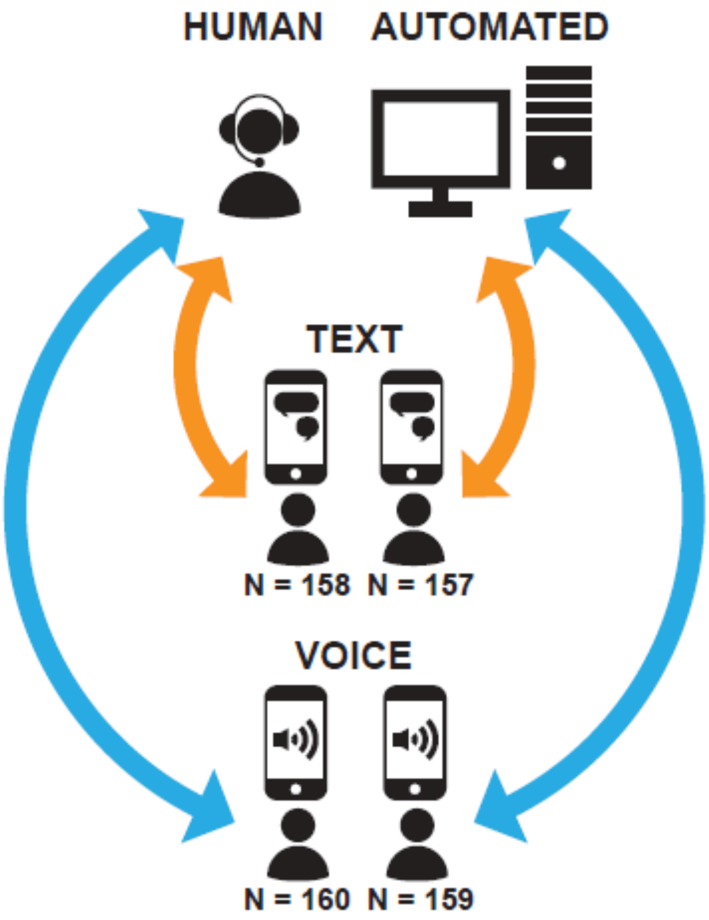
NEW QUESTIONS ABOUT INTERVIEWERS AND THEIR EFFECTS

- In traditional survey modes, how are these transformations changing effects of interviewers?
  - E.g., as more Rs choose text or video for both informal and transactional purposes, and avoid answering incoming calls, how will they treat FTF or phone interviews?
- What are potential effects of interviewers—positive and negative—in popular communication modes not yet widely deployed for surveys (e.g., texting, video)?
  - E.g., will interviewers enhance participation and R motivation?
  - E.g., will interviewers reduce Rs' willingness to disclose sensitive info?
- How will automated "interviews" in this new landscape compare with human-administered interviews?
  - And will differences be greater in some modes than others?

CURRENT STUDY

- Explores dynamics of interviewer-respondent interaction in corpus of interviews
- Four existing or plausible survey modes that work through native apps on the iPhone
    - As opposed to specially designed survey apps
    - As opposed to web survey in phone's browser
    - Uniform interface for all Rs
        - As opposed to mix of platforms (Android, Windows, etc.)

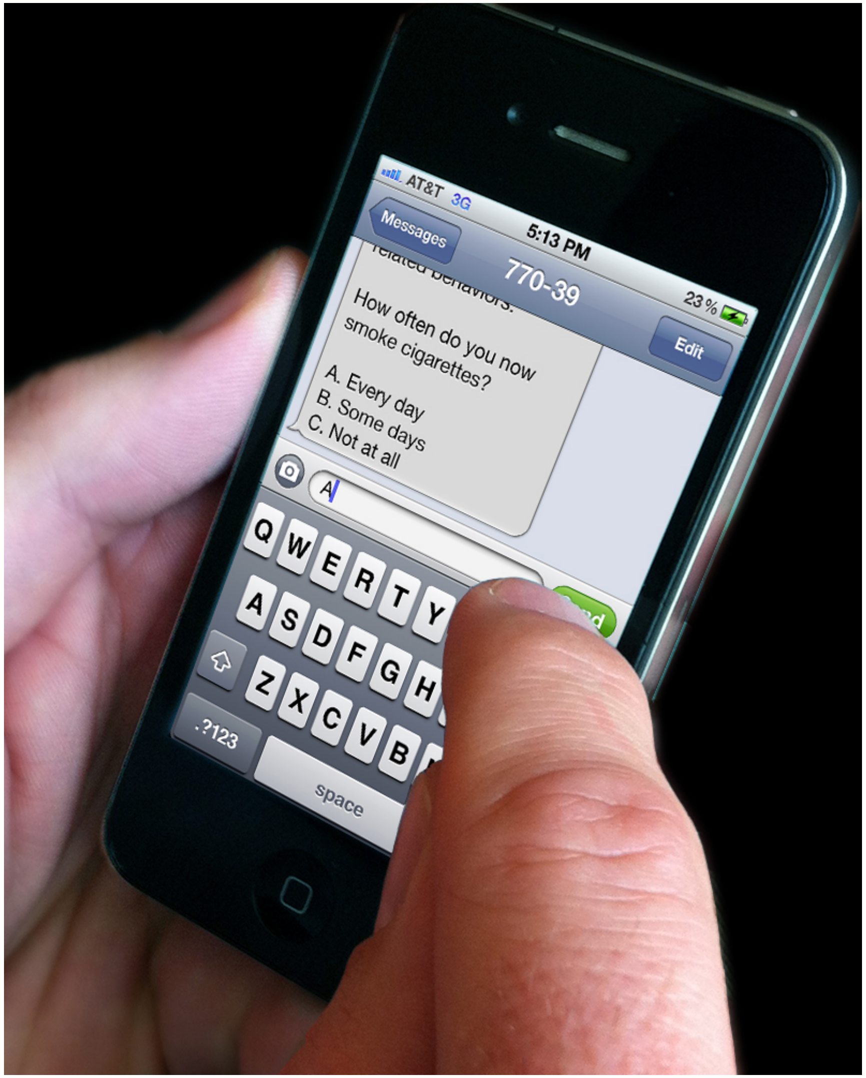
SCHOBER ET AL., 2015: EXPERIMENTAL DESIGN



- 4 Modes on iPhone:
  - Human Voice
  - Human Text (SMS)
  - Automated Voice
  - Automated Text (SMS)
- 32 Q's from ongoing US surveys
- *R*s (convenience sample) screened in
  - age ≥ 21; US area code
  - $20 iTunes gift code

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0128337

TEXT RESPONDENT



related behaviors.

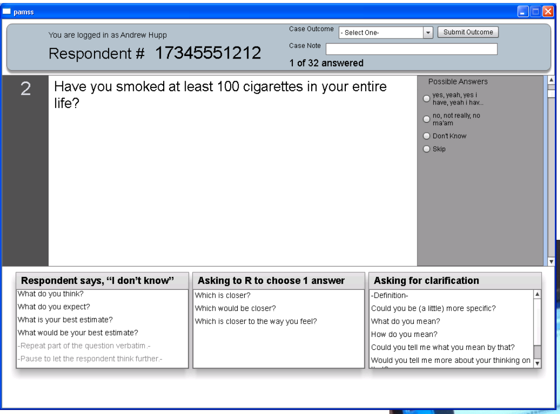
How often do you now smoke cigarettes?

A. Every day
B. Some days
C. Not at all

A

IMPLEMENTATION: HUMAN VOICE

- 8 interviewers (*I*s) from U Mich survey research center
- custom designed CATI interface that supports voice and text interviews



You are logged in as Andrew Hupp

Respondent # 17345551212

Case Outcome: - Select One-

Case Note:

1 of 32 answered

2 Have you smoked at least 100 cigarettes in your entire life?

Possible Answers

○ yes, yeah, yes i have, yeah i hav...
○ no, not really, no ma'am
○ Don't Know
○ Skip

Respondent says, "I don't know"

What do you think?
What do you expect?
What is your best estimate?
What would be your best estimate?
-Repeat part of the question verbatim.-
-Pause to let the respondent think further.-

Asking to R to choose 1 answer

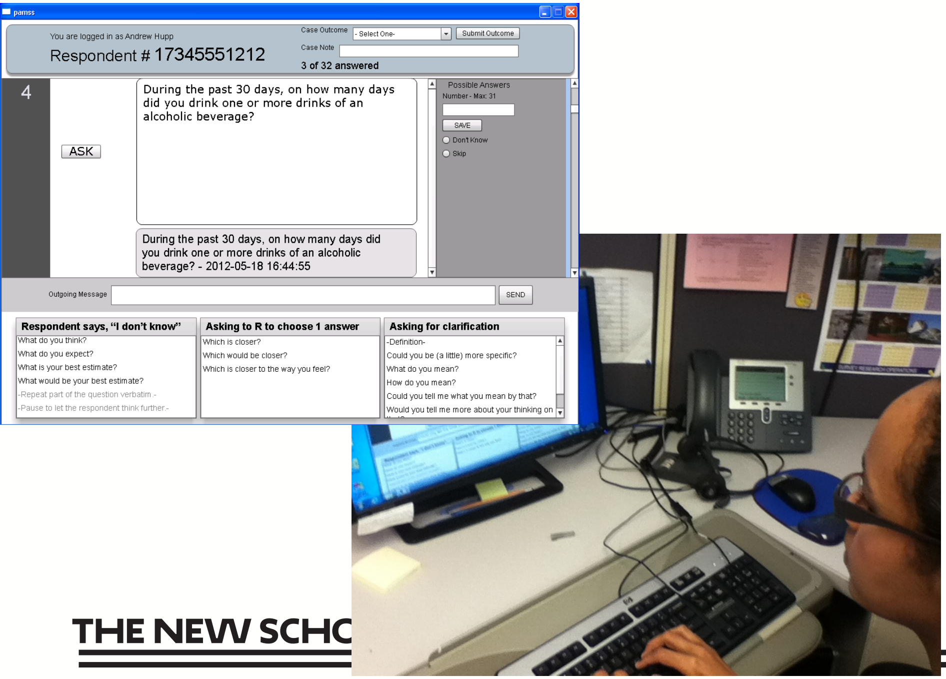
Which is closer?
Which would be closer?
Which is closer to the way you feel?

Asking for clarification

-Definition-
Could you be (a little) more specific?
What do you mean?
How do you mean?
Could you tell me what you mean by that?
Would you tell me more about your thinking on

IMPLEMENTATION: HUMAN TEXT

- Same 8 *I*s from U Mich survey research center
- Same custom designed CATI interface
  - *I* selects, edits, or types (personalizes) questions/prompts, and clicks to send
- Text messages sent through third party (Aerialink)
- Rs can answer with single character: Y/N, letter (a/b/c), or number



You are logged in as Andrew Hupp

Respondent # 17345551212

Case Outcome: - Select One-

Case Note:

3 of 32 answered

4

During the past 30 days, on how many days did you drink one or more drinks of an alcoholic beverage?

During the past 30 days, on how many days did you drink one or more drinks of an alcoholic beverage? - 2012-05-18 16:44:55

Possible Answers

Number - Max: 31

○ Don't Know

○ Skip

Outgoing Message:

Respondent says, "I don't know"

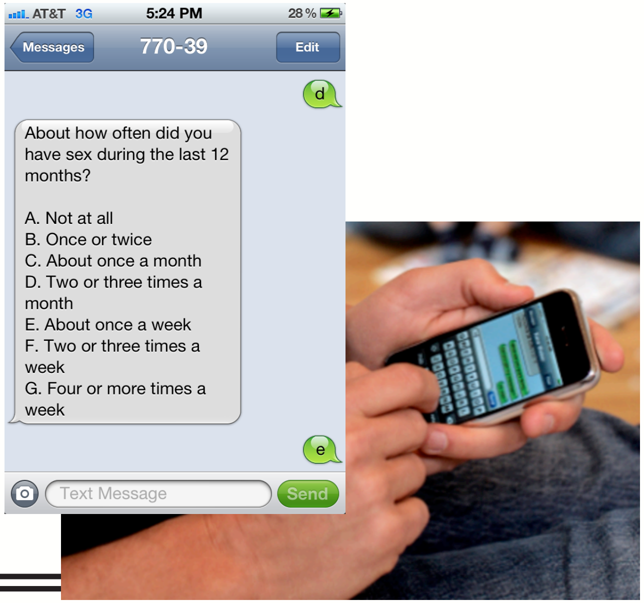
What do you think?
What do you expect?
What is your best estimate?
What would be your best estimate?
-Repeat part of the question verbatim.-
-Pause to let the respondent think further.-

Asking to R to choose 1 answer

Which is closer?
Which would be closer?
Which is closer to the way you feel?

Asking for clarification

-Definition-
Could you be (a little) more specific?
What do you mean?
How do you mean?
Could you tell me what you mean by that?
Would you tell me more about your thinking on



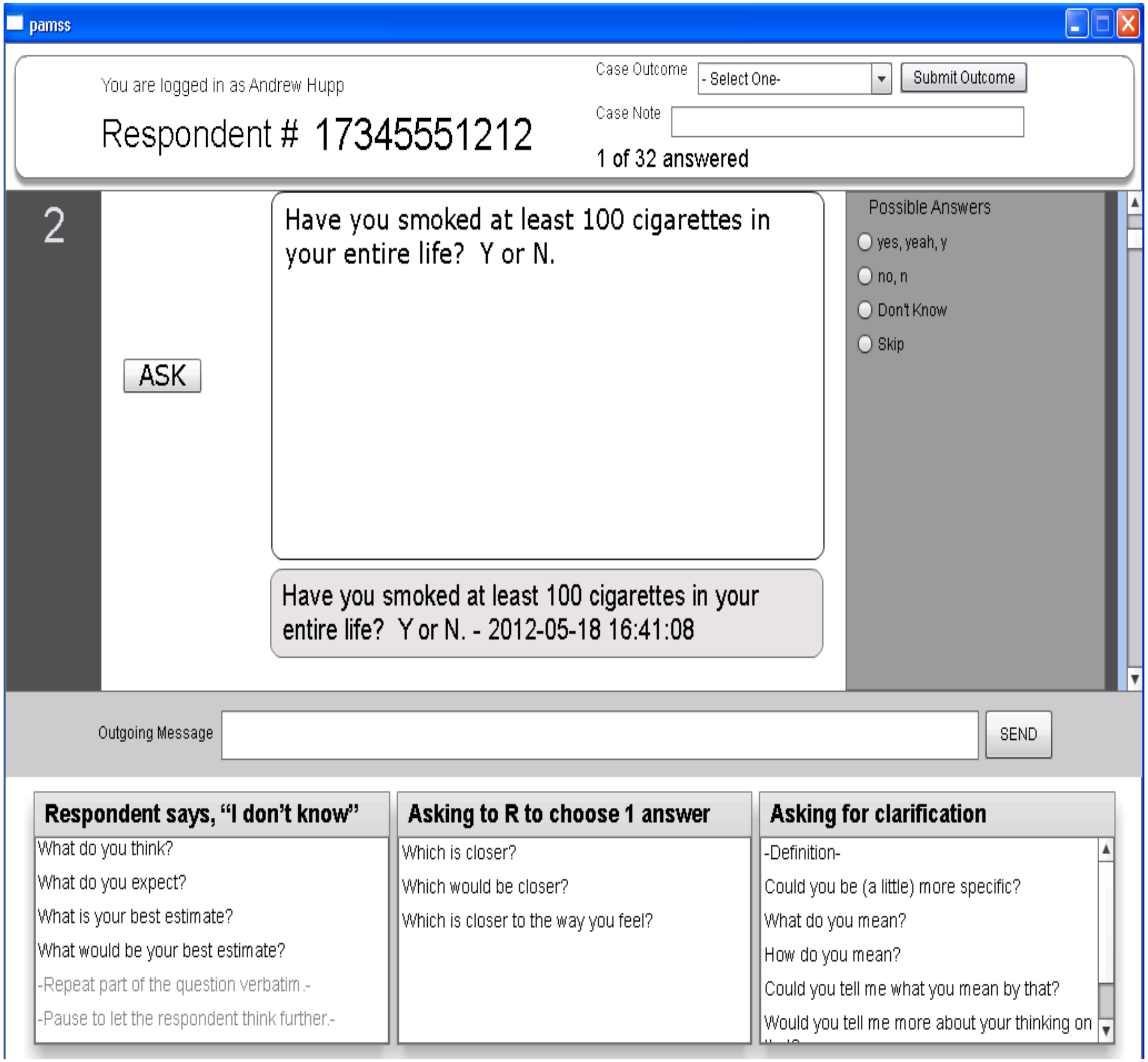
d

About how often did you have sex during the last 12 months?

A. Not at all
B. Once or twice
C. About once a month
D. Two or three times a month
E. About once a week
F. Two or three times a week
G. Four or more times a week

e

HUMAN TEXT INTERVIEWER INTERFACE



You are logged in as Andrew Hupp

Respondent # 17345551212

Case Outcome: - Select One-

Case Note:

1 of 32 answered

2

Have you smoked at least 100 cigarettes in your entire life? Y or N.

Have you smoked at least 100 cigarettes in your entire life? Y or N. - 2012-05-18 16:41:08

Possible Answers

- ○ yes, yeah, y
- ○ no, n
- ○ Don't Know
- ○ Skip

Outgoing Message:

Respondent says, "I don't know"

What do you think?
What do you expect?
What is your best estimate?
What would be your best estimate?
-Repeat part of the question verbatim.-
-Pause to let the respondent think further.-

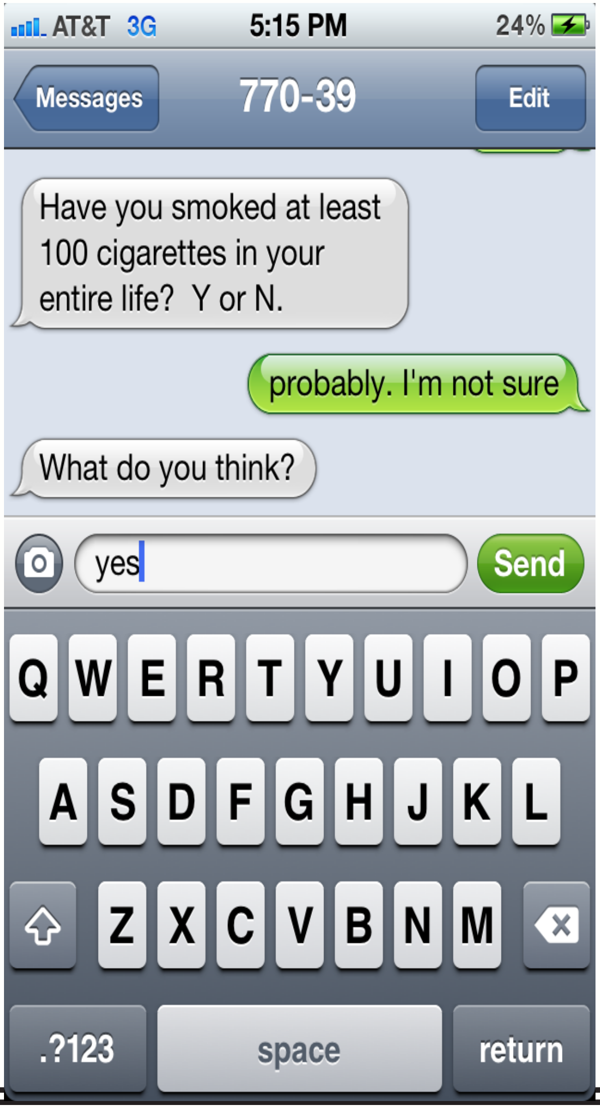
Asking to R to choose 1 answer

Which is closer?
Which would be closer?
Which is closer to the way you feel?

Asking for clarification

-Definition-
Could you be (a little) more specific?
What do you mean?
How do you mean?
Could you tell me what you mean by that?
Would you tell me more about your thinking on

TEXT INTERACTION BETWEEN HUMAN INTERVIEWER AND RESPONDENT



Have you smoked at least 100 cigarettes in your entire life? Y or N.

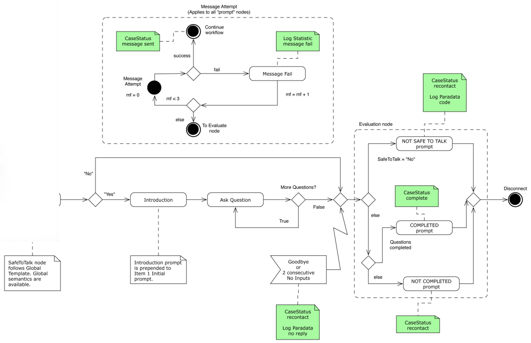
probably. I'm not sure

What do you think?

yes

IMPLEMENTATION: SPEECH IVR

- Custom built speech dialogue system
- Uses ATT's Watson speech recognizer, Asterisk telephony gateway
- Recorded human interviewer, speech responses (not touchtone)



EXAMPLES FROM AUTOMATED VOICE

Categorical

I: How often do you read the newspaper? 'Every day', 'a few times a week', 'once a week', 'less than once a week', or 'never'?

R: Every day

I: Got it.

Numerical

I: Thinking about the time since your eighteenth birthday (including the recent past that you've already told us about), how many male partners have you had sex with?

R: None → First Hypothesis: "Nine"

I: I think you said '9'. Is that right? Yes or No.

R: No

Explicit Confirmation

I: Thinking about the time since your eighteenth birthday (including the recent past that you've already told us about), how many male partners have you had sex with?

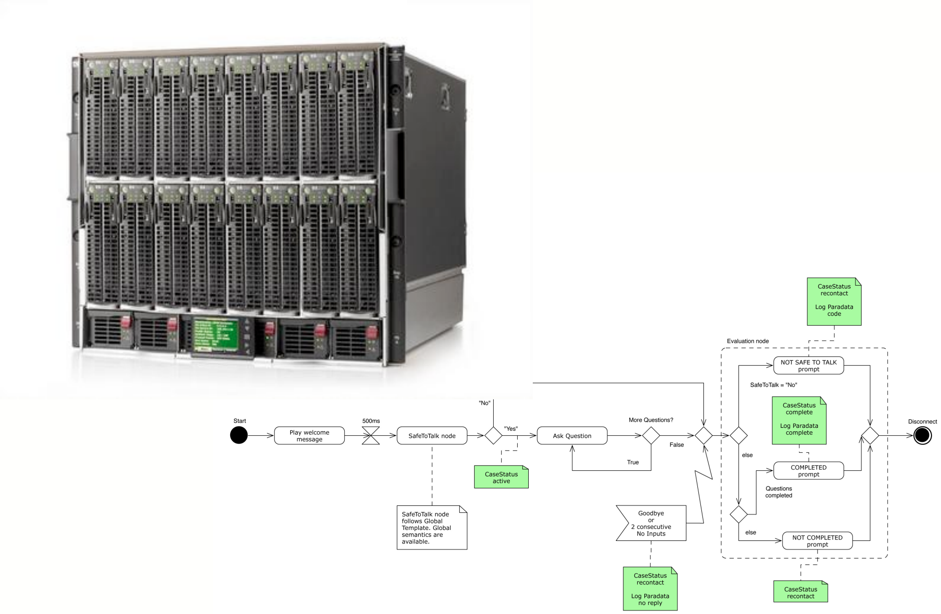
R: Zero → Last Hypothesis: "Zero"

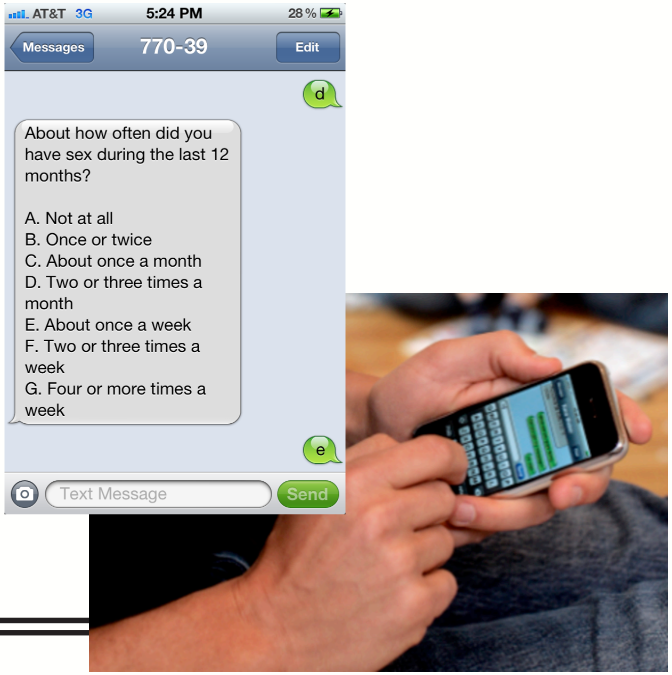
I: Thanks

Last Annotation: "Zero"

IMPLEMENTATION: AUTO-TEXT

- Custom built text dialogue system
- Text messages sent through third party (Aerialink)
- Rs can answer with single character: Y/N, letter (a/b/c), or number





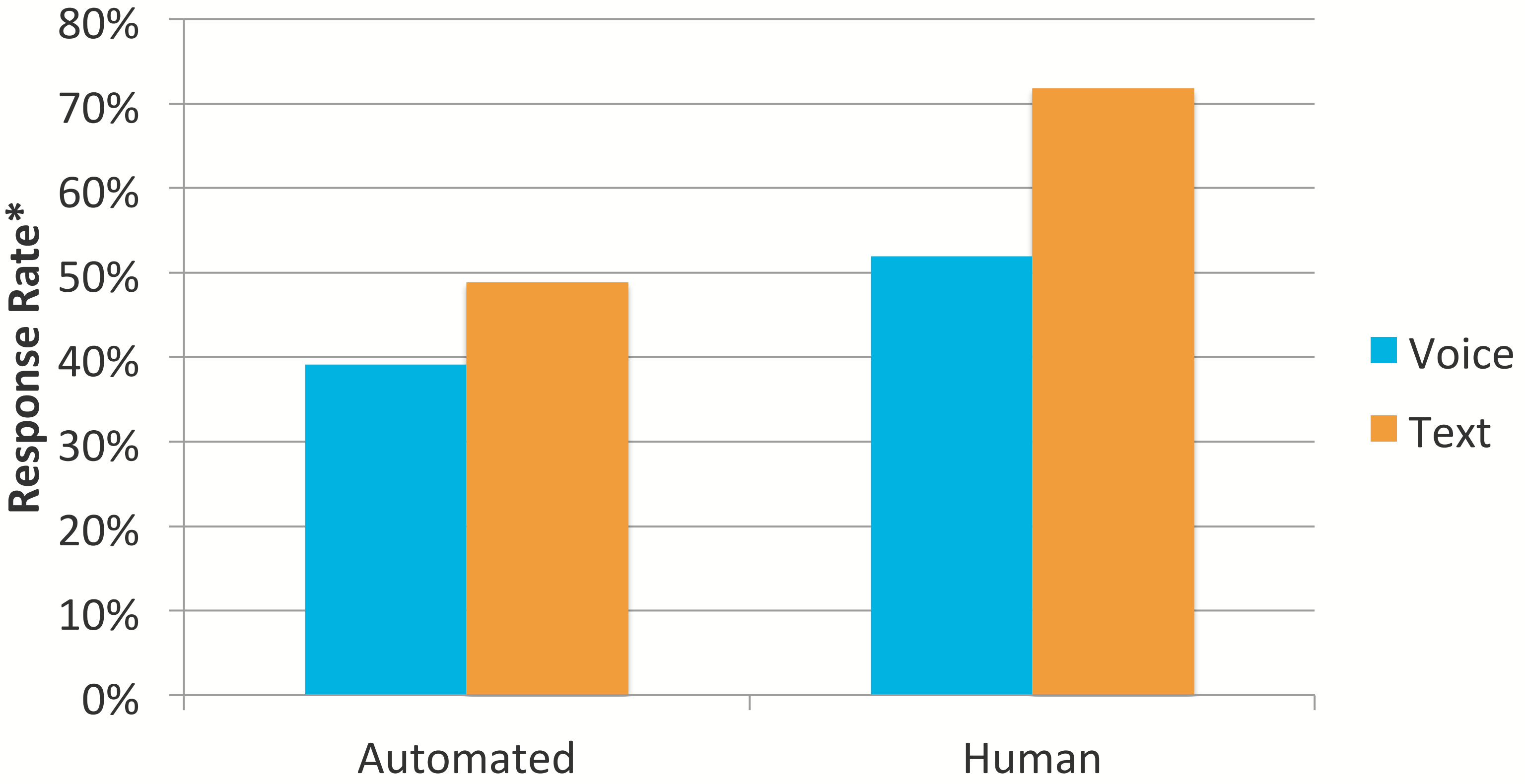
d

About how often did you have sex during the last 12 months?

A. Not at all
B. Once or twice
C. About once a month
D. Two or three times a month
E. About once a week
F. Two or three times a week
G. Four or more times a week

e

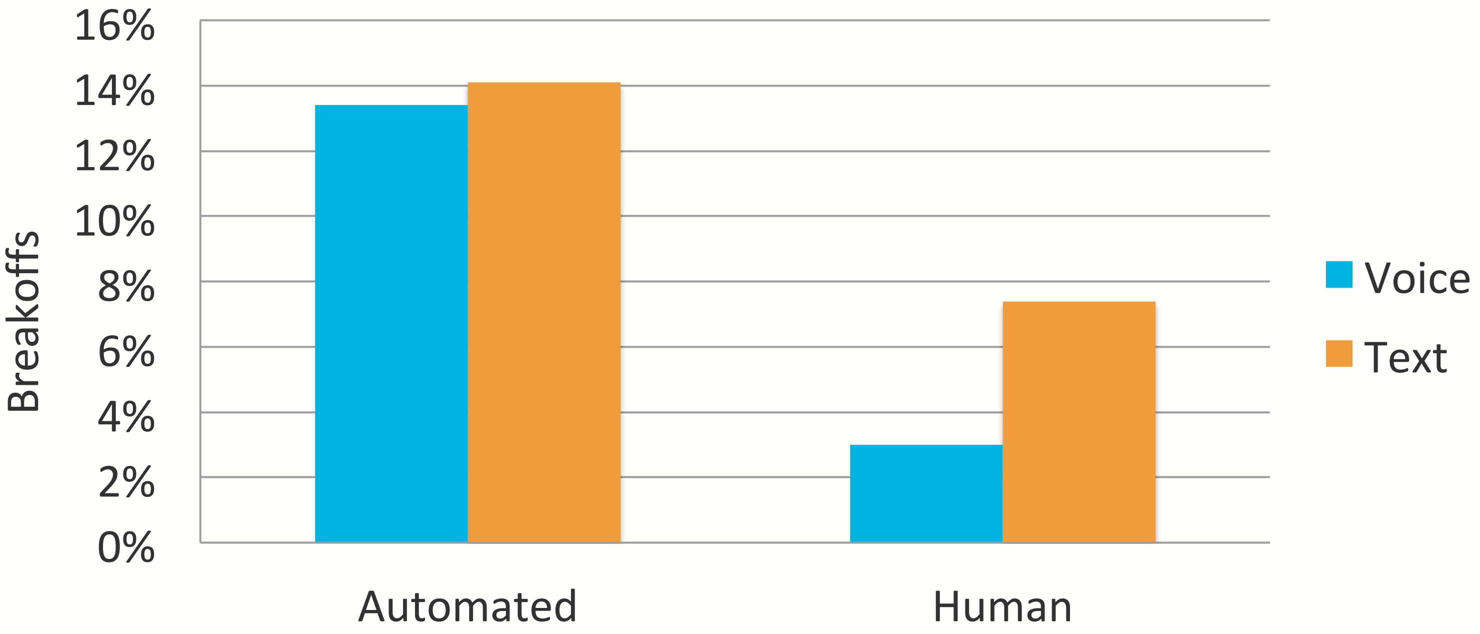
RESPONSE RATES* ACROSS MODES



- Higher response rate in text could be due to (1) persistence of invitation (different kind of noncontact), (2) ability to respond when convenient, (3) more time to decide

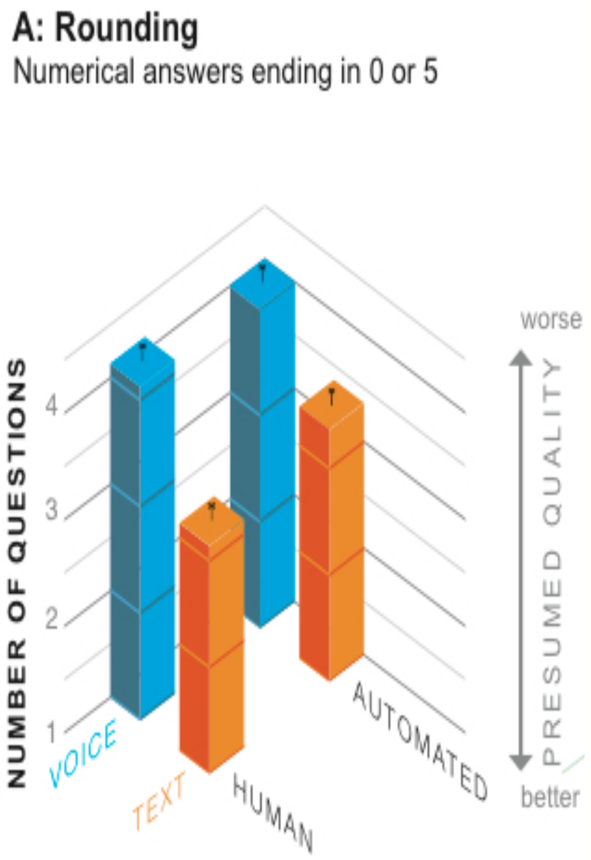
*AAPOR RR1: # complete interviews / # invitations

BREAKOFFS ACROSS MODES



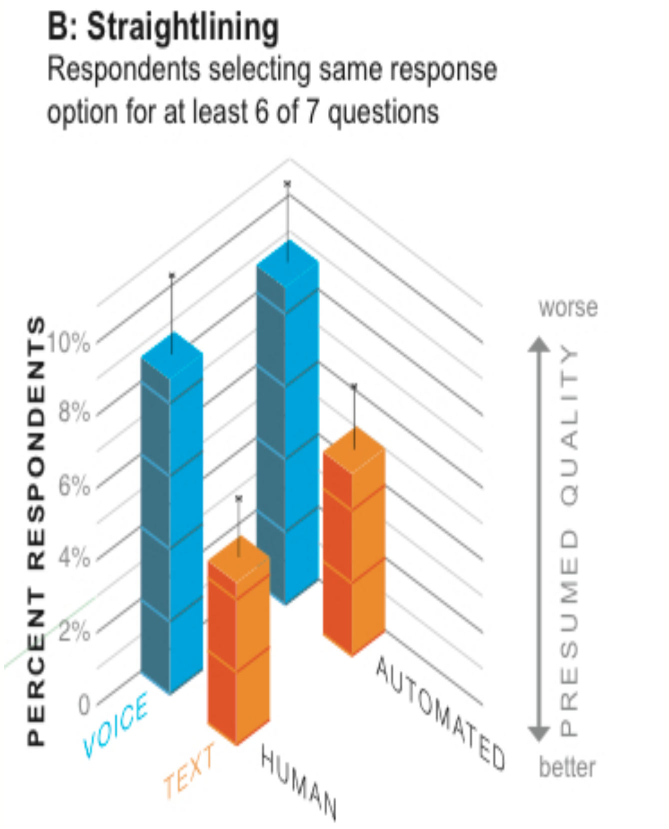
- More breakoffs in Text could be due to (1) no human voice to keep *R*s engaged, and (2) asynchronous character reducing need to answer *Q*s quickly ··· or ever
- Despite more breakoffs in text, response rates (starting and finishing) are higher in text interviews
- Substantially higher breakoff rates in Automated than Human modes likely due to absence of human interviewer

TEXT VS. VOICE: SATISFICING



A: Rounding

Numerical answers ending in 0 or 5



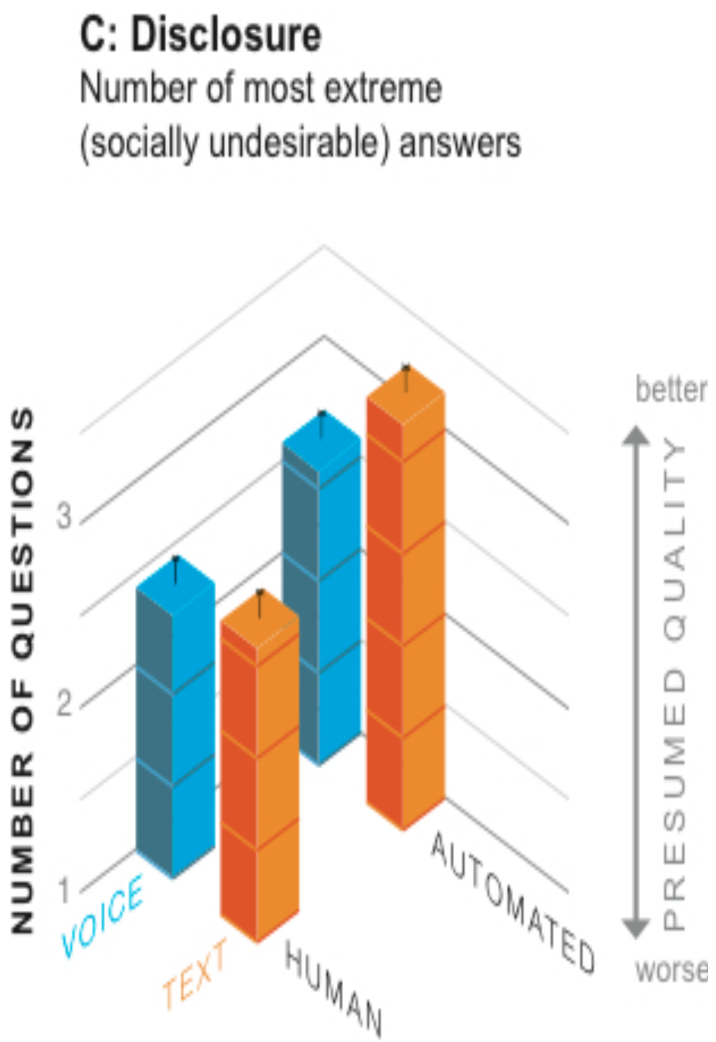
B: Straightlining

Respondents selecting same response option for at least 6 of 7 questions

TEXT VS. VOICE: DISCLOSURE

TEXT VS VOICE

- Similar pattern reported in West et al.'s (2015) study in Nepal
- Suggests greater disclosure in text is robust across populations and implementation



AUTOMATED VS HUMAN-ADMINISTERED

- Replicates widely-observed finding of greater disclosure in self- than interviewer-administration (e.g., Tourangeau & Smith, 1996)

WHAT ACCOUNTS FOR TEXT VS. VOICE DIFFERENCES IN PRECISION AND DISCLOSURE?

- Could be any or all of the many differences in timing and behavior between text and voice interviews
  - alone or in combination
- Plausible contributing factors include:
  - Text reduces immediate time pressure to respond, so R has more time to think or look up answers

    → Could explain greater precision (less rounding) in text
  - Text reduces "social presence"
    - Reduced salience of I's ability to evaluate or be judgmental?
    - No immediate evidence of I's reaction?

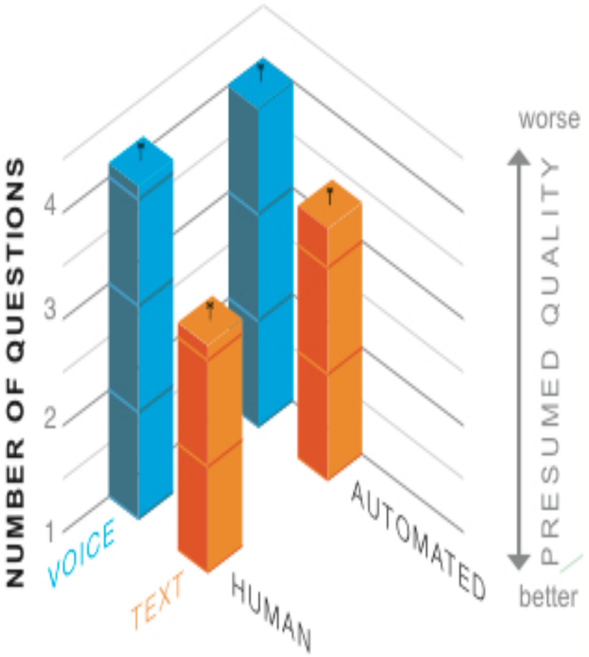
    → Could explain more disclosure in text

EXPERIMENTAL DESIGN HELPS RULE IN OR RULE OUT ACCOUNTS

- e.g., maybe R’s round less in text because text I’s never laugh (no *LOL*’s or *haha*’s)
  - – Maybe laughter in voice interviews suggests that casual responses are sufficient
  - – But that can’t be it because R’s round just as much in Human and Auto Voice interviews, and automated “interviewer” never laughed

A: Rounding

Numerical answers ending in 0 or 5



EXAMPLES: HUMAN TEXT VS. HUMAN VOICE INTERACTIONS

HUMAN TEXT				HUMAN VOICE		
1	I:	During the last month how many movies did you watch in any medium?		1	I:	During the last month, how many movies did you watch in ANY medium.
2	R:	3		2	R:	OH, GOD. U:h man. That's a lot. How many movies I seen? Like 30.
				3	I:	30.
Total elapsed time until next Q:						
1:21				**0:12**		

EXAMPLES: HUMAN TEXT VS. HUMAN VOICE INTERACTIONS

HUMAN TEXT		
1	I:	During the last month how many movies did you watch in any medium?
2	R:	Medium?
3	I:	Here's more information. Please count movies you watched in theaters or any device including computers, tablets such as an iPad, smart phones such as an iPhone, handhelds such as iPods, as well as on TV through broadcast, cable, DVD, or pay-per-view.
4	R:	3
Total elapsed time until next Q:		
2:00		

HUMAN VOICE		
1	I:	*During the last*
2	R:	Huh?
3	I:	Oh, sorry. Um, during the last month, how many movies did you watch in ANY medium.
4	R:	Oh! Let's see, what did I watch. Um, should I say how many movies I watched or how many movies watched me? [laughs] All right let's-let me think about that. I think yesterday I watched u:m, not in its entirety but you know, coming and going. My kids are watching in. Um, I don't know maybe 2 or 3 times a week maybe?

EXAMPLES: HUMAN TEXT VS. HUMAN VOICE INTERACTIONS

HUMAN VOICE		
5	I:	Uh, so what would be your best estimate on how many, um, you saw in the whole month.
6	R:	[pause] Um, I don't know I'd say maybe 3 movies if that many.
7	I:	3?
8	R:	Is that going to the movies or watching the movies on tv. Like you said *any medium* right?
9	I:	That's *any movies.* Yep.
10	R:	Maybe 1 or 2 a month I'd say.
11	I:	1 or 2 a month? [breath] Uh, so what would be *closer*

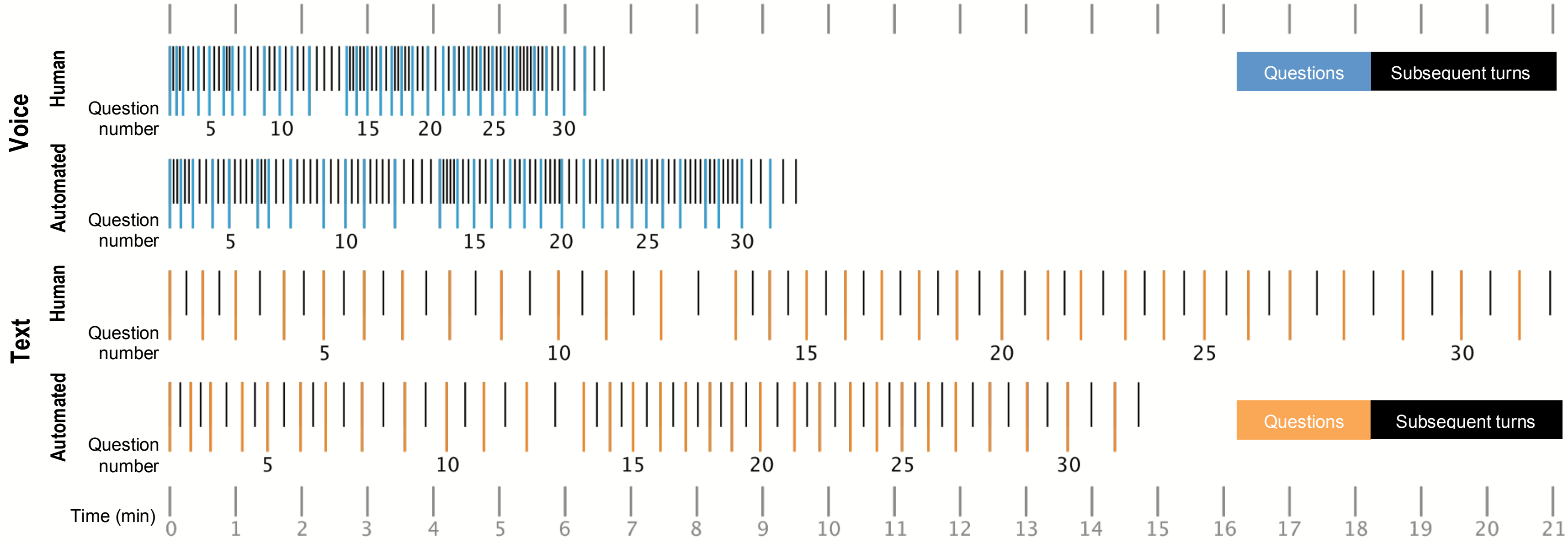
T

EXAMPLES: HUMAN TEXT VS. HUMAN VOICE INTERACTIONS

Total elapsed time until next Q: ←		
1:36		

HUMAN VOICE		
12	R:	*Yeah, because* I uh, um, occasionally I take the kids on a Tuesday to see a movie, depending on what's playing. So I'd maybe once or twice a month
13	I:	Which would be closer, once or twice.
14	R:	I would say twice.
15	I:	Twice?
16	R:	R: Mhm. Because it runs 4 Tuesdays which is cheaper to go
17	I:	Right
18	R:	R: so I'd say twice, yah. Because I do take them twice. Not last month but the month before

INTERVIEW DYNAMICS: TIMING

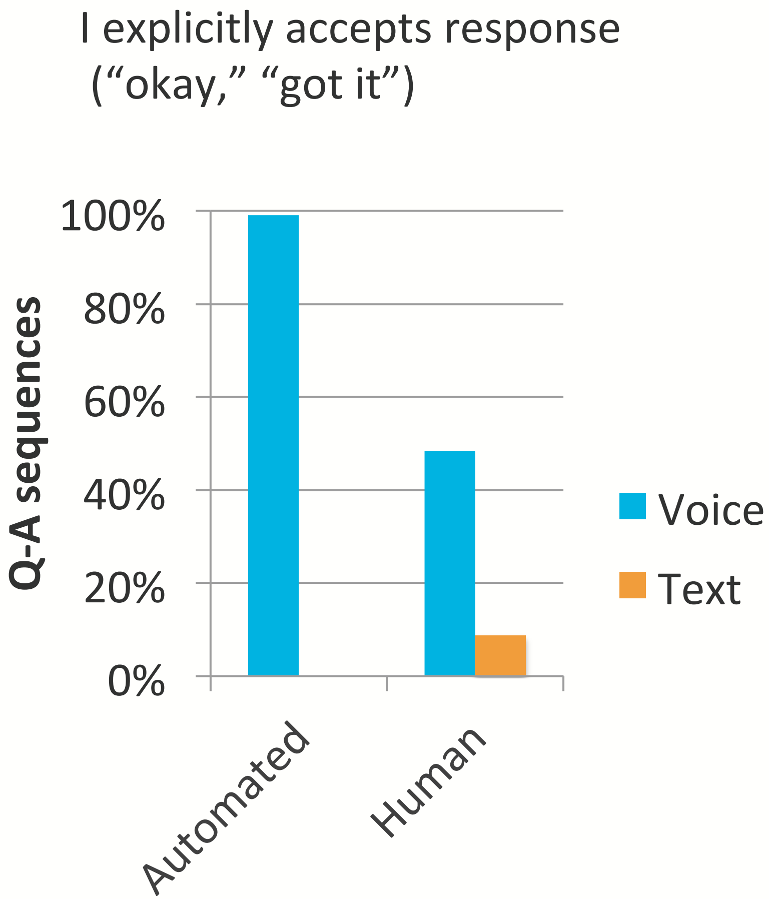


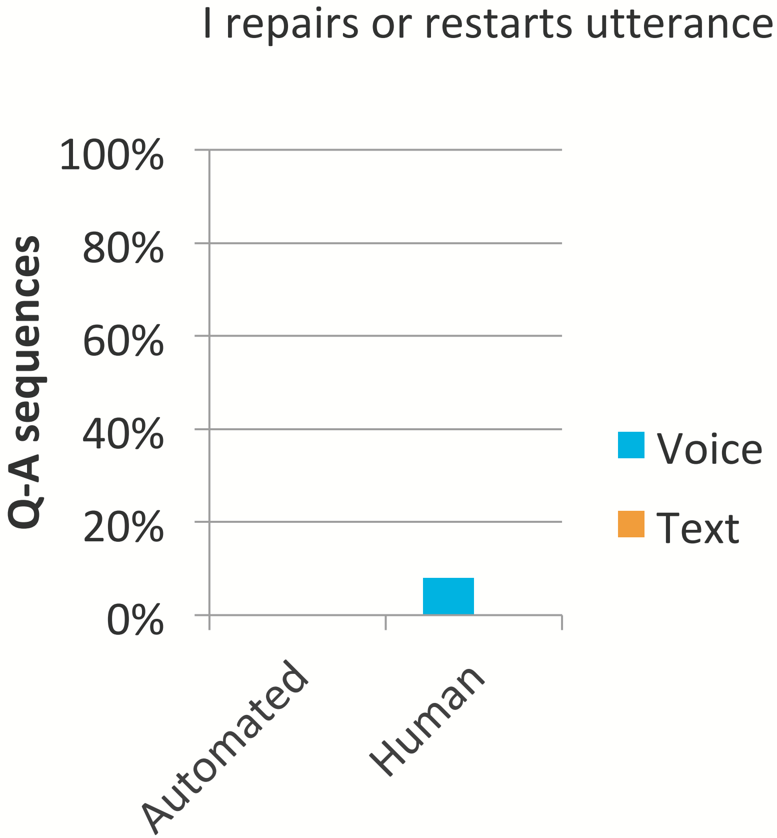
- From data quality evidence, *R*s may be using the time between turns productively
- Could involve checking records and thinking about answer before answering

PROFILE OF INTERVIEW DYNAMICS IN EACH MODE

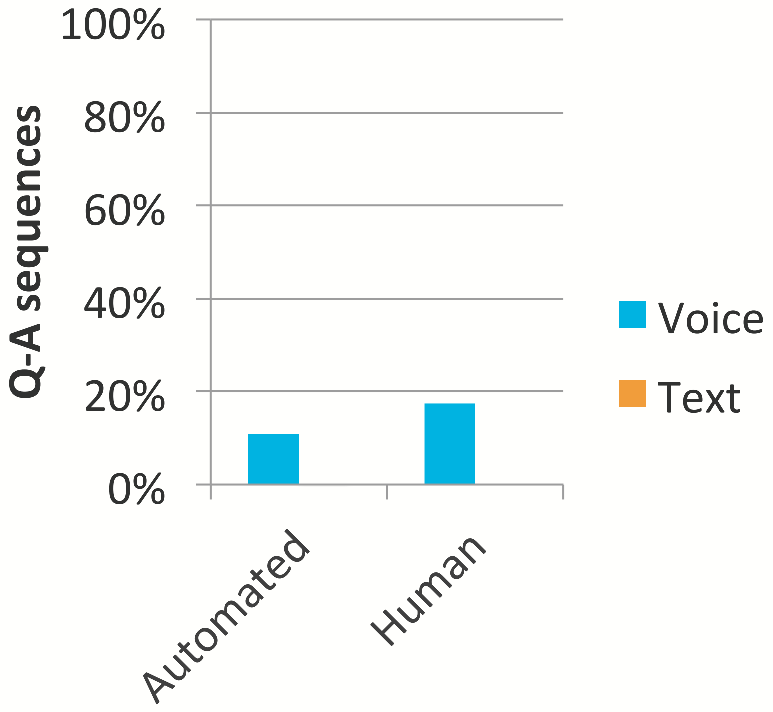
- Coding scheme developed for I and R interview “moves” and interactional paradata in all four modes
  - 25 interviewer moves
    - e.g., ask Q as worded, present response alternatives, no-input (“I didn’t hear that”), no-match (“I didn’t understand that”)
  - 30 respondent moves
    - e.g., answer Q not using exact response alternatives, report behavior instead of answering, ask for clarification
  - Additional behaviors
    - e.g., speech disfluencies and typos, laughter, hedges
- High interrater reliability among 3 coders (Cohen’s kappas = .91-.99) on subset of 400 Q-A sequences from 619 interviews

MODE-SPECIFIC PATTERNS OF MANY CODED BEHAVIORS, E.G.:

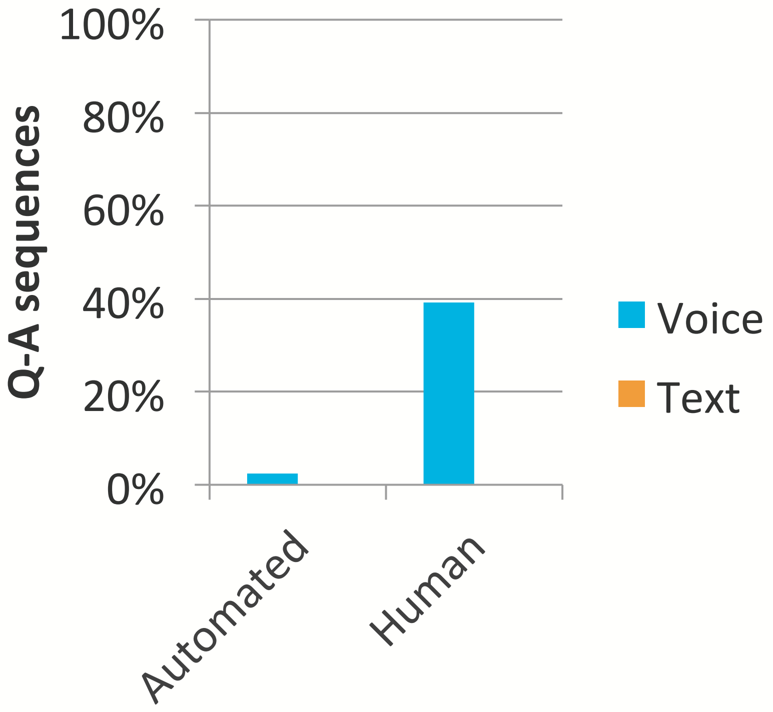




MODE-SPECIFIC PATTERNS OF MANY CODED BEHAVIORS, E.G.:



R gives a synonym of response option



R produces a filler (e.g., “um”)

TEXT (VS. VOICE): SIMPLER INTERACTION (MORE "PARADIGMATIC"* SEQUENCES)

Respondent

- Fewer variable and unacceptable answers
- Less reporting of behavior
- Fewer backchannels ("uh-huh")
- Almost no requests for repeat of survey Q
- Fewer "Don't Know" answers
- Fewer requests for time to find answer
- Less commentary
- Fewer hedges
- No speech disfluencies, few typos

Interviewer

- No misstatements of Q
- Almost no repeats of Q or response alternatives
- Fewer neutral probes
- Almost no laughter (LOL)
- No speech disfluencies (fillers, repairs), few typos
- Less commentary

* Schaeffer & Maynard (1996)

AUTOMATED (VS. HUMAN) INTERVIEWER: SIMILAR (NOT IDENTICAL) PATTERN

Respondent

- Fewer variable and unacceptable answers
- No "reporting" of behaviors
- More changed answers (Auto-Voice)
- Fewer backchannels ("uh-huh")
- Fewer requests for repeat of survey Q
- Fewer "Don't Know" answers
- Less commentary
- Fewer hedges
- Fewer disfluencies

Interviewer

- No misstatements of Q
- Almost no repeats of Q or response alternatives
- No neutral probes
- No laughter (LOL)
- No speech disfluencies (fillers, repairs) or typos
- No commentary

BEHAVIORS AND DATA QUALITY?

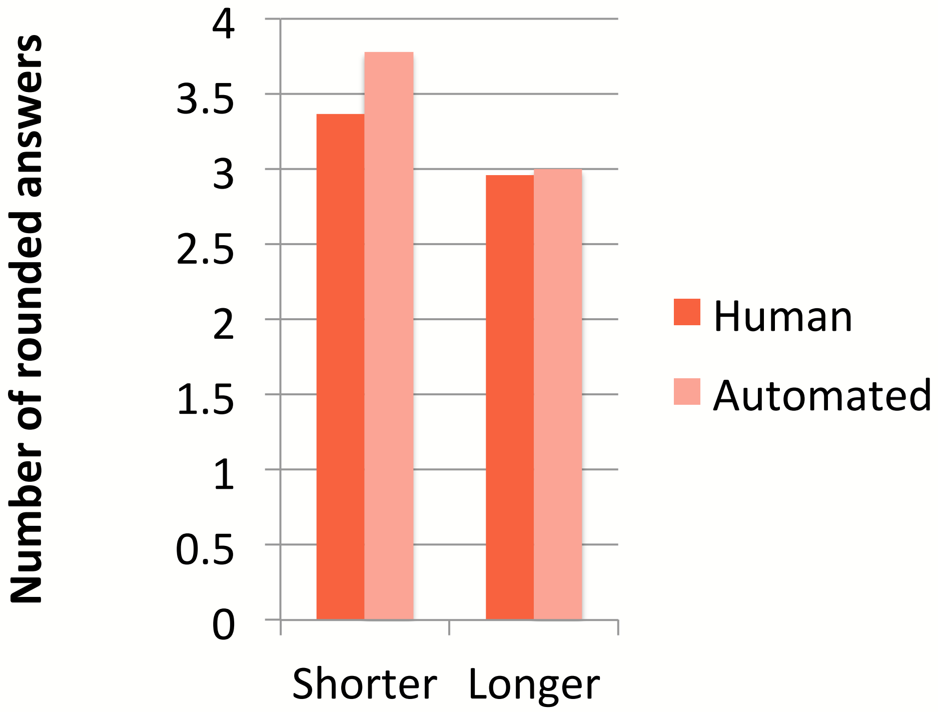
- Many of coded behaviors are plausibly associated with interviewers' "human touch" or "social presence"
- They may also be (though don't have to be) correlates of interviewer-respondent rapport (e.g., Garbarski, Schaeffer, & Dykema, 2016)
- Is there any evidence in this corpus that "humanizing" behaviors are linked with data quality?
- For example, does interviewer laughter, disfluency, or commentary predict Rs' level of disclosure?
  - More disclosure because of increased comfort?
  - Less disclosure because underlines potential that interviewer could be judgmental?

LINKS WITH DISCLOSURE?

- No evidence of difference in disclosure in interviews with more interviewer laughter, disfluency or commentary
- But recall that there WAS more disclosure in text (vs. voice) and automated (vs. human) interviews
  - which had no such interviewer behaviors
- → Consistent with a view that the interviewer behaviors that differ across these modes are part of what causes the data quality differences
  - Maybe *are* what defines the modes
- → Interviewer's "humanness" and social presence can reduce disclosure (relative to automated system), but "more humanness" may not reduce disclosure further

LINKS WITH PRECISION?

- No consistent evidence that interviewer behaviors in voice interviews predict levels of rounding



- But clear evidence in **text** interviews that there is more rounding in faster-paced interviews (shorter interturn interval)
- → Slower is better

than median interturn interval (15.75 sec)

Effect of interturn interval: F (1,309)=11.79, p<.001

SUMMARY: TEXTING

- Text interviews have quite different dynamics than voice interviews on same device
  - Take longer overall but with fewer turns of interaction
  - More “to the point,” less small talk
  - Allow Rs to answer when convenient for them and while multitasking
    - Other evidence: Many Rs reported preferring text to voice interview
- Nonetheless, text interviews led to better data quality (more precision, more disclosure) than voice interviews
  - both in human and automated interviews
  - must be because of features of medium
- → **Decreased social presence** of interviewer and **asynchrony of interaction** may have important benefits

SUMMARY: AUTOMATION

- Automated “interviews” in voice and text have quite different dynamics than interviewer-administered in both modes
  - Schober et al. (2015) analyses: Same effects of automation on precision of answers in both voice and text
  - Independent effect of automation (improvement) on disclosure
  - Reduction in participation with automation
- → Effects of interviewers in new modes differ for different measures of data quality

TOTAL SURVEY ERROR PERSPECTIVE?

- In this corpus, texting clearly improved measurement
- Texting also improved participation
- Can’t tell from this corpus how texting affects potential interviewer effects (assignment of R’s to I’s was not systematic), but worth testing
- In principle, texting could well reduce interviewer effects
  - To the extent that interviewer variance is related to interviewer behavior, texting simply has *less* interviewer behavior
  - Largely streamlines the interview to its essential question-asking and -answering elements
  - Probably leads to more standardized interviews than when interview is conducted in voice

CAVEATS AND CHALLENGES

- Do patterns of findings extend to other implementations of these modes?
  - Other respondent populations, differently incentivized?
  - Different survey questions?
  - Different subpopulations of Rs with different levels of experience in particular modes?
- Challenge: moving target
  - Modes keep changing
  - Adoption trajectories for different populations
  - Evolving norms (e.g., not taking voice calls!)

IMPLICATIONS

- Interviewer effects may look quite different in different modes
- As people's communication habits evolve—including increased interaction with automated systems—previous wisdom about effects of interviewers may change
  - Systematic study over time and in multiple modes will be needed
- Interviewers with particular experience or comfort in particular modes may need to be selected
- "Human touch" in interviewing may have not only important benefits (e.g., motivation, rapports) but also drawbacks (reduction in privacy, intrusiveness)

THANK YOU!

Some publications (thus far):

- https://umich.box.com/s/gctog47xqlhjk0yzfrazfzgkyn8edj9n
- https://doi.org/10.1371/journal.pone.0128337
- https://doi.org/10.1093/poq/nfw097
- http://www.aclweb.org/anthology/W13-4050
- https://www.emeraldinsight.com/doi/abs/10.1108/QAE-06-2017-0033

Data at ICPSR:

- http://doi.org/10.3886/E100113V2
- http://doi.org/10.3886/E100429V1