Using Verbal Paradata Monitoring and Behavior Coding to Pilot Test Gender Identity Questions in the California Health Interview Survey: The Role of Qualitative and Quantitative Feedback

Matt Jans[1], David Grant[1], Royce Park[1], Jane Kil[1], Joe Viana[1],
Nicole Lordi[2], Sue Holtby[2],
Bianca D.M.Wilson[3], Jody L. Herman[3]
[1]UCLA Center for Health Policy Research
10960 Wilshire Blvd. Ste 1550, Los Angeles, CA, 90024
[2]Public Health Institute
555 12th Street, 10th Floor, Oakland, CA 94607
[3]The Williams Institute
UCLA School of Law, Box 951476 Los Angeles, CA 90095-1476

Abstract

This study reports on interview monitoring conducted as part of gender identity question testing in the California Health Interview Survey (CHIS). Four methods of asking gender identity or transgender status were randomly assigned across approximately 3,000 respondents in production telephone interviewing. Two methods used a two-step process, first asking the respondent's sex assigned at birth followed by the sex/gender they identify with now. The other two methods used one question, beginning with a short definition of transgender, and asked if the respondent identified as transgender. We used a combination of qualitative monitoring and question duration coding to assess the relative difficulty of asking each type of question. Two-hundred and twenty (165 English; 55 Spanish) recordings were monitored (50% with men, and a range of ages from 18-70 years). The two-step version was shown to be the easiest to administer by qualitative assessments and quantitative metrics. We discuss theoretical and survey practice motivations for the coding approach and process, and explain how monitoring results were used in conjunction with other pilot test information to evaluate the four methods.

Key Words: gender identity, question pretesting, interviewer-respondent interaction

1. Introduction

Measuring gender identity and transgender identification in the general public poses interesting survey measurement challenges. People who are gender minorities (e.g., transgender, gender queer, non-binary) may have very specific and unique identities, and terminology around gender identity is continually evolving. It seems uncontroversial to assert that self-identified transgender people are probably quite self-aware and reflective about their gender, their physiological sex, and any discrepancy between the two, how that makes them feel, and how they are perceived by others. In other words, gender is likely a very salient and personal topic for them. On the other hand, almost everyone else in the

general population likely reflects on their gender identity rarely if ever, and thinks of themselves as either "male" or "female", with little thought about their "gender" independent of their assigned sex. This creates a challenge for survey measurement because we are faced with measuring a construct that must be defined very precisely for some of the population (in this case, a fairly small proportion), but must also be understood by the rest of the population. If the survey question is not specific or meaningful enough for transgender respondents, we could either create false negatives (i.e., truly transgender people who do not identify based on our question), or we might capture their status accurately, but in a way that limits nuanced analyses. At worst, we might offend transgender people and cause them to stop the survey at that point. On the other hand, if the question requires too much prior knowledge about transgender identities, is too confusing to people not familiar with transgender identities, or is, somehow, offensive to non-transgender people, we could create false positives (e.g., non-transgender people identifying as transgender accidentally) or item nonresponse. Symptoms of such response problems might arise as interaction difficulties during question asking/answering sequences. Thus, we observed and recorded interviewer-respondent interactions, and analyzed question durations during our experimental test.

This paper focuses on the interaction monitoring and question duration measurement portions of this experiment. For more details about the test in general, and response distributions of the questions tested, see Grant et al., (2015), also in the 2015 Survey Research Methods Section proceedings.

1.1 Approaches to Question Pretesting and Interaction Monitoring Coding

There are many ways to qualitatively and quantitatively pretest survey questions. One is to monitor or code interviewer-respondent (I-R) interactions. This can capture interviewer difficulties administering the questions and respondent difficulties answering them. This type of pretesting can range from simple as listening, without coding any behavior, to more complex behavior and verbal paradata coding (e.g., Jans, 2010; Maynard, Houtkoop-Steenstra, Schaeffer, & Van der Zouwen, 2002; Olson & Smyth, 2015; Ongena & Dijkstra, 2006, 2007). The steeply-rising time and effort commitment with the amount of information recorded means that the latter approaches are generally only used in long-term coding projects and basic I-R research, despite early visions that behavior coding could be a standard practice in question pretesting (Cannell, Lawson, & Hausser, 1975). We tried to balance these two approaches in our pretesting.

We began with the assumption that "interaction problems are heard, not seen," so listening to recordings closely would be at least as important as reviewing response distributions. We knew that a combination of qualitative and quantitative assessment methods would provide a deeper sense of the interviewer-respondent interaction, and respondent problems that would not be evident in item nonresponse and break-off rates alone.

We also had little time and resources to commit to the development and implementation of the monitoring. We were hopeful that the interaction problems about which worried most would be so immediately-noticeable that even simply monitoring a few interactions would reveal them. On the other hand, we were concerned that if the problems were rare or unanticipated, we could miss them if we did not monitor a large sample systematically.

For the quantitative component of the assessment, we considered a range of coding schemes including classical behavior coding (Cannell et al. 1975), and more recent research focusing on paralanguage such as, fillers (like "um" and "uh"), answering too

quickly or too slowly, non-answer verbalizations with answer content (i.e., reports), mid-utterance answer repairs (e.g., "Fort- no fifty times."), hedges and qualifications (e.g., "It's about fifty times"), vocal inflection (e.g., answering with rising intonation at end like a question; i.e., "up-speak"), and direct comments about confusion or unease (Conrad, Schober, & Dijkstra, 2008). Indicators from both traditions were used.

2. Method

2.1 Monitoring Sessions

The monitoring protocol had two prongs: a) group monitoring sessions, and b) individual monitoring sessions. Two group sessions were completed with 6-7 listeners at each session, and a total of 66 recordings over both sessions. The sessions lasted 60-90 minutes each. English and Spanish interviews were both monitored as a group (some listeners spoke Spanish and some did not). The only formal documentation of these sessions was meeting minutes circulated after the session.

Individual monitoring sessions were conducted by 3 listeners across about 12 sessions of 60-90 min each. Two of these monitors made detailed measurements of question durations, discussed below. Individual monitoring sessions began as relatively qualitative listening, but also used a form designed to capture relevant interviewer and respondent behavior and paralanguage. Eventually, this form was reduced to one capturing only question duration and a few other variables (discussed more below). Duration analyses here are based on 165 English interviews and 55 Spanish interviews.

2.1.1 Question recordings

Recordings were created in Westat's CATI system by programming recording onset and offset commands to automatically record the relevant interview section unless the respondent asked not to be recorded. Respondents were notified at the beginning of the interview that they may be recorded for quality control purposes, at which point the potential to record would be blocked by the CATI system if the respondent opted out.

For this question test, the gender identity questions (which were randomly assigned to respondents) were placed in a question series on sexual orientation and sexual behavior. To maximize the usefulness of the recordings, we programmed them to begin at the start of the sexual orientation question and end after the gender identity questions. This meant that they also included questions on self-identified sexual orientation, the sex of the respondent's sexual partners, HIV testing/status, and same sex domestic partnership and marriage. Only the gender identity questions are discussed here.

The four questions tested were:

One-step Version 1

> "Some people describe themselves as transgender when they experience a different gender identity from their sex at birth. For example, a person born into a male body, but who feels female or lives as a woman. Do you consider yourself to be transgender?" (Yes/No/DK/Refused)

One-step Version 2

> "Sex is what a person is born. Gender is how a person feels. When a person's sex and gender do not match, they might think of themselves as transgender. Are you transgender?" (Yes/No/DK/Refused)

Two-step Version 1

> Q1 "What sex were you assigned at birth, on your original birth certificate?"
> (Male/Female/DK/Refused)
>
> Q2 "Do you currently describe yourself as male, female, or transgender?'
> (Male/Female/Transgender/DK/Refused)

Two-step Version 2

> Q1 "What sex were you assigned at birth, on your original birth certificate?"
> (Male/Female/DK/Refused)
>
> Q2 "Do you currently describe yourself as male, female, transgender, are you not sure yet, or do you not know what this question means?"
> (Male/Female/Transgender/Not sure yet/Don't know what question means/DK/Refused)

The qualifying phrase "are you not sure yet, or do you not know what this question means" in two-step v2 comes from a question originally tested with teens. Given the novelty of asking transgender questions in general population surveys, we decided to keep it in case respondents did not understand what the question was about but would be afraid to say so.

2.1.2 Group-session monitoring logistics

The group-session monitoring was coordinated by one staff member at UCLA and one at Westat. UCLA staff gathered in a conference room and connected to Westat by via WebEx® video conference software. Recordings were played through the video conference. Staff from the Public Health Institute connected to the meeting via WebEx® as well. This proved to be a very effective way to coordinate monitors across multiple sites.

2.1.3 Individual-session monitoring logistics

Individual monitoring sessions were conducted by 3 UCLA staff one-on-one with the same Westat supervisor. The goal of most of these sessions was to record verbal paradata and other response difficulties systematically.

2.2 Monitoring Documentation in Group and Individual Sessions

It is not essential to keep copious records when monitoring, but it does aid in problem identification, replicability, and formalizing and disseminating design choices. For this test we had different types of documentation for each phase of the test.

2.2.1 Group session monitoring notes

The group monitoring sessions resulted in memos circulated by a UCLA project staff member, and session participants were allowed to take notes in whatever method and level of detail they wanted. They could contribute to the memos, but there not an intensive effort to aggregate notes, nor was any coding of interviewer or respondent behavior done.

2.2.2 Individual session monitoring form, spreadsheet, and record keeping

The process was a bit more complex for the individual monitoring sessions, and evolved over the course of the test. Individual monitoring sessions, like the group sessions, were flexible enough to allow monitors to take notes in a personal way, but were also structured in what was recorded. Two of the monitors used this structured scheme and revised it over time from a broad paralinguistic paradata recording tool (based on paralinguistic indicators from Conrad [2008] mentioned above), to a tool for just recording question durations. The individual-session coding protocol (i.e., our coding form or spreadsheet) evolved over the following stages. All of the forms were created in Excel®.

Stage 1: Many things to code

We started with a form that incorporated the paralanguage and respondent/interviewer behaviors discussed above, so that it was a combination of traditional behavior coding (e.g., read question as worded) and more contemporary survey methodology research (e.g., fillers, long pauses, and reports). It was designed to contain information from one recording (e.g., one I-R pair) on each page. The reason for using Excel® was that data entered into the page-style form could be easily processed into a standard row/column format for analysis. Also, designing the form to fit on one page meant that it could be printed and used as a hard copy if needed. Readers who are interested in learning more details about the coding forms and process can contact the first author directly.

Stage 2: Fewer things to code

After one or two rounds of individual and group monitoring, we realized that there was very little paralanguage to code, and very little variation on the affective expression of respondents and interviewers than we anticipated. The initial form was also hard to complete because of the ordering of items. We removed the least frequent and least variable behaviors, and re-arranged the form to be structured so that coded items appear in the order you would hear them in the interview. This improved form use ease significantly.

Stage 3: Reduction to duration only

As monitoring continued, it became clear that the two-step versions were quicker and smoother to administer than the one-step versions. We decided that it would be important to document and test any duration difference quantitatively, so we revised the goal of the individual monitoring sessions to only record the duration of each question and a few incidental characteristics that could be easily coded. We also found that this information was easier to enter in a standard row/column spreadsheet format (i.e., one I-R pair per row), rather than the page-per-pair format originally used. We added some additional formatting, including bolding and underlines, and summary numbers by question type to make the coding form easier to use while monitoring and easier to obtain quick statistics from. We also added question text to the spreadsheet as an anchor and a check to be used while monitoring. Duration was coded from the beginning of the interviewer reading the gender identity question to the end of the respondent's answer. Respondent age and gender were part of the method that Westat used to identify cases (so that full case IDs would not be used), so those variables were added to the coding form as well.

3. Results

3.1 Coding Efficiency

Taking into account only the staff time used for the individual monitoring sessions (i.e., excluding materials prep, etc.), we were able to monitor about 22 cases per hour with the more structured page-style form that collected a lot of paradata and behavior. The later sessions that used a basic spreadsheet and focused on question duration had a rate of 29 cases per hour. Comparatively, the group listening sessions had a rate of 26 cases per hour. The original individual sessions moved more slowly than the later individual sessions due to the complexity of the form that we reduced over time. The group sessions likely moved more slowly due to the discussions that we had about the interactions we heard. Comparing the operational efficiency of the two methods does not mean that one should be used over the other. They each added unique value to the effort, so we recommend using both in pretesting studies like this.

3.2 Overwhelming Qualitative Findings of Question Performance

Between individual and group sessions it was clear that the vast majority of cases had no confusion and no problem answering, and interviewer problems reading the questions were very rare (too rare to bother recording). Even requests for clarification or repeat, which we thought might be frequent for respondents who were not used to thinking about birth gender v. self-identified gender, were rare. The most frequent voice inflection we heard was rising intonation or "upspeak" on the answer to "What sex … on your original birth certificate," suggesting that respondents might be confused by the question, or at least found it odd. However, it never seemed to impede answering the question or the follow-up question, and so we did not record this systematically.

3.2.1. Qualitative findings in Spanish

We heard a few things in the Spanish interviews that may point to cultural considerations in the measurement of gender identity. For example, the Spanish translation used a term that literally means "transsexual" because there is no commonly-accepted Spanish term equivalent to "transgender." These terms have different connotations in English, so it is difficult to tell if there is semantic equivalence in the Spanish version. Further, the impression of the monitor who conducted in-depth independent review of Spanish interviews was that the Spanish translation felt "clunkier" to read by interviewer. This was a good reminder that parallel testing and re-translation are important if semantic equivalence is desired.

3.3 Question Duration

Question duration includes the entire reading of the question (from the start of the question by the interviewer to the completion of the respondent's answer), not to be confused with response latency (e.g., Bassili & Fletcher, 1991; Couper, 2000; Olson & Smyth, 2015; Yan & Tourangeau, 2008). Our duration measure reflects interviewer pace, question length (in words), the delay between question asking and the start of the respondent's answer, and the respondent's answer itself. Thus they are adequate as measures of the practicality of asking these questions, and tell a combined story of interviewer difficulty reading and respondent difficulty answering. But they do not speak to respondent difficulty exclusively. Extreme outliers in the figures below were removed so that no individual cases were identified.

Figures 1 and 2 show that two-step version 1 (panel c) was the quickest to read in English and Spanish. The differences in means of the four questions was significant at the $\alpha = 0.05$ level for both languages. Figure 2 shows the same relative pattern of question duration across versions for Spanish interviews. Panel c is also the shortest version in Spanish. There were no differences in question duration between men and women within each language, nor was there a correlation of age with question duration.

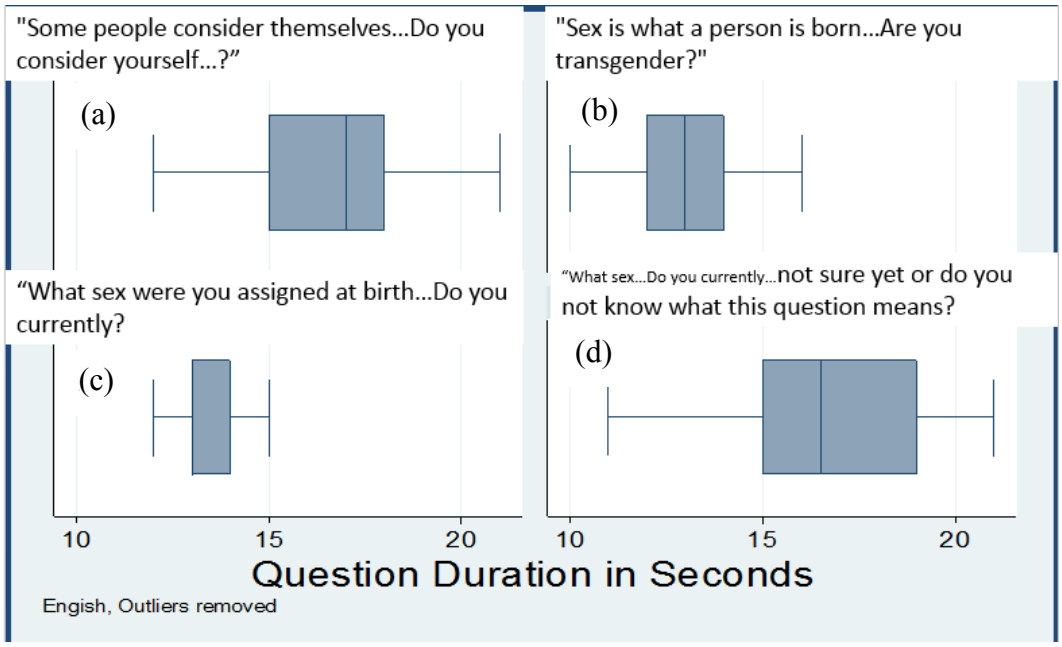

Figure 1: English interview gender identity question duration in seconds (outliers removed)

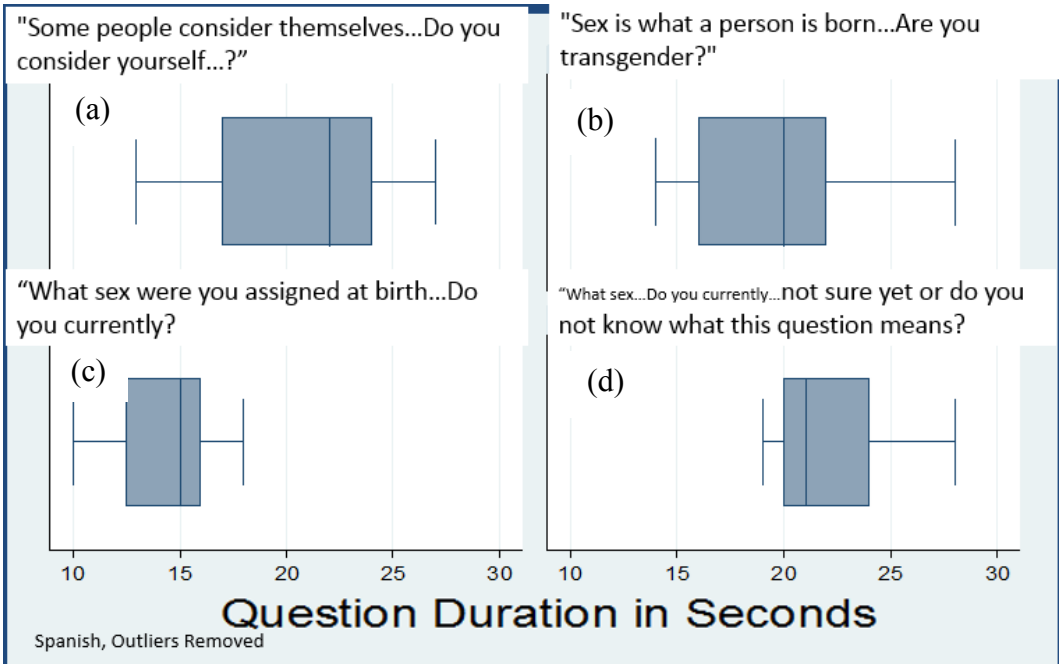

Figure 2: Spanish interview gender identity question duration in seconds (outliers removed)

4. Discussion

4.1 Review of Monitoring Findings

Our monitoring efforts, overwhelmingly, showed us that these questions work well (see Grant, et al. 2015 for details about the observed transgender identification rates). There was very little behavior (paralinguistic or otherwise) on the questions themselves, suggesting that the questions can be asked and do not harm the I-R interaction or overall survey responses. They are not confusing to respondents (at least not confusing enough to cause break-off or large amounts of item nonresponse. A two-step version that asks directly about the respondent's gender on their birth certificate in the first question and their current self-identified gender in the second seems to work the best. It is important to note that, despite appearing longer on paper, the two-step version was significantly quicker to administer, reinforcing the importance of including listening in all question development efforts. There was no duration difference by sex or correlation between duration and age (not shown). Spanish questions took longer to administer, but our current analysis does not allow us to say if that was due to the reading of the question, length of answers, or baseline differences in rate of speech of our Spanish-speaking interviewers and respondents. It did not seem to be due to major comprehension problems, despite the caveats noted above. This could be a fruitful area of future research.

As an aside, we heard more long pauses, fillers (e.g., *um*s and *uh*s), qualifications, and other signs of difficulty, as well as more refusals, on our sexual orientation question. This question has been in the CHIS interview since 2003, and research shows that item nonresponse to it has declined significantly for some groups over that time (Jans et al., 2015). We are planning to explore this in future work, but it was eye-opening that the concept of gender identity, which is probably more foreign to most people than sexual orientation, appears to be easier to ask.

4.2 Review of the Monitoring Approaches and Lessons Learned

The group and individual sessions each had pros and cons that together made a comprehensive interaction review process. Pros of the group listening process include a) immediate discussion of the phenomena heard with our collaborators, and b) having a shared experience/common ground about the questions and issues which aided in editing the questions when needed. The only potential con is that it would be difficult to implement systematic coding or note-taking that requires focus and self-pacing in this setting. The major pro of the individual monitoring sessions is that monitors can self-pace, re-listen to interaction segments as needed, and take detailed and systematic notes that can be analyzed quantitatively, like the question durations we recorded here. Cons include that all notes/experiences must be explained/written/communicated to the other team members, as there is no directly-shared experience. We recommend using both approaches whenever possible.

The other major lesson learned was that sometimes just listening provides enough information to make design conclusions, even without conducting in-depth, detailed, and time-consuming coding. Despite past evidence of paralanguage and respondent behavior that can indicate confusion or difficulty answering, we heard very little in our questions. In retrospect, this is likely due to the personal and, relatively, absolute nature of the questions we tested, compared to the factual questions and complex mappings often used to uncover these problematic behaviors and interactions. It would have saved us some time if we had done more qualitative listening first (individually or as a group) before engaging in recording anything.

It is also enlightening to reflect on the fact that, in this situation, our design conclusions and recommendations would have probably been the same had we only done the group listening sessions and no individual monitoring or duration analyses. Although this could not have been known ahead of time, it suggests that sometimes just listening is enough. Even when there is suspicion that detailed coding may be needed, we recommend listening to 10-20 interactions first, distributed across relevant demographic characteristics if possible (e.g., language, age, sex) to hear what the most obvious and common issues are, and then develop the coding scheme, rather than developing the scheme first. We also take this as support for advice that any question pretesting effort should include listening, even when it is certain that detailed coding will not be done.

If detailed coding is required, much can be done in Excel®, without resorting to specialized software. However, Excel® should be treated as an interface, not just a spreadsheet, and the design of that interface matters. In our early versions we used a spreadsheet that looked like a form or data entry interface. But in the end we decided that a simple row-column display with minor formatting worked fine.

4.5 Limitations and Next Steps

CHIS is a phone survey, and the dynamics of asking gender identity are likely quite different for in-person interviews, just as the dynamic of asking gender is different in that mode. Our findings suggest that further research into the Spanish translation could help identify why they take longer than English questions, and why they appeared rougher in some exchanges. Also, since we only tested English and Spanish versions, future tests should include translations of the questions into the other CHIS languages to ensure measurement equivalence. Finally, our questions were placed in a sequence on sexual

orientation and sexual health. We are curious to know whether our standard gender question, which is asked much earlier in the interview with other demographics questions, could be replaced with the two-step gender identity questions. We are very optimistic about that possibility based on our results.

Acknowledgements

The authors thank Sherman Edwards, Susan Fraser, and Denise Buckley from Westat for scheduling and running the monitoring sessions, and for their commitment to CHIS data quality. We also thank the Arcus Foundation, The Bohnett Foundation, Ford Foundation, The Gil Foundation, and Mr. Weston Milliken for funding this research secured by the Williams Institute. A complete list of CHIS funders can be found at http://healthpolicy.ucla.edu/chis/about/Pages/funds.aspx

References

Bassili, J. N., & Fletcher, J. F. (1991). Response-time measurement in survey research: A method for CATI and a new look at nonattitudes. *Public Opinion Quarterly*, *55*(3), 331–346.

Cannell, C. F., Lawson, S. A., & Hausser, D. L. (1975). *A technique for evaluating interviewer performance: a manual for coding and analyzing interviewer behavior from tape recordings of household interviews*. Survey Research Center, Institute for Social Research, University of Michigan.

Conrad, F. G., Schober, M., & Dijkstra, W. (2008). Cues of communication difficulty in telephone interviews. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. de Leeuw, L. Japec, P. J. Lavrakas, … R. L. Sangster (Eds.), *Advances in telephone survey methodology* (Vols. 1–Book, 1–Section, pp. 212–230). New York: John Wiley & Sons.

Couper, M. P. (2000). Usability evaluation of computer-assisted survey instruments. *Social Science Computer Review*, *18*(4), 384–396.

Grant, D., Jans, M., Park, R., Ponce, N., Kil, J., Wilson, B. D. M., … Gates, G. (2015). Putting the "T" in LBGT: A transgender question pilot test in the California Health Interview Survey. In *Proceedings of the Survey Research Methods Section (AAPOR)*. Hollywood, FL.

Jans, M. (2010). Verbal paradata and survey error: respondent speech, voice, and question-answering behavior can predict income item nonresponse. Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/75932

Jans, M., Viana, J., Grant, D., Cochran, S. D., Lee, A. C., & Ponce, N. A. (2015). Trends in sexual orientation missing data over a decade of the California Health Interview Survey. *American Journal of Public Health*, e1–e8. http://doi.org/10.2105/AJPH.2014.302514

Maynard, D. W., Houtkoop-Steenstra, H., Schaeffer, N. C., & Van der Zouwen, J. (2002). Standardization and tacit knowledge. *Interaction and Practice in the Survey Interview, New York*. Retrieved from https://www.ssc.wisc.edu/soc/faculty/pages/DWM_page/PDF%20files/2002DWM_Schaef_Conversion.pdf

Olson, K., & Smyth, J. D. (2015). The effect of CATI questions, respondents, and interviewers on response time. *Journal of Survey Statistics and Methodology*, *3*(3), 361–396. http://doi.org/10.1093/jssam/smv021

Ongena, Y., & Dijkstra, W. (2006). Methods of behavior coding of survey interviews. *Journal of Official Statistics*, *22*(3), 419–451.

Ongena, Y., & Dijkstra, W. (2007). A model of cognitive processes and conversational principles in survey interview interaction. *Applied Cognitive Psychology*, *21*(2), 145–163. http://doi.org/10.1002/acp.1334

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: the effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, *22*(1), 51–68. http://doi.org/10.1002/acp.1331