Article

Beyond the In-Person Interview? How Interview Quality Varies Across In-person, Telephone, and Skype Interviews

David R. Johnson[1], Christopher P. Scheitle[2], and Elaine Howard Ecklund[3]

Abstract
Conducting qualitative interviews in-person is usually presented as the gold standard, with other modes being seen as inferior. There have been arguments, however, that remote interviews, such as those conducted using the telephone or videoconference technologies, should be seen as equivalent to or even superior to in-person interviews. Evaluations of these claims have been limited by the small number of interviews used to compare modes. We analyze over 300 interviews conducted using three modes: in-person, telephone, and Skype. Our analyses find that in-person interviews have clear advantages when it comes to producing conversation turns and word-dense transcripts and field notes but do not significantly differ from the other two modes in interview length in minutes, subjective interviewer ratings, and substantive coding. We conclude that, although remote interviews might be necessary or advantageous in some situations, they likely do often come at a cost to the richness of information produced by the interviews.

Keywords
mode effects, mode comparison, qualitative research, interviews

We once received a review of a grant proposal in which a reviewer questioned, "Why is it necessary to actually travel to interview a respondent in-person? The interviews could be performed via Skype, for example, at much less cost." The rationale underlying this critique is valid: Researchers have a fiduciary obligation to be cost-effective in how they deploy scarce resources for research. And if one can just as easily record an interview remotely, why bother meeting your respondent in-person?

Many qualitative researchers would recoil in response to this suggestion that researchers "go remote." Indeed, in their widely used textbook on qualitative interviewing, Rubin and Rubin (2011,

[1] Associate Professor of Educational Policy Studies, Georgia State University
[2] West Virginia University, Morgantown, WV, USA
[3] Rice University, Houston, TX, USA

Corresponding Author:
Christopher P. Scheitle, West Virginia University, 29 Beechurst Ave., Morgantown, WV 26506, USA.
Email: cpscheitle@mail.wvu.edu

p. 125) assert that conducting an interview remotely is "not a preferred way." Hermanowicz (2002, p. 497) similarly cautions researchers to "carry on long-distance only as a last resort." There seems to be a consensus among scholars that in-person interviews represent the most desirable format for conducting interviews, while views of remote methods (e.g., such as telephone or Skype) vary from acceptable and necessary (cf. Cachia & Millward, 2011; Holt, 2010) to inherently inferior (cf. Gilham, 2005; Hermanowicz, 2002; Novick, 2008; Rubin & Rubin, 2011).

The assumption behind these scholars' claims about the superiority of in-person interviews is one of mode effects, which refers to how differences in the administration of a given research instrument (e.g., a survey or interview guide) creates differences in participant experiences and data collection. The underlying concern for many is that characteristics of remote interviews undermine the quality of data collected. There is a large literature on mode effects in survey research (cf. De Leeuw, Mellenbergh, & Hox, 1996; Groves & Fultz, 1985; Schaeffer, 1980), examining for instance how effects such as social desirability (presenting oneself in a positive light) and "recency effects" (favoring the last response option offered) vary across self-administered, in-person, and phone surveys. Research on mode effects in qualitative research, by contrast, is less developed (but see Abrams, Wang, Song, & Galindo-Gonzales, 2015; Gravlee, Beranrd, Maxwell, & Jacobsohn, 2013) and there is little consensus about operationalizing "quality," making it difficult to assess whether one interview mode is indeed superior to the other.

The paucity of evidence is understandable. To compare these two approaches, one would need a sample size of interviews large enough to allow statistical comparison of in-person versus remote interviews. One should be hesitant in asserting a claim about what constitutes a "typical" sample size in qualitative research, particularly across disciplines, but recommendations for reaching saturation in the field—which rarely exceed 50 participants (cf. Bernard, 2000; Bertaux, 1981; Creswell, 1998; Kuzel, 1992)—underscore the point that most samples are not large enough for systematic comparisons of quality.

In this article, we examine how in-person and remote interviews generate mode effects in data collection. To do so, we draw on a broader study of science careers (Ecklund, Johnson, Scheitle, Matthews, & Lewis, 2016; Ecklund et al., 2019) that included more than 600 in-depth interviews conducted with physicists and biologists in eight countries around the world. Our analysis focuses on 310 interviews conducted exclusively in English with scientists in the United States, the United Kingdom, and India. We consider five categories of mode effects: word counts, duration, conversational turns, interviewer ratings, and substantive data content. Controlling for interviewer, respondent, and other characteristics, we find that in-person interviews are more conversational and more detailed than remote methods, but that telephone and Skype methods do no clearly lead to differences in interview ratings or substantive data codes generated or used from analysis. In the conclusion, we speculate about what these differences may indicate regarding interview quality.

Beyond the In-Person Interview

The goal of the qualitative interview is to generate data that inform how researchers understand social organization and processes. Done correctly, the interview method pursues this goal by exploring how respondents assign meaning to particular experiences, events, and themes. Most qualitative researchers—at least as evidenced by the interview literature—begin with the assumption that interviews will be conducted in-person. Indeed, the clear "gold standard" for conducting interviews is the in-person interview (McCoyd & Kerson, 2006).[1] The perceived superiority of this mode of interview can be attributed to scholarly tradition and the apparent advantages of meeting a respondent in-person where they live, work, or play. In-person interviews may be seen as advantageous because they provide the most natural conversational setting, the strongest foundation for

building rapport, and the best opportunity to observe visual and emotional cues (Irvine, Drew, & Sainsbury, 2013).

Remote methods (e.g., telephone and Skype) of conducting interviews, by contrast, tend to be ignored or presented as inherently inferior (cf. Hermanowicz, 2002; Novick, 2008; Rubin & Rubin, 2011). Because of the long-standing availability of telephone interviewing relative to online remote methods, such as Skype, critiques of telephone interviews are more voluminous. Scholars perceive telephone interviews as difficult to manage, more likely to result in misunderstandings, limited in their ability to generate meaningful conversations, and challenging contexts for knowing when and whether to ask sensitive or threatening questions (Hermanowicz, 2002; Rubin & Rubin, 2011).

There have been, however, calls for changing the perceptions of acceptable modes for collecting qualitative interviews. Scholars like Holt (2010) and Cachia and Millward (2011) have presented cases for conducting semistructured interviews by telephone. Beyond practical advantages (e.g., more flexible scheduling, reduced travel costs), they both argue that telephone interviews offer distinct methodological advantages. For instance, each points to Sturgis and Hanrahan's (2004, p. 114) observation that some settings are "often loud, public, and uncomfortable." Telephone interviews, on the other hand, provided a calmer and more private setting and therefore produced richer information. Alternatively, one can imagine that unusual time demands or geographic distribution of some target populations, such as elites or professionals, may lead respondents themselves to prefer or insist on remote interviews. These methodological advantages, however, might be better understood as methodological necessities.

Cachia and Millward (2011) also argue that many of the perceived disadvantages of qualitative interviews collected over the phone are overstated or nonexistent. For instance, they state that the loss of information obtained from observing nonverbal cues in-person are often naturally replaced with additional verbal cues or can be replaced by specific probing questions. Holt (2010, p. 120) concludes that "there is no need to consider the use of telephones for narrative interviewing as a 'second-best option: indeed, there may be sound, ideological, methodological, and practical reasons why it may be a more favorable mode than the often 'default mode' of face-to-face interviewing."

The telephone is not the only remote mode that has received attention, as other scholars have advocated for other remote formats to be seen as legitimate means for collecting qualitative interviews. Scholars like Glassmeyer and Dibbss (2012) as well as Janghorban, Roudsari, and Taghipour (2014) have highlighted the potential benefits of using web-based videoconference programs like Skype for conducting interviews. Janghorban et al. (2014, p. 1) note that the use of a web camera makes the interaction "comparable to the onsite equivalent for the presence of nonverbal and social cues." Hanna (2012) makes a similar case. Some have pointed out potential problems with Skype interviews, such as technical issues like dropped calls or poor audio clarity, an inability to see a subject's full body language depending on the position of the camera, and a potential loss of intimacy between subject and interviewer (Seitz, 2016). Yet, others argue that technical issues are less of a problem as technology and connections have improved and that other potential disadvantages can be overcome, making Skype not only an acceptable alternative but also sometimes a better option for producing high-quality interviews (Iacono, Symonds, & Brown, 2016).

Mode Comparison in Qualitative Research

While most discussions of the in-person versus remote interview debate consist primarily of assertions based on practical considerations and researcher experience, a handful of studies have compared interviews collected in-person to those collected using other modes. Claims about quality and analysis of mode effects are often intermingled in such work, but it is debatable whether various observed differences (e.g., word count or interview length) in the data are indeed tied to quality. Accordingly, while we organize our review of existing mode effects research in terms of existing

claims about data quality, our view is that such analyses only indicate various ways data collection differs across in-person and remote interviews. Nevertheless, a case may be made for what such differences in data collection may mean for interview quality. We reserve such speculation for the conclusion.

In the interview literature, one of the most common claims is that high-quality interviews are characterized by depth of detail (cf. Hermanowicz, 2002). The value assigned by social scientists to richly detailed data is seen in the use of strategies such as probing, carefully crafted and sequenced interview guides, listening to what is said and unsaid, and other techniques thought to generate a fine-grained understanding of respondents (Fontana & Frey, 1994; Hermanowicz, 2002; Holstein & Gubruim, 1995, Merton, Fiske, & Kendal, 1956; Rubin & Rubin, 2011). In the mode effects literature, researchers primarily discuss detail in relation to three mode effects: *word count*, *interview duration*, and the *presence of topic-related data*. Word count is the most prevalent mode effect analyzed. Vogl (2013), for example, compared telephone and in-person interviews conducted with 56 children in Germany (a total of 112 interviews) and found the total number of words spoken by the subject were larger on the telephone (1,544) versus in-person (1,451). Sturges and Hanrahan (2004), by contrast, compared transcript length of 21 in-person interviews and 22 phone interviews with individuals in correctional facilities and found that in-person interviews were slightly longer than telephone interviews. In comparing the interviews, they concluded that "the number of responses did not vary greatly relative to each question. More importantly for our purposes, the nature and depth of responses did not differ substantially by type of interview" (Sturges & Hanrahan, 2004, p. 112).

Depth of detail may also be captured by mode effect researchers' analyses of interview duration. This operationalization of detail reflects McCracken's (1988) emphasis on the "long interview," which "is designed to capture the famous 'richness' of qualitative data" (p. 65). Vogl (2013) found that the duration of the interviews did not differ significantly between the two modes, as the in-person interviews averaged 24 min and the telephone interviews averaged 25 min. Irvine, Drew, and Sainsbury (2013) examined 11 interviews, 5 of which were conducted in-person and 6 of which were conducted over the telephone, and found that phone interviews were shorter in terms of the number of minutes (80 min for phone interviews and 101 min for in-person interviews). Based on these findings, they conclude that telephone interviews might tend to produce more "businesslike" conversations, in which the subject is more aware of their role as information provider.

One finds the most innovative operationalization of detail in the work of Abrams, Wang, Song, and Galindo-Gonzales (2015), who compared six focus groups conducted in three settings: face-to-face, online audiovisual, and online text-only. While focus groups represent a different dynamic than one-on-one in-depth interviews, their approach to capturing data richness is worth noting: They operationalized data richness in part by comparing the prevalence of topic-related data in transcripts (using qualitative coding). They found, for example, that face-to-face focus groups generated more topic-related data relative to online audiovisual and text-only interfaces.[2] What sets this approach apart from other qualitative mode effects research is the focus on the content of the data rather than numeric properties of it such as word count.

Another mode effect found in existing research is *researcher ratings* of the interview. Researcher ratings can constitute subjective claims about the interaction between interviewer and respondent. Whether they are good measures of differences in data produced by one mode or another is debatable. As a mode effect, researcher ratings call our attention to how the interviewer's experience varies across in-person and remote modes of data collection. Most frequently, one encounters claims about the superiority of a given mode based on holistic comparisons rather than concrete measures. Iacono, Symonds, and Brown (2016, p. 11), for example, concluded based on comparing their interviews conducted in-person, over e-mail, and using Skype that "the data gathered using Skype, in our personal experience, was just as good as the data gathered using face to face interaction. In

some cases better in fact." Yet, the offer no concrete evidence behind such a claim. The literature does include one exception that relies on subjective ratings of data quality rather than the interview itself. Abrams et al. (2015), during analysis, had coders subjectively assign a "richness" score and found that all face-to-face and online audiovisual focus groups were rated as having high richness, while text-only focus groups were rated as having low richness.

One mode effect rarely considered in existing research is *turn-taking*, which is fundamental to conversation and social interaction. While sociologists and linguists in the field of conversation analysis view turn-taking as an essential concept (cf. Sacks, Schegloff, & Jefferson, 1978), to our knowledge it has received only indirect attention in mode effects research. Vogl (2013), for example, measures requests to speak. Irvine et al. (2013), by contrast, examined types of turns, such as adequacy checks (e.g., asking the interviewer if they answered the question effectively) and acknowledgment tokens (e.g., "uh huh" and "right"). Irvine et al. (2013) found that acknowledgment tokens occur less frequently in phone interviews relative to in-person interviews and speculate that interviewers pay less attention on the telephone and thus provide less feedback to respondents. Turn-taking, of any type, could therefore vary across interview modes and based on Irvine et al. (2013) one might anticipate that turn-taking is less frequent over the phone relative to in-person or Skype interviews.

Despite there being some studies comparing interviews collected using different modes, there remains "a need for additional well-designed studies comparing interview modalities in qualitative research" (Novick, 2008, p. 394). Indeed, with the exception of work by Abrams et al. (2015), Irvine et al. (2013), and Vogl (2013), many of the conclusions drawn in comparisons of in-person and remote interviews are largely of an impressionistic nature. That is, the analysis is often based simply on a sense or feeling of how each mode performed. More broadly, interview mode research is characterized by small samples of interviews in a given mode, which makes it difficult to conduct systematic comparisons.

In the study presented here, we utilize data from over 300 semistructured interviews that were collected using three different modes: in-person, telephone, and Skype. The relatively large number of interviews and multiple modes of data collection allow us to examine much more systematically the potential impact (or lack thereof) of different interview modes.

Data and Measures

The data for this analysis come from a larger project that was broadly designed to understand the social context of scientific work, particularly as it relates to gender, ethics, religion, and policy. The project collected data in eight nations/regions: France, Hong Kong, India, Italy, Turkey, Taiwan, the United Kingdom, and the United States.[3] In each region, the project began with a sample survey of scientists working at universities or research institutes. Across all eight regions a total of 9,422 scientists responded to the survey out of 22,525 who were invited. Additional background and methodological details concerning the survey component of this project can be found in Ecklund, Johnson, Scheitle, Matthews, and Lewis (2016) as well as in Ecklund et al. (2019).

In addition to collecting the survey data, the project conducted follow-up in-depth interviews with some survey respondents in each region. All survey respondents were asked whether they would be willing to be contacted for a follow-up interview. From those that indicated a willingness to be interviewed the project selected potential interviewees with the goal of minimizing travel costs by conducting multiple interviews in particular locations and maximizing the diversity of interview subjects, particularly based on their reported religiosity from the survey data. A total of 609 interviews were conducted across the eight regions.

Subjects were offered the choice of completing the interview in English or the official language of their region. Each interview followed a semistructured format, as interviewers utilized an

interview guide to ensure that key issues were covered, but we also allowed for flexibility to ask follow-up questions or pursue relevant lines of inquiry as they arose. All interviews were ultimately translated into English, but the translation process and the original differences in language make it more difficult to compare interviews that were not originally conducted in English. We therefore limit our analysis here to the three regions in which all of the interviews were conducted in English: the United States, the United Kingdom, and India. This leaves 306 interviews, which is still a sizable sample for an in-depth interview project.

The key interviewers on our team possessed significant experience conducting in-depth interviews and have PhDs in sociology that emphasized qualitative methods. Other interviewers included doctoral students in sociology and two scholars with doctorates in other fields. All team members participated in fieldwork training emphasizing interview strategies, familiarity with the interview protocol and subject matter, and standard operating procedures for recording field notes after each interview. Graduate students interviewed scientists in doctoral and postdoctoral training, while team members with PhDs interviewed scientists of all career stages.

Independent Variable: Mode of Interview

Beyond the large number of interviews collected, a somewhat unique aspect of the interview part of the larger project is that it featured three modes of collection. Specifically, interviews were collected in-person, over the phone, or through Skype (a web-based video calling program). In each region we began data collection with a first stage of in-person interviews in locales with the highest concentration of scientists who indicated on the survey that they were willing to participate in follow-up interviews. If during this stage, a respondent was still willing to meet but unable to do so in-person while researchers were in the region, we scheduled remote interviews for a later date. In the second stage, we conducted interview remotely and allowed respondents to select Skype or telephone. During this stage, we sought to ensure balance in our interview sample across dimension such as gender, institutional type, career stage, discipline, and geographic locale. Thus, while many respondents that we interviewed remotely were not situated in major metropolitan centers, geography was only one factor influencing whether a respondent participated in the interview in-person versus remote. The mode of the interview collection serves as the key independent variable of interest for this study.

Dependent Variables

We examine a broad variety of mode effects identified in our literature review above: interview duration, word count, the presence of substantively important data, interviewer ratings, and conversational turns. Interview duration simply represents the total minutes of the interview (starting with the first question asked and ending after the final response to the last interview question).[4] With respect to word count, we measure the (1) *overall word count of an interview transcript*, (2) *the count of the interviewer's words in the transcript*, (3) *the count of the respondent's words in the transcript*, and (4) *the word count of an interviewer's field notes for each interview*. We differentiate between interviewer and respondent words on the premise that details from the respondent are more important than the commentary of the interviewer. Without such differentiation, it is possible that an especially verbose interviewer could inflate the overall word count of an interview transcript and thereby muddle understanding of the nature of details generated. Interviewer field notes capture a different locus of detail than the actual words spoken during the recorded conversation. These "context notes"—as we referred to them—were not summaries of the interviews; they captured significant details, notable quotes, and offered a preliminary attempt at interpreting themes or patterns emerging in the interview process. They also described occasional nonvisual characteristics

of the interview, such as nonverbal characteristics of the respondent or the setting in which the interview was situated (when relevant). Accordingly, if a study participant was long-winded and produced an excessively long interview transcript with few relevant details, the context notes for this interview might be relatively short.

With respect to the presence of substantively important data, we include two measures tied to data analysis. One measure, *number of codes in transcript*, examines the number of times a given interview generated codes in two of the core sections of our interview guide. One section focuses on ways in which religion comes up in the workplace and includes 14 possible coding categories. The other section focuses on respondents' views of the relationship between science and religion and included five categories. Categories were coded dichotomously with 1 indicating the presence of interview data related to a given coding category and 0 indicating absence of data.[5] We thus generated *number of codes in transcript* by summing these dichotomous variables. Given that the codes across these sections represent themes of substantive importance to our research, this measure provides an indication of whether a given interview mode produces more substantively important data relative to another. A second measure, *interview appears in book*, indicates whether a given interview generated data that were quoted in the book based on this project (Ecklund et al., 2019). This measure is dichotomous, with 1 indicating the book quotes a given interview (any number of times) and 0 indicating it does not. This measure helps assess whether one mode is more likely to produce data that best illustrate substantive themes relative to another.

To capture researcher ratings of interviews, we include the measure *interviewer rating of "excellent."* After each interview, the interviewer was asked to rate the interview as poor, average, or excellent. Our protocol asked interviewers to base their rating on whether the interviews generated interesting information. We also discouraged interviewers from allowing extraneous criteria to shape the rating by training them to note factors such as mistakes or fatigue in their context notes, as necessary. In the analyses below, we focus on a dichotomous measure representing whether the interview was rated as excellent or not (i.e., 0 represents poor/average and 1 represents excellent). This measure at best provides a holistic and indirect assessment of interview quality without and direct measure of differences in data produced. Strictly speaking, as a mode effect, the measure indicates whether one mode of interview generates a different experience for the interviewer.

Finally, to capture conversation analysts' emphasis on turn-taking as an essential component of conversation, we include the measure *number of conversational turns*. This is a continuous measure that is the sum of the total number of statements by the interviewer and the respondent during the recorded interview.

Controls

To isolate the differences in the dependent variables across the modes of interview collection, we must account for other factors that could influence those outcome measures. A total of 11 interviewers collected the 306 interviews, although the number of interviews collected by each interviewer varied. Rather than try to control for every potentially relevant characteristic of interviewers that could impact the outcome measures, we include an indicator representing each interviewer in the analysis except for one interviewer who serves as the reference category. These indicators effectively absorb all differences *between* the interviewers, leaving the analysis to examine differences across the independent variables *within* each interviewer.

In addition to the interviewer indicators, we include dichotomous indicators representing the sex, academic discipline, professional rank, and region of each respondent. For the sex indicator, male serves as the reference category. Professional rank consists of three categories: graduate student, postdoctoral researcher, tenure track/tenured professor. The graduate student group serves as the reference category. As noted earlier, this analysis consists of interviews conducted in three nations:

the United States, the United Kingdom, and India. Respondents in the United States serve as the reference group in the analysis. Finally, we include a measure of *interviews completed by interviewer at time of respondent.* It is possible that an interviewer's interview engagement, context notes, and subjective rating of an interview may differ if it is one's fifth interview conducted for the project versus their fiftieth. This measure helps control for this possibility. If "Interview Case A" has a score of 20 on this measure, the value indicates that this was the 20th interview in the entire project conducted by the interviewer.

Results

Table 1 presents descriptive statistics for all the measures included in this analysis. Looking at the outcome measures, we see that the mean transcript word count is 9,119 words, which consists of a mean of 2,320 interviewer words and 6,799 respondent words. The mean context notes word count is 2,047. The average temporal duration is 65.1 min, with a mean of 216.8 conversational turns. The average transcript was coded with just under seven codes. Just over 40% of the interviews were given a rating of excellent by the interviewers, and just under 40% of interviews appear in the book. It is worth noting that there is sizable variation in these outcomes across the interviews. For example, the shortest interview in terms of time was 23 min long, while the longest interview lasted 139 min. Similarly, the shortest transcript consisted of 3,452 words, while the longest contained 26,659 words. The question of interest for this study is whether interview mode is associated with such variation.

We see in Table 1 that two thirds of the interviews were collected in-person, while about 24% were collected over the phone and 9% were collected using Skype. Four interviewers collected the bulk (73.8%) of the 306 interviews, while the remaining seven interviewers each conducted between 1.6% and 5.9% of the interviews. The interview subjects were relatively evenly divided between the sexes and disciplines, although slightly more subjects were men (56.9%) and biologists (56.9%). A little over half of the subjects were in tenure track or tenured positions (52.3%), while 27.1% were graduate students and 20.6% were in postdoctoral research positions. Lastly, 31.3% of the subjects were in the United States, 43.5% were in the United Kingdom, and 25.2% were in India.

Table 2 begins to examine differences in the outcome measures across the three modes of interview collection. We see that the transcripts for interviews conducted in-person are on average 1,624 words longer than those conducted over the phone and 1,157 longer than those conducted over Skype. We similar differences for the interviewer- and respondent-specific word counts. Furthermore, both phone and Skype interviews have fewer average conversational turns relative to in-person interviews. We also see that in-person interviews are the most likely to have been rated as excellent, as 43.4% of in-person interviews received this rating compared to 35.1% of phone and 28.6% of Skype interviews. These outcomes would all seem to point to in-person interviews as being higher quality than either phone or Skype interviews—to the extent that more detailed and conversational interviews, and interviewer perceptions of quality, are proxies for actual quality.

Other outcomes, though, offer more mixed conclusions. The context notes for in-person interviews are longer than those for phone interviews by an average of 260 words. However, the context notes for interviews conducted using Skype actually average a little longer (66 words) than those conducted in-person. We find a similar pattern with the length of interviews in minutes. The in-person interviews average 66.0 min compared to 61.4 for phone interviews, but Skype interviews actually average a little longer than in-person interviews at 68.1 min. On the other hand, Skype interviews seem less likely to have been used in the book, but phone interviews seem slightly more likely to have been used in the book.

Of course, the patterns seen in Table 2 do not account for interviewer or respondent effects. For this we turn to Table 3, which presents results from ordinary least squares (OLS) regression analyses

Table 1. Descriptive Statistics for All Measures.

	Mean or Percentage	Std. Dev.	Min	Max
Outcome measures				
Word count	9,119	3,088	3,452	26,659
Interviewer word count	2,320	735	900	4,682
Respondent word count	6,799	2,992	1,519	24,791
Context notes word count	2,029	818	608	5,209
Interview length in minutes	65.1	19.0	23	139
Number of conversational turns	216.8	72.6	68	460
Number of codes in interview	6.92	2.10	0	12
Interview rated as "excellent" by interviewer	40.2%	—	—	—
Interview appears in book	38.4%	—	—	—
Mode				
In-person	66.7%	—	—	—
Phone	24.2%	—	—	—
Skype	9.1%	—	—	—
Interviewer				
0	18.3%	—	—	—
1	23.5%	—	—	—
2	21.9%	—	—	—
3	4.2%	—	—	—
4	4.2%	—	—	—
5	4.6%	—	—	—
6	10.1%	—	—	—
7	1.0%	—	—	—
8	5.9%	—	—	—
9	1.6%	—	—	—
10	4.6%	—	—	—
Respondent sex				
Male	56.9%	—	—	—
Female	43.1%	—	—	—
Respondent discipline				
Biology	56.9%	—	—	—
Physics	43.1%	—	—	—
Respondent rank				
Graduate student	27.1%	—	—	—
Postdoctoral	20.6%	—	—	—
Tenure track/tenured	52.3%	—	—	—
Nation				
United States	31.3%	—	—	—
United Kingdom	43.5%	—	—	—
India	25.2%	—	—	—
N		306		

of the word count measures, interview length in minutes, and number of conversational turns and content codes. The OLS models present unstandardized coefficients, as these can be interpreted naturally (e.g., net word count differences). Looking first at the model for the overall transcript word count, we find that, relative to in-person interviews, interviews conducted on the phone and Skype produce transcripts that are 1,454 and 1,486 words shorter, respectively. Both of these differences are statistically significant ($p < .05$) and net of the interviewer and respondent characteristics in the model. We find similar differences when breaking this overall word count into interviewer- and

Table 2. Descriptive Statistics of Interview Outcome Measures by Mode of Interview.

	Mode			
	In-Person (N = 204)	Phone (N = 74)	Skype (N = 28)	Overall (N = 306)
Mean overall word count	9,600	7,976	8,443	9,119
Mean interviewer word count	2,386	2,128	2,320	2,320
Mean respondent word count	7,213	5,847	6,123	6,799
Mean contexts notes word count	2,104	1,844	2,170	2,029
Mean interview length in minutes	66.0	61.4	68.1	65.1
Mean number of conversational turns	228.0	187.6	209.4	216.8
Mean number of codes in transcript	6.82	7.17	7.00	6.92
Percentage of interviews rated as "excellent" by interviewer	43.4%	35.1%	28.6%	40.2%
Percentage of interviews appearing in book	39.5%	40.5%	25.0%	38.4%

respondent-specific counts, although only the phone mode reaches the level statistical significance for the interviewer word count outcome.

Turning to the context notes outcome, we find that, relative to in-person interviews, interviews conducted on the phone and Skype produce context notes that are 503 and 304 words shorter, respectively. Both of these differences are statistically significant ($p < .05$). Note that the difference associated with Skype interviews emerges after accounting for the interviewer and respondent characteristics, as this difference was not seen in Table 2.

The content of researcher context notes provide some clues into why in-person interviews are more detailed than remote interviews. One is that technological problems could take away time for probing or even completing a full interview. Some of the problems that team members noted when performing Skype or phone interviews include respondents with poor phone reception or faulty Internet connections. Were these to occur at the outset of an interview, one can imagine that researchers may be less inclined to probe for more details out of a concern for completing the interview during the scheduled time frame. Another is that remote conversations have lower vocal clarity than in-person interviews. One team member noted in her context notes, for example, "There were some things the respondent said that I absolutely couldn't understand, and this made it difficult to probe at some points in the interview" (RASIC_UK90). Even when conversations proceed without technological barriers, accents—even slight British or Indian accents—can be more difficult for American researchers to understand remotely relative to in-person conversations. This can produce misunderstandings and preclude researchers from hearing important data that might require further probing (for instance because they are reluctant to ask respondents too frequently to restate something just said). Finally, the notes suggest that respondents in remote interviews could be distracted or not offering their full attention. For example, a team member conducted a telephone interview with a scientist in his home and noted in her context notes: "The interview was interrupted twice and R told me that she needs to stop her dog barking (RASIC_UK83). We also recall conducting remote interviews in which background noise suggests that respondents are multitasking—for example, looking at something on their computer.

Examining the other outcomes in Table 3, we do not find any statistically significant differences in the length of interviews in minutes for phone and Skype interviews relative to in-person interviews. We do find that phone and Skype interviews had significantly fewer conversational turns relative to in-person interviews. On the other hand, we do not find significant mode effects for the number of content codes in the interviews.

Table 3. Ordinary Least Squares Models Predicting Interview Outcome Measures (Unstandardized Coefficients).

Outcome	Overall Interview Word Count	Interviewer Word Count	Respondent Word Count	Context Notes Word Count	Interview Length in Minutes	Number of Conversational Turns	Number of Codes in Transcript
Mode							
In-person (ref.)	—	—	—	—	—	—	—
Phone	−1,454.4**	−291.5**	−1,162.8**	−503.0**	−4.6	−47.1**	0.5
Skype	−1,486.7**	−222.1	−1,264.5*	−304.6*	−2.6	−41.6**	−0.1
Interviewer							
0 (ref.)	—	—	—	—	—	—	—
1	1,109.5*	402.9**	706.5	166.5	5.8	31.7*	0.3
2	646.8	−240.9	887.8	−13.3	1.5	7.2	1.3**
3	2,395.6*	609.9*	1,785.6	930.7**	22.6**	131.8**	2.5**
4	4,590.7**	771.6**	3,819.0**	1,135.0**	35.5**	87.8**	0.5
5	−907.5	32.2	−939.8	432.6*	0.1	13.1	0.4
6	−275.9	185.2	−461.1	714.7**	0.7	48.1**	−0.4
7	−1,632.0	−692.7	−939.3	−93.4	−12.1	−24.0	0.9
8	654.1	167.6	486.4	1,603.4**	16.1*	74.8**	0.9
9	−30.1	−143.4	113.3	−214.9	−1.7	−2.7	1.0
10	2,224.2*	−529.8**	2,754.0	1,537.4**	19.4**	7.6	0.6
Interviews completed by interviewer at time of respondent	−11.0	6.1*	−17.2	2.6	0.0	0.4	0.0
Respondent sex							
Male (ref.)	—	—	—	—	—	—	—
Female	469.2	166.9*	302.3	32.9	1.3	14.9	0.2
Respondent discipline							
Biology (ref.)	—	—	—	—	—	—	—
Physics	−452.5	−21.5	−431.0	91.4	−1.2	−9.0	0.0
Respondent rank							
Graduate student (ref.)	—	—	—	—	—	—	—
Postdoctoral	−221.8	−248.8*	26.9	43.1	−2.6	−11.7	−0.3
Tenure track/ tenured	6.8	−266.2**	273.0	182.4	0.5	−11.1	−0.2
Nation							
United States (ref.)	—	—	—	—	—	—	—
United Kingdom	−41.9	503.5**	−545.5	534.3**	0.8	58.9**	1.8**
India	36.8	557.9**	−521.1	106.9	4.3	62.1**	0.8
N	306	306	306	306	306	306	306
Adjusted R^2	.15	.25	.12	.44	.18	.22	.12

Table 4 shows the results from logistic regression models for the two dichotomous outcomes representing the interviewer's subjective rating of the interview as "excellent" or not and whether the interview was used in the book produced from this project. Results are presented as odds ratios,

Table 4. Logistic Regression Models Predicting Interview Outcome Measures (Odds Ratios).

Outcome	Interviewer Rating of "Excellent"	Interview Appears in Book
Mode		
In-person (ref.)	—	—
Phone	0.81	1.24
Skype	0.50	0.53
Interviewer		
0 (ref)	—	—
1	0.78	2.41*
2	1.01	1.93*
3	0.28	4.06
4	0.32	3.56
5	0.19*	3.20
6	0.24*	0.91
7	1.15	[b]
8	0.46	1.89
9	[a]	2.84
10	1.59	3.96*
Interviews completed by interviewer at time of respondent	0.99	1.01
Respondent sex		
Male (ref.)	—	—
Female	1.32	1.31
Respondent discipline		
Biology (ref.)	—	—
Physics	0.57*	0.55*
Respondent rank		
Graduate student (ref.)	—	—
Postdoctoral	2.07	0.90
Tenure track/tenured	3.38**	2.23*
Nation		
United States (ref.)	—	—
United Kingdom	0.78	0.70
India	0.76	1.23
N	301	303
Pseudo R^2	.10	.08

[a]All five of Interviewer 9's interviewers were rated as excellent. Due to this lack of variation, these five cases cannot be included in this model. [b]None of Interviewer 7's interviews were cited in the book. Due to this lack of variation, these three cases cannot be included in this model.

so that numbers above 1 represent an increase in the odds of the interview being rated as excellent and numbers below 1 represent a decrease in the odds of the interview being rated as excellent. We do not find statistically significant differences in the odds of a phone or Skype interview being rated as excellent by the interviewer relative to the odds of an in-person interview being rated as excellent. Similarly, we do not find statistically significant difference across modes of collection in whether an interview was used in the book.

While our focus is on mode differences, it is worth taking a moment to examine the other findings in Tables 3 and 4. Examining the interviewer indicators finds a number of differences across interviewers. For instance, we see that, relative to Interviewer 0, Interviewers 3 and 4 produced significantly more detail and longer interviews across most of the measures presented in Table 3. In

Table 4, we see that Interviewers 5 and 6 have reduced odds of rating an interview as excellent relative to Interviewer 0. Such differences could be a function of factors like interviewer experience or interviewer demographics, but given that these were not our main concern in this study the interviewer indicators provided a powerful means of accounting for all measured and unmeasured differences between the interviewers.

To summarize the mode effects, there is strong evidence that interviews conducted on the phone or on Skype produce fewer details as measured by the word count of transcripts, word count of contextual notes, and number of conversational turns. The evidence does not clearly show, however, that interviews conducted on the phone or Skype lead to temporally shorter interviews, poorer ratings by interviewers, the appearance of content codes in analysis, or the use of the interviews in the book.

Discussion

In this article, we examined variation in mode effects across in-person, telephone, and Skype. To do so, we used 306 interviews conducted in English with physicists and biologists in the United States, the United Kingdom, and India. OLS and logistic regression models allowed us to assess whether—net of interviewer and respondent characteristics—measures related to depth of detail, interviewer ratings, and conversation turns vary significantly across different modes of the qualitative interview.

Given the assumption that in-person interviews are inherently better than "going remote," it is worth speculating what our results may indicate about differences in data quality across in-person, phone, and Skype interview modes. Given that such claims are often made subjectively from holistic assessments of one's experience conducting interviews, it is worth noting that our analysis found no differences across interview modes in how members of our research team rated the interviews. While we asked interviewers to base their ratings on whether they received interesting or useful information, we did not develop a standardized rating procedure beyond these guidelines. Thus, one can imagine that an interview in which one section is richly detailed while other sections are superficial and general is viewed as average by one researcher and excellent by another. The difference could depend, for example, on the importance assigned to the richly detailed section by the interviewer or their interest in the topic. Future mode analyses that rely on subjective ratings by interviewers would benefit from standardized instructions on how to rate interviews. Recall that—as a mode effect—interviewer ratings at best indicate differences in the researcher's experience across interview modes. Nevertheless, to the extent that such a measure provides a proxy or indirect reflection of interview quality, it does not seem to be the case that in-person interviews are any different from remote methods.

If we turn to the interview literature, one claim we see is that the hallmark of a high-quality interview is depth of detail in the data produced. Here, our mode effects analysis implies that there is considerable evidence in support of the argument that in-person interviews produce more richly detailed (and therefore higher quality) interviews than those conducted by telephone online: Remote interviews produced fewer details in terms of overall word counts, respondent word counts, and context notes. Can we assume more words indicate finer grained details about a phenomenon of interest? If when asked about job satisfaction an interview participant replies, "I am more or less happy, yes"—These seven words tell us almost nothing about how they understand their work. An interview characterized by rich details, by contrast, would have followed this response with a probe such as, "Can you tell me about a time you were less happy at work?" This, in turn, would require more specificity, more details, and thus more words from the interviewee. Of course, it is possible that a respondent can be long-winded and produce details of no value. Such events are not pervasive in our data, however, because interviewers on our team were trained to recognize occasions when respondent narratives wander afar and steer the narrative back on track. As scholars who have

collectively interviewed several hundred respondents, we place the most stock in depth of detail as measured by transcript and field note word count, as they are the only record of the encounter once one returns from the field, hangs up the phone, or turns off the computer. On the other hand, the lack of differences across modes in terms of substantive codes generated and data usage in our book undermines our speculation that in-person interviews are superior to remote methods.

Another claim in the interview literature is that high-quality interviews are like conversations (Silverman, 2004). This ideal of interview data collection entails exchange, in which a person provides a narrative in exchange for the "interested attention" and—at times—the views of a researcher (Czarniawska, 2004, p. 48). Interviews with lower turn-taking counts suggest that the interviewer is sticking relatively close to the interview guide, with fewer probes, and limited engagement with the participant. We found that in-person interviews generate significantly more rounds of turn-taking relative to remote interviews, which suggests to us that distance or technology may place a constraint on interaction. There are limits to probes (interviews are not interrogations) and researchers should be cautious about interjecting too much (which may distort or bias views received from respondents). The picture that emerges from less conversational remote methods is one in which respondents are treated as "vessels of answers" (Holstein & Gubrium, 1995) in which the interview guide does all of the work rather than treating them as conversation partners. In this respect, we conclude that in-person interviews are superior to the remote approach.

It is our speculation that in-person interview, by virtue of their conversational and more detailed nature, are indeed superior to telephone and Skype interviews. And while this study has its strengths, it also has its weaknesses. The main limitation is that we did not randomly assign interview modes to subjects. This means that what might seem to be mode effects could be confounded with unmeasured respondent characteristics that led to the utilization of particular modes in the first place. For instance, one could imagine that reluctant or less interested subjects tended to favor conducting interviews over the telephone or over Skype. While this could be a possible cause of the lower transcript word counts, rather than the mode itself, we do not believe it explains our finding that in-person interviews produce more details. Recall that our interview participants indicated on a prior survey whether they would be willing to participate in a follow-up interview. While it is true that there may be variation in the level of interest in participation, it would be unusual for someone who is reluctant to nevertheless participate in the interview portion of the study. Additionally, among the respondents that we could not interview in-person, we did not track preferences for one remote mode over another. We did, however, note during scheduling that strong Internet access would be important to Skype interviews. It is thus possible that some of our results do not apply well to categories of individuals who do not have consistent or strong digital access.

It is also worth noting that our sample is comprised of scientists—who are highly educated and generally accustomed to talking about their work. On one hand, one might question whether the applicability of our findings to other populations by virtue of these characteristics. On the other, the focus of our conversations with these individuals did not consist of questions about science itself, they generally consisted of questions that could be asked of anybody: religious beliefs, ethics in the workplace, and work–family balance, for example. While we believe these findings have broad applicability, they perhaps apply best to other groups that share characteristics in common with scientists, such as professions, semiprofessions, lines of work with strong occupational identity, and other creative and intellectual endeavors.

The chief implications of our results are that researchers should place a premium on in-person interviews, while using remote interviews selectively. Members of our team were no less likely to rate a remote interview as excellent than an in-person interview, meaning telephone and online interviews can be productive and economically advantageous way to collect data. This is an important finding, particularly for researchers with minimal or no funding who have also have research designs that preclude conducting all interviews in-person. Nevertheless, scholars who utilize remote

interviews should be cautious about characteristics of telephone and online interviews that may undermine their ability to produce richly detailed interviews. Preemptive strategies for telephone interviews may include requesting use of land lines rather than cell phones. For online interviews, researchers should request that participants situated themselves in environments with strong Internet connections, preferably over Ethernet connections. And for all remote interviews, asking respondents to situate themselves in private environments that are free of distraction is critical.

Returning to the question posed by the grant reviewer at the beginning of this article, we would say that there does indeed seem to be some benefit to traveling to conduct in-person interviews. Of course, this benefit must also be weighed with the practical concerns raised by in-person interviews (e.g., cost). We by no means suggest that remote interviews are unacceptable, nor do we deny that there may be some situations for which remote interviews are actually preferable (e.g., interviews with individuals in correctional facilities). In many situations, however, the benefits of remote interviews must be seen as potentially coming with some costs to the details obtained in the interviews.

Authors' Note

Declaration of Conflicting Interests

Funding

Notes

1. We refer to "in-person" versus "face-to-face" as the latter's meaning is less precise with technologies like Skype, which could be seen as offering at least a literal "face-to-face" experience.
2. Vogl (2013) also uses a coding scheme to examine differences in the content of statements across interview modes, but her focus is on assessing decision making, communication, and other interaction processes and not depth of detail.
3. Because Hong Kong and Taiwan are not technically nation-states, from here forward we refer to each country case as a region.
4. Because we relied on different actors for interview transcription (two professional firms and undergraduates on our research team), we developed specific protocols to insure that interview transcripts accurately reflect the conversations that took place with respondents. We also included quality control measures to assess accuracy after transcription training (in the case of undergraduates) or contract agreements (in the cases of the two firms).
5. To ensure the accuracy of these codes, our qualitative data set separate cells that contained the actual data indicative of a given code that allowed our team to check quality of coding.

Software Information

Analyses were conducted in Stata SE 15. The analyses utilize the regress and logit commands.

References

Abrams, K. M., Wang, Z., Song, Y. J., & Galindo-Gonzalez, S. (2015). Data richness trade-offs between face-to-face, online audiovisual, and online text-only focus groups. *Social Science Computer Review*, *33*, 80–96.

Bernard, H. R. (2000). *Social research methods*. Sage.

Bertaux, D. (1981). From the life-history approach to the transformation of sociological practice. In D. Bertaux (Ed.), *Biography and society: The life history approach in the social sciences* (pp. 25–29). Sage.

Cachia, M., & Millward, L. (2011). The telephone medium and semi-structured interviews: A complementary fit. *Qualitative Research in Organizations and Management*, *6*, 265–277.

Creswell, J. W. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. Sage.

Czarniawska, B. (2004). *Narratives in social science research*. Sage.

De Leeuw, E. D., Mellenbergh, G. J., & Hox, J. J. (1996). The influence of data collection method on structural models: A comparison of a mail, a telephone, and a face-to-face survey. *Sociological Methods & Research*, *24*, 443–472.

Ecklund, E. H., Johnson, D. R., Scheitle, C. P., Matthews, K. R., & Lewis, S. W. (2016). Religion among scientists in international context: A new study of scientists in eight regions. *Socius*, *2*, 1–9. doi:https://doi.org/10.1177/2378023116664353

Ecklund, E. H., Johnson, D. R., Vaidyanathan, B., Matthews, K. R., Lewis, S., Thompson, R., & Di, D. (2019). *Secularity and science: What scientists around the world really think about religion*. Oxford University Presss.

Fontana, A., & Frey, J. J. (1994). Interviewing: The art of science. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 361–376). Sage.

Gilham, B. (2005). *Research interviewing: The range of techniques*. Open University Press.

Glassmeyer, D. M., & Dibbs, R. (2012). Researching from a distance: Using live web conferencing to mediate data collection. *International Journal of Qualitative Methods*, *11*, 292–302.

Gravlee, C. C., Bernard, H. R., Maxwell, C. R., & Jacobsohn, A. (2013). Mode effects in free-list elicitation: Comparing oral, written, and web-based data collection. *Social Science Computer Review*, *31*, 119–132.

Groves, R. M., & Fultz, N. H. (1985). Gender effects among telephone interviewers in a survey of economic attitudes. *Sociological Methods & Research*, *14*, 31–52.

Hanna, P. (2012). Using Internet technologies (such as Skype) as a research medium: A research note. *Qualitative Research*, *12*, 239–242.

Hermanowicz, J. C. (2002). The great interview: 25 strategies for studying people in bed. *Qualitative Sociology*, *25*, 479–499.

Holstein, J. A., & Gubrium, J. F. (1995). *The active interview*. Sage.

Holt, A. (2010). Using the telephone for narrative interviewing: A research note. *Qualitative Research*, *10*, 113–121.

Iacono, V. L., Symonds, P., & Brown, D. H. K. (2016). Skype as a tool for qualitative research interviews. *Sociological Research Online*, *21*, 1–15.

Irvine, A., Drew, P., & Sainsbury, R. (2013). 'Am I not answering your questions properly?' Clarification, adequacy and responsiveness in semi-structured telephone and face-to-face interviews. *Qualitative Research*, *13*, 87–106.

Janghorban, R., Roudsari, R. L., & Taghipour, A. (2014). Skype interviewing: The new generation of online synchronous interview in qualitative research. *International Journal of Qualitative Studies on Health and Well-being*, *9*, 1–3.

Kuzel, A. J. (1992). Sampling in qualitative inquiry. In B. F. Crabtree & W. L. Miller (Eds.), *Doing qualitative research* (pp. 31–44). Sage.

McCoyd, J. L. M., & Schwaber Kerson, T. (2006). Conducting intensive interviews using email: A serendipitous comparative opportunity. *Qualitative Social Work*, *5*, 389–406.

McCracken, G. (1988). *The long interview*. Sage.

Merton, R. K., Fiske, M., & Kendal, P. L. (1956). *The focused interview: A manual of problems and procedures*. Free Press.

Novick, G. (2008). Is there a bias against telephone interviews in qualitative research? *Research in Nursing & Health*, *31*, 391–398.

Rubin, H. J., & Rubin, I. S. (2011). *Qualitative interviewing: The art of hearing data*. Sage.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In J. Schenkein (Ed.), *Studies in the organization of conversational interaction* (pp. 7–55). Academic Press.

Schaeffer, N. C. (1980). Evaluating race-of-interviewer effects in a national survey. *Sociological Methods & Research*, *8*, 400–419.

Seitz, S. (2016). Pixilated partnerships, overcoming obstacles in qualitative interviews via Skype: A research note. *Qualitative Research*, *16*, 229–235.

Silverman, D. (2004). *Qualitative research: Theory, method, and practice* (2nd ed.). Sage.

Sturgis, J. E., & Hanrahan, K. J. (2004). Comparing telephone and face-to-face qualitative interviewing: A research note. *Qualitative Research*, *4*, 107–118.

Vogl, S. (2013). Telephone versus face-to-face interviews: Mode effect on semistructured interviews with children. *Sociological Methodology*, *43*, 133–177.

Author Biographies

David R. Johnson (drj@unr.edu) is an assistant professor of higher education leadership at the University of Nevada, Reno. He completed his doctorate in sociology at the University of Georgia. Johnson's research examines the professional, organizational, and policy contexts of postsecondary institutions and scientific research.

Christopher P. Scheitle (cpscheitle@mail.wvu.edu) is an assistant professor of sociology at West Virginia University. He received his PhD from Penn State University. Broadly speaking, his research examines the social dynamics of religion in the United States.

Elaine Howard Ecklund (ehe@rice.edu) is the Herbert S. Autrey Chair in Social Sciences, professor of sociology, and director of The Religion and Public Life Program at Rice University. Ecklund received her PhD from Cornell University. Theoretically, Ecklund explores how individuals and small groups bring changes to larger institutions that constrain them. Substantively, her work explores this topic in relationship to religion, science, gender, race, and immigration in different national contexts.