(in press) In F.G. Conrad & M.F. Schober (Eds.),
Envisioning the Survey Interview of the Future.
New York: John Wiley & Sons.

Is it Self-Administration if the Computer gives you Encouraging Looks?

Justine Cassell & Peter Miller
Northwestern University

The movie *Kinsey* opens with a scene of the great sex researcher being interviewed by one of his students. During the course of the interview, Kinsey not only reveals some surprising facts about his own sexual history, but also takes the opportunity to train the student in the fine art of interviewing. Much of the training concerns body language – eye gaze is important to indicate that you are listening, sitting far away creates a perception of distance, frowning will prevent the subject from relaxing. Kinsey's unfortunate student, however, has a hard time controlling his nonverbal behavior as the interview delves more and more deeply into the personal life of his mentor.

Survey researchers have long worried about the unconscious effects of interviewer appearance, including nonverbal behaviors such as these, on the responses of survey interviewees. For this reason (as well as to reduce costs) various communication technologies that allow partial or total self-administration of the interview have been adopted by survey researchers in attempts to objectify the survey process. Self-administered paper and pencil questionnaires, the telephone, computer assisted self interviews and then the Web survey have been thought to hold the answer to the sorts of bias that might be introduced into surveys by the effects of the face-to-face contact of two humans in conversation. The telephone transmits only voice information about the interviewer, and other methods (in the text format most often used to present questionnaires) removes the interviewer altogether. A paradox exists, however, in the debate surrounding the use of these technologies. On the one hand, many researchers continue to

hold to the belief that *rapport* is essential to the successful survey interview. On the other hand, it is also thought that the appearance of the interviewer as well as his/her unconscious nonverbal responses may affect respondents' answers, as the respondents seek to give the answers that they believe are expected. If, however, *rapport* is in fact primarily carried by nonverbal behavior – is induced by the very set of non-conscious behaviors that the telephone and internet are thought to suppress – then there is a problem. For this reason the newest communication technology on the block, embodied virtual humans, are a possible answer to the survey interviewer's paradox Embodied virtual humans might allow us to control the appearance of an interviewer as well as exactly which verbal and nonverbal behaviors are displayed, allowing us to then collect, in an atmosphere of utmost scientific accuracy, the information from citizens upon which the modern democracy depends.

In this chapter we examine the literature on survey interviewing, and the lessons we can learn about the effects of appearance and nonverbal behavior. We then turn to embodied virtual humans and the role that they might potentially play in better understanding these effects in survey interviewing, and perhaps even in mitigating them in practice.

Discussions of interviewer effects often obliquely implicate nonverbal communication between interviewer and respondent. The socio-economic or racial divide between interviewers and respondents has produced worries about respondent deference, manifesting itself in socially desirable answers. Interviewers have been taught that they must establish "rapport" with respondents, but maintain a "professional distance" in order to get accurate information. Survey directors over the decades have speculated on what combination of demographic and attitudinal traits make for a "good interviewer." Interviewing lore is suffused with beliefs about how skin color, gender, dress, demeanor, attitude and "body language" can influence responses for good or

ill. But the contribution of these nonverbal factors to response error has not been identified in most cases, or even studied systematically. One often needs to "read between the lines" of methodological reports to infer how nonverbal communication may have played a role in the findings.

The following selective review looks at examples of research involving the physical characteristics of the interviewer and the idea of rapport in face-to-face interviews, before turning to an examination of research comparing face-to-face and other modes of respondent contact. The review concludes with some observations from studies that have closely examined within-interview interaction, before we turn to the topic of adapting embodied virtual humans to the task of the survey interview.

Physical Appearance of the Interviewer

"The interviewing staffs employed by the polls come from a rather narrow range of the socio-economic scale. They are well above the median of the population in income, they are mostly people with some college education, they are better dressed, more academic in speech and more bourgeois in outlook than the lower income groups they interview" (Katz, 1942).

Social Class. A prominent theme in early studies of interviewer effects was the impact of the interviewer's social class on responses from their lower class counterparts. Katz's (1942) paper is an early example of such articles published between World War II and the 1960s. The basic hypothesis – implied in the quotation above – is that interviewers, who were universally middle to upper class, may obtain socially desirable answers from lower class respondents. Lower class respondents were thought to infer from the dress and mannerisms of interviewers what sort of answers they would like to hear and then "unconsciously" shade responses in that direction. Katz's study involved a quasi-experimental comparison of interviews administered by typical

"white collar" Gallup interviewers and working class interviewers hired specially for the study. The "white collar" interviewers obtained responses that were less supportive of union activity than did their "blue collar" counterparts.

Lenski and Leggett's (1960) "Caste, Class and Deference in the Research Interview" furthers this theme in their examination of the social class and race differential between interviewers and respondents in the 1958 Detroit Area Study. Going beyond the notion that white middle class interviewers would elicit deferential responses on items related to unionism or race, these investigators documented that lower class and black respondents were more likely to give responses characteristic of acquiescence.

Age and Gender. Benney, Riesman and Starr (1956) and Ehrlich and Riesman (1961) examined the effect due to age and gender differences between interviewers and respondents. In the former study, respondents gave different answers to sensitive sexual questions depending on the age and gender of the interviewer. In the latter case, adolescent girls were more likely to report disobedience to younger than older interviewers.

Race. Research on how the race of an interviewer affects responses, mainly focusing on black respondents, also has a lineage stretching back to the 1940s. Hyman (1954) reported the results of two experiments undertaken in 1942, one in Memphis and one in New York, in which white and black interviewers were randomly assigned to interview black respondents. Black respondents in the Memphis sample gave substantially different answers to white and black interviewers, even on innocuous items like automobile ownership. The differences observed in the New York experiment were notably smaller, leading to the belief that the oppressive segregationist environment in Memphis exacerbated the race of interviewer effect.

Robinson and Rohde (1945) conducted an experiment in which interviewers with varying degrees of "Jewishness" (appearance and names) administered interviews that included items measuring anti-Semitism. The more "Jewish" the interviewer (in appearance and name), the fewer anti-Semitic responses received. Athey, et al. (1960) compared responses obtained in interviews conducted by white, Asian and black interviewers. Compared with whites, Asian interviewers got more positive answers to questions about Asian acceptance. Again, compared with whites, black interviewers received more positive views of housing integration. Williams (1964; 1968) reported the findings of a similar experiment in North Carolina in 1960. In this case, the white and black interviewers were randomly assigned to interview black respondents. The study found that respondents tailored racial attitudes in the direction of the interviewers' assumed preferences, particularly low SES respondents.

Another randomized experiment involving race of interviewer was conducted in Detroit in 1968 and the findings reported by Schuman and Converse (1971). Hatchett and Schuman (1975) examined the impact of race of interviewer on white respondents in an experiment in Detroit in 1971. Schuman and Converse's oft-cited findings led to the conclusion in the methodological community that the primary site of race of interviewer effects is in responses to questions about racial matters, and that black interviewers obtain more valid responses on such items from respondents of the same race (though the authors caution that the picture is more complex and is apt to change over time as the nature of race relations changes). Hatchett and Schuman's study of white respondents showed a similar pattern of tailoring responses to what the respondent may perceive that interviewers want to hear. The study called into question the deference or fear explanation commonly offered for interviewer effects involving black

respondents – suggesting that respondents – black or white -- generally seek to get through a potentially difficult social exchange with the least conflict.

Political Affiliation. A unique study reported by Bischoping and Schuman (1992) was conducted during the campaign leading up to the Nicaraguan elections of 1990. This election, following years of civil war, pitted the ruling Sandinista party against a coalition of opposition parties under the UNO label. The campaign was bitter and took place in a volatile context in which political polls, frequently affiliated with one political camp or the other, varied markedly on the predicted outcome. They sought to explain how respondents came to tailor their responses toward the perceived preferences of interviewers who worked for even ostensibly neutral organizations. Schuman designed a randomized experiment in which the same cadre of interviewers, using the same questionnaire, sponsored by a local university, conducted interviews with only one variation in approach – the color of the pen they used to record responses. When interviewers used a red pen, a major color of the Sandinista party, reported political preferences tended to favor that group. When they used a blue pen, the color of the UNO coalition, the response pattern was reversed. This is the only study reviewed here in which a single nonverbal characteristic was the sole manipulation, and it demonstrates the extent to which putatively small changes in the interviewing situation can apparently lead to major differences in response.

Multiple Social Characteristics. Some investigations have studied interviewer-respondent disparities on multiple social characteristics. Dohrenwend, Colombotos and Dohrenwend (1968) analyzed income and race differences between interviewers and respondents in a public health survey in New York. Weiss (1968) examined age, education and socioeconomic status disparities between interviewers and respondents in a survey of welfare mothers. As noted

above, Williams (1964) examined race of interviewer effects for high and low SES black respondents.

What do we learn from this body of research into the effects of demographic characteristics on interviewing? With the exception of Katz's early quasi-experiment, studies examining the effect of social class differences between interviewers and respondents are based on post-hoc coding of interviewer and respondent characteristics. Katz's study, comparing white collar and blue collar interviewers, could not make clear inferences about interview effects because the interviewers were responsible both for recruiting respondents and questioning them. In none of the social class studies was there within-interview evidence of the process by which the social class differential produced the effect it was claimed to produce. Similarly, studies of age, gender or race of interviewer effects infer a process of communication that is not documented through observation. Respondents may tell interviewers what they think they want to hear for a complex of reasons, including nonverbal cues, but the studies in this area were not designed to "unpack" how social class, age, gender or race worked their effects in the interview, nor how the communication might vary in different interview contexts. Nevertheless, the studies presented above are highly suggestive of effects due to the appearance of survey interviewers.

Rapport

Along with interviewer demographic characteristics, "rapport" is a frequently cited factor that may affect the quality of data obtained in the interview. It is an elusive concept, as Weiss (1970) pointed out. A dictionary definition is "a relationship, especially one of mutual trust or emotional affinity." Establishing "rapport" was traditionally thought to be a prerequisite of gaining interview cooperation. Interviewing texts mention the importance of establishing trust so that respondents will be forthcoming in their answers, but, as Hyman (1954) famously noted,

it can be overdone, leading to ingratiating rather than honest responses. For Cicourel (1964) rapport is incompatible with a task orientation to the interview. This concept has been the subject of much speculation but little research. If it is useful in the interview – and Weiss (1970) and Goudy and Potter (1975) suggest that it is not – nonverbal behaviors are likely to play a major role in establishing it. In her brief review, Weiss (1970) states – but does not cite the relevant studies – that nonverbal communication such as eye contact, smiles and head nods contribute to rapport between interviewer and respondent.

Sheatsley's (1950; 1951) exhaustive description of the characteristics of the NORC field staff begins and ends with the concern that respondents who are questioned by dissimilar interviewers will provide erroneous answers. For Sheatsley, an interviewer's "rapport" with the respondent was the same thing as demographic similarity. See also Williams (1964), who states that, "Status characteristics directly affect communication, and similar status tends to reduce bias, especially by reducing inhibitions." (p.340) Riesman and Glazer (1948) in a wide-ranging essay on "the meaning of public opinion," following the failure of the polls prior to the 1948 US presidential election, also equate "rapport" in part with demographic similarity, noting that "the polls try to use an interviewer of [the respondent's] own ethnic group, though they are rarely able to use one of his own class (p. 643)."

Other investigators went beyond demographic similarity to measure "rapport" in different self-reported measures. Dohrenwend et al. (1968) classified interviewers according to whether they had preferences for interviewing particular kinds of respondents and whether they reported suffering embarrassment when asking about any of the topics in the interview. These scores, applied to the interviewers, were expected to capture how they may have behaved when talking with "un-preferred" people or when working through parts of the questionnaire they did not like

to administer. A similar interviewer-based scoring system was employed by Williams (1968), who had interviewers complete measures designed to capture their general capacity for "rapport" and "objectivity." Weiss (1968), by contrast, had interviewers score each respondent on a scale measuring how cordial and disclosing she had been. This interview-specific measure, though not the product of an analysis of the interview interaction and based only on interviewer assessment, is a step closer toward the goal of capturing the affective flavor of an interview. Importantly (and ironically) Weiss found that, overall, the higher the rapport on average, the lower the validity of the responses (ascertained through matching responses with administrative records).

Studies of Interaction in the Face-to-Face Interview

The studies reviewed above identified factors involving nonverbal communication in face-to-face interviews that may lead to response error. The studies do not spell out a clear mechanism for the effects observed and do not offer interview observations of nonverbal communication. We do not know, for example, if white interviewers behaved differently nonverbally when asking questions of black respondents about racial matters than they did when addressing other topics. We don't know, similarly, if younger interviewers exhibited different nonverbal behaviors when they obtained more reports of disobedience from adolescent girls than did older interviewers. We know that interviewers must establish some sort of relationship with respondents if they are to complete an interview, but the studies of "rapport" leave us in the dark about the behaviors that produce the sort of "rapport" that leads to better reporting.

Although "race of interviewer" effect studies still appear in the literature now and then, there is little research on other aspects of interviewer physical appearance and "rapport" today. In part, the shift away from this line of research is the result of the very technological change we described above – away from face-to-face contacts and to telephone, self-administered and Web

surveys. In part, the change also reflects lessening concern about some social fissures in American society (social class). Also, the field has come to believe, based on studies like Schuman and Converse (1971) and Hatchett and Schuman (1975), that we need only worry about response effects from interviewer appearance in special cases like racial attitudes. Finally, survey organizations recognized from the time of Sheatsley's report in the early 1950s that, due to the vagaries of the survey business and its labor pool, it was impossible on a general basis to physically match respondents with interviewers.

In concert with these observations, the study of interviewing shifted its focus to more micro-analysis of interviewer-respondent interaction. Aided by the development of small, battery-powered audiotape recorders, records of question-asking and answering could be examined closely for characteristics related to better or worse reporting. The programmatic research by Cannell and colleagues, beginning in the 1960s and continuing for three decades, is the key exemplar of this approach. Like the earlier work on interviewer characteristics, it seldom addressed nonverbal behavior. But some of these studies offer clues into how nonverbal behavior might work in the interview and also offer a template for studies that take on the topic directly.

A formative investigation in this program of research was Cannell, Fowler and Marquis (1968), in which reporting in health interviews was examined in relation to a wide variety of independent variables – interviewer and respondent demographic, knowledge and attitudinal variables as well as all of the verbal behaviors that took place in the interviews. A key lesson of the study was that the behaviors predicted response quality better than the demographic variables that had characterized earlier interviewing studies. Interviewers and respondents tended to balance one another's verbal output. Some of the output produced useful answers – as when the

interviewer asked a question correctly and the respondent gave an answer that met the question's objectives – and much verbal activity was devoted to other sorts of task relevant and irrelevant objectives.

The focus on micro-analysis of behavior offered the possibility of "unpacking" interviewer effects attributed to demographic differences with respondents. In a following study, Marquis and Cannell (1969) examined the exchanges between white interviewers with older and younger white and black respondents. They found, as in the earlier investigation (Cannell et al., 1968), that there was more verbal activity in total in the interviews with older respondents, and that there was more "task orientation" in the interviews with younger and black respondents – less overall verbal activity. The interviews with older and white respondents featured more non-task communication, including comments on the questions, "polite behavior," elaborations, suggestions to the interviewer and refusals to answer. In other words, the white interviewers appear to have experienced something more like a "conversation" when interviewing white and older respondents. The addition of nonverbal observations would have allowed this hypothesis to be tested.

One more experiment in this research program suggests how the micro-level study of interviewing effects could benefit from observations of nonverbal behavior. Henson, Cannell and Lawson (1973) compared two interviewing treatments for eliciting information about automobile accidents: a "personal" style emphasizing incipient friendship (like some earlier ideas of "rapport") and an "impersonal" approach focused on administering the questionnaire in a standard manner. The two approaches produced basically equivalent results in the accuracy of accident reporting. Van der Zouwen, Dijkstra and Smit (1991) also report inconclusive results in a similar comparison.

Lessons and Shortcomings of Interview Interaction Studies

Early studies of interviewing effects based on interviewer appearance begged the question of how physical characteristics of the interviewer worked the effects they were said to produce. More recent research, which has focused carefully on interviewer respondent interaction, has improved on the "black box" of earlier studies, but its utility is limited by the absence of data on nonverbal behavior (with a few exceptions, e.g., Schober & Bloom, 2004 on paralinguistic indicators of misunderstanding). We learn from studies of interviewer personal characteristics and "rapport" that respondents sometimes give answers that appear to be ones that would be well received by interviewers, rather than ones that seem more accurate – an effect that has been referred to as "social desirability". Rapport has been equated with similarity between the interviewer and respondent, or has been conceived of as a characteristic of the relationship between the two, in which case it may mediate or interact with physical similarity.

For the most part research on rapport in the survey interview has not drawn from research in the field of interpersonal interaction. Thus, while there is a fairly substantial body of literature examining the verbal and nonverbal components of rapport, it has not influenced survey research. A particularly influential theory (Tickle-Degnen & Rosenthal, 1990) describes rapport as comprised of three components, which change in importance over time, and each of which is conveyed by both verbal and nonverbal means. *Positivity* is particularly important early in rapport-building, and can be indicated by smiles. *Mutual attention* remains important throughout the development of rapport, and is conveyed by eye gaze among other behaviors. *Coordination* becomes increasingly important over the course of a budding relationship, and is marked by features such as quick turn-taking, frequent acknowledgement of the other's utterances, and so forth. Other researchers have found important contributions in assessment of rapport played by

posture (LaFrance, 1982) and certain kinds of nonverbal mimicry (Lakin, Jefferis, Cheng, & Chartrand, 2003). We do not know, however, the extent to which survey interviewers are consciously or unconsciously engaging in these behaviors and for this reason the process by which demographic similarity or rapport affects respondents' answers remains a "black box."

In sum, it is clear that the appearance of the interviewer, including interviewer demographics and moment-to-moment nonverbal reactions, have some effect on the interview, but neither the extent nor the mechanism is clear.

These questions became largely moot for many survey applications as the telephone replaced the face-to-face interview in commercial, academic and, increasingly, government surveys. The energy that might have gone into studying nonverbal behavior in the face-to-face interview was expended instead on aspects of the telephone contact – e.g. how respondents would react to the lack of visual cues provided in person (viz. Miller, 1984) or how the interviewer's tone of voice might affect nonresponse in telephone contacts (e.g.Oksenberg, Coleman, & Cannell, 1986). Early consideration of the telephone interview produced the hypothesis that the audio-only mode might encourage more respondent disclosure because the interviewer was not physically present (e.g. Hochstim, 1967). But the issue of social desirability did not disappear with the growth of telephone interviewing. Instead, newer technology, operationalized in computerized self-administered questionnaires (computer-assisted self interviewing or CASI), or in audio computer-assisted self interviewing (A-CASI) methods employed in telephone contacts was introduced to remove the effects of the interviewer's presence and lessen social desirability bias. A-CASI has shown some notable results (viz. Villarroel et al., 2006) and more mixed findings (e.g. Currivan, Nyman, Turner, & Biener, 2004; Moskowitz, 2004). In any case, the general trend in methodological research has been to

examine ways in which the impact of the interviewer, whatever it may be, can be lessened or removed and self-administration can be increased, in the interest of obtaining honest answers from respondents.

Embodied Conversational Agents

Enter the embodied virtual human (here forth called by the name by which it is most commonly known to researchers in Human-Computer Interaction, *Embodied Conversational Agent* or ECA). To be clear on our terms, an ECA is a graphical projection of a full-body human on a screen. The ECA may be life-size and projected onto a giant screen, or may be small enough to fit on a hand-held device such as a cell-phone. The depiction of the human can fall anywhere between absolute realism and cartoon depiction. But for it to be called an Embodied Conversational Agent it must be able to speak and also to display many of the nonverbal behaviors that humans do in face-to-face conversation, such as eye gaze, head nods, facial displays, body posture, and hand gestures. And it must be capable of simulating many of the same responses that humans give, such as happiness and sadness, attentiveness and boredom, desire to take the floor, and acknowledgement that the other's words have been understood. In fact, much research on ECAs is directed towards *autonomous* embodied conversational agents, where some or all of those responses are automatically generated on the basis of underlying models of human behavior (Cassell, Sullivan, Prevost, & Churchill, 2000). That is, autonomous ECAs may nod when their speech recognition algorithm has actually understood what the real human interlocutor has said, and look happy when their model of emotion has detected the fulfillment of a hoped-for goal. During the development phase, however, the effects of ECAs on interaction are often studied using what is called the "wizard of oz mode" where a scientist hidden behind a curtain controls the behaviors of the ECA. In any case, and importantly for our

purposes here, ECAs are implemented on the basis of research into human behavior, and they are most often evaluated with respect to a gold standard of real human behavior. They are also the result of decades of research in Artificial Intelligence (AI) and Natural Language Processing (NLP) that allows the construction of systems that understand language, reason on the basis of that understanding, and produce responses appropriate to the context. However, the ECA marked an important departure from previous work in AI and NLP, in that it recognized the importance of the social context in interaction – that reasoning and understanding are not sufficient for human conversation, which also requires a *display* of having understood, and the means to regulate the conversation through devices such as nods and lean-forwards. ECAs are attempts to make interaction with a computer more natural, intuitive, and like human interaction.

In this vein, the methodology underlying ECAs is quite specific. Researchers collect data on human-human conversation, analyze those data in such a way as to build a formal or predictive model of the human behavior observed, implement a computational system of a virtual human on the basis of the model, evaluate the system both by observing what it looks like themselves, and by observing how it interacts with real humans, evaluate the results, and go back to collect more data on human-human communication to fill the lacunae in the model.

As an example both of how ECAs are implemented and how they function, Figure 1 shows an ECA named REA (for Real Estate Agent) who was programmed on the basis of a set of detailed studies into the behavior of realtors and clients. Over a period of roughly 5 years, various graduate students, post-docs and colleagues in Cassell's research group studied different aspects of house-buying talk, and then incorporated their findings into the ECA. Hao Yan looked at what features of a house description were likely to be expressed in hand gestures, and what features in speech (Yan, 2000). Yukiko Nakano discovered that posture shifts were

correlated with shifts in conversational topic and shifts in whose turn it was to talk (Cassell, Nakano, Bickmore, Sidner, & Rich, 2001). Tim Bickmore examined the ways in which small talk was employed to establish trust and rapport between realtor and client (Bickmore & Cassell, 1999). Earlier work by Scott Prevost on intonation (Prevost, 1996) and by Obed Torres on patterns of eye gaze (Torres, Cassell, & Prevost, 1997) also went into the implementation. As research into human conversation progressed, the group also came to better understand some of the overall properties of human conversation, and the system was iteratively re-designed to incorporate those insights.

The result was a virtual woman who would try to sell a home to whoever approached her. A small camera on top of the screen allowed her to detect the presence of real humans and initiate a conversation with them. Her knowledge of appropriate realtor conversation led her to ask questions about the person's housing needs and then nod, seem to reflect, and pull up data on appropriate properties, describing them using a combination of descriptive hand gestures, head movements and spoken language.

Figure 1: REA, the virtual Real Estate Agent

A number of experiments evaluated REA's performance in terms of how well it compared to human-human interaction. Thus, Nakano's work looked at anecdotal evidence of how closely patterns of eye gaze during a conversation between a real person and REA resembled eye gaze between two humans (Nakano, Reinstein, Stocky, & Cassell, 2003) and she found that implementing a model of nonverbal behavior in conversational acknowledgement (grounding) appeared to improve the extent to which the human-machine conversation resembled human-human conversation. Other experiments compared how well the ECA functioned with and without particular human-like "modules," such as facial displays of emotion, hand gestures, and small talk.

Social Dialogue in Embodied Conversational Agents

The small talk study is particularly revealing of the kinds of complex effects that ECA researchers are likely to find. Cassell and Bickmore initially posited that small talk would be a particularly successful addition to the ECA for the very reasons described above: that it can be used to provide such social cues as attentiveness, positive affect, and liking and attraction, and to mark shifts into and out of social activities (Argyle, 1988). And, people who interact with ECAs seem to wish to engage in such social dialogue, as shown by a naturalistic study of an ECA embedded in an information kiosk, where roughly 1/3 of the 10,000 utterances from over 2,500 users were social in nature (Gustafson, Lindberg, & Lundeberg, 1999). In Cassell and Bickmore's work, then, two versions of the ECA were implemented, of which one was capable of using small talk in the way that researchers in conversational analysis have documented, to mitigate face threat and increase trust, while the other version simply had REA get down to business.

An initial study compared how likely users were to trust REA to sell them an apartment, when she used small talk vs. task talk only. Since prior literature has suggested a mediating effect due to personality, the study further compared introverts and extroverts in their use of the two versions of REA. Results demonstrated that extroverts trusted the system more when it engaged in small talk, while introverts were not affected by the use of small talk (Cassell & Bickmore, 2002). Insofar as these results mirror research in social psychology, they seemed to indicate that people were reacting to REA as they would react to another human. They also lent support to our model of the role of small talk in task-oriented conversations such as selling real-estate.

A subsequent study (Bickmore & Cassell, 2005) added the additional condition of medium, comparing users communicating with REA by standing in front of the screen on which she was displayed to users communicating with REA by telephone. In this second study, however, results concerning the role of small talk were strongly mediated by the role of the body in the interaction. That is, main effects showed that subjects in the *phone* condition felt that they knew Rea better, liked her more, felt closer to her, felt more comfortable with the interaction, and thought Rea was more friendly than those in the embodied condition. In addition, social dialogue was more fun and less tedious on the phone while only task-limited dialogue was judged to be more fun and less tedious when embodied. That is, subjects preferred to interact, and felt better understood, face-to-face when it was a question of simply "getting down to business," and preferred to interact, and felt better understood, by phone when the dialogue included social chit-chat.

Looking back at the implementation of REA, Cassell and Bickmore came to believe that these results were a condemnation of REA's nonverbal behaviors, which may have inadvertently

projected an unfriendly, introverted personality that was especially inappropriate for social dialogue, and that was at odds with the model of small talk that had been implemented. Rea's model of non-verbal behavior, at the time of this experiment, was limited to those behaviors linked to the discourse context, and had not been changed for the small-talk version. Thus, Rea's smiles were limited to those related to the ends of turns, and she did not have a specific model of immediacy or other nonverbal cues for liking and warmth typical of social interaction (Argyle, 1988). The results obtained indicate that adding social dialogue to embodied conversational agents requires a model of social nonverbal behavior consistent with verbal conversational strategies. To this end, in more recent work Cassell and her students have begun to examine how rapport is signaled nonverbally with an eye towards revising REA's model of nonverbal behavior. Although some literature exists on this subject, it is incomplete. And so, in particular, they are interested in how friends differ from strangers, and people meeting for the first time differ from those same people meeting a third or fourth time; in situations where the participants can see one another (and thus nonverbal behaviors can play a role) and situations where there is no visual access. When these data are analyzed, REA's nonverbal behavior will be updated. But, until good process data can be collected on the user's behavior, REA will not be able to know whether her rapport-building behaviors are successful (see Person, this volume).

But the results with REA also demonstrate the extent to which subjects quite unconsciously respond to REA as if she is a real person – for example, judging her as unfriendly when she does not display nonverbal behaviors linked to rapport. In this context, an analysis of the users' speech behavior in talking with REA is revealing. People tend to "hyper-articulate" their speech when they talk to computers (Oviatt, MacEachern, & Levow, 1998) whereas their speech to other people contains more slurred speech, interruptions and disfluencies. Looking at

rates of disfluency in users communicating with REA showed that interactions with REA were more similar to human-human conversation than to human-computer interaction (Bickmore & Cassell, 2005). This subject's responses during debriefing made the same point:

> REA exemplifies some things that some people, for example my wife, would have sat down and chatted with her a lot more than I would have. Her conversational style seemed to me to be more applicable to women, frankly, than to me. I come in and I shop and I get the hell out. She seemed to want to start a basis for understanding each other, and I would glean that in terms of our business interaction as compared to chit chat. I will form a sense of her character as we go over our business as compared to our personal life. Whereas my wife would want to know about her life and her dog, whereas I really couldn't give a damn.

However, these results also serve as a warning of the extent to which *all* of the ECA's behaviors need to be considered in designing an experiment because "people cannot not interpret". That is, as Kinsey told his student, every behavior of an agent – an artificial agent just as much as a human being – is the basis for an inference about that person's personality, emotions, and stance towards the conversation. Reeves and Nass (Reeves & Nass, 1996) have demonstrated much the same thing. Replicating several decades of social psychological research, systematically substituting a computer for one of the participants, their findings show that computers can evoke behaviors quite similar to those evoked by another human, including behaviors of social desirability. Thus, for example, when subjects were asked to rate the performance of a piece of computer software, they were harsher in their assessment when the questionnaire was filled out on an adjoining computer, rather than the computer on which they had used the software -- presumably because they didn't want to hurt the computer's feelings (Nass, Moon, & Carney, 1999). And on the basis of these findings, they conclude that people treat computers as social actors rather than as tools.

With this caveat in place, how can ECAs be used in survey interviewing?

ECAs in Survey Interviewing

The ECA has already begun to be used as an survey tool, but most often to train survey interviewers (c.f., inter alia, Link, Armsby, Hubal, & Guinn, 2006). In this context, the ECA plays a patient or other interviewee, so that interviewers can practice their skills. ECAs have not yet come to serve as interviewers, or to replace CASI or A-CASI technologies, most probably for the reasons that we have just outlined. First of all, the human science that ECAs depend on is still in its relative infancy. Autonomous ECAs are built on models of human behavior, and we do not yet know enough about the nonverbal aspects of rapport, or the effects of certain kinds of questions on the interview, to build an ECA that can autonomously work its way through a questionnaire, responding to feedback appropriately, and making decisions about when to ask the same question again and when to go on to the next part of the interview.

Secondly, and more damningly, it may be difficult to *ever* use ECAs as interviewers because, as Tourangeau, Couper & Steiger point out, if ECAs (and other computer systems) are treated as social actors, then they – and the interview methodologies that depend on them -- may be as subject to social desirability effects as are real humans (Tourangeau, Couper, & Steiger, 2003). Tourangeau and colleagues go on to examine the effect of representations of humans on social desirability, by showing images of people (the authors of the study) above the text of the questions on an Internet-based survey. They find little social desirability in their results, but in fact what they may have found is that simple photos may not suffice as representations of the agency of the computer. Or, that when you show the authors of the study above the questions, the subjects will not attribute agency to the *computer* but to the author of the questions asked by the computer, which will not lead them to feel that the *computer* might be judging them (Sundar & Nass, 2000).

We started this chapter discussing the paradox of survey interviewing – that rapport is felt by some to be essential for a good interview, while others see it as underlying social desirability effects. The current twist on that paradox is that the more embodied conversational humans come to resemble real humans, the more they may evoke identical responses to human interviewers. If this is the case, they will more resemble human interviewers than CASI or A-CASI technologies, and we imagine the day when ECAs will need to learn how to wear a poker face . . .

On the other hand, as we indicated above, ECAs can serve as powerful tools in social science research. They can help us to understand fundamentals of human interaction that shape interview contacts. More concretely, they can allow us to manipulate aspects of interaction – e.g. nonverbal cues – that are difficult or impossible to examine systematically in human-to-human interaction. In this context, we would suggest that ECAs can help us tease apart the very effects of social desirability on survey interviewing, and provide the kind of results that will both allow us to better train human interviewers, and to build more adequate interviewing technologies. Because non-verbal behaviors are largely unconscious, the vast majority of research on their deployment in survey interviewing has been post-hoc and anecdotal. ECAs, on the other hand, in their wizard of oz mode where an experimenter chooses each response, are infinitely controllable. For this reason they may allow us to differentiate the effects of social class, race, gender, age, and conversational style on the interview context. In fact, analogously, some research has begun to look at the effects of ECA appearance when the ECA plays the role of a tutor. Thus, Baylor & colleagues (Baylor, Rosenberg-Kima, & Plant, 2006) found that young women's stereotypic views about women engineers changed after interacting with a female engineer ECA. However, motivation to actually study engineering was more likely to be

changed when the young women interacted with a male engineer ECA. Likewise, Person (REF??) had students receive tutoring from the four ECAs pictured in Figure 2. She discovered that whereas college students preferred to learn from a tutor of their own race and gender, their learning gains were greatest when the ECA tutor was a white male.

Figure 2: Effects of Race & Gender on Tutoring

As well as research on demographics, some research (see Person, this volume) has also begun to target the effect of particular verbal and nonverbal styles of ECAs, once again on learning. In this experiment some tutors were polite and some were rude. Surprisingly, rude tutors were judged as trustworthy, interesting and better able to teach than were polite tutors, and learning gains were identical in the two conditions.

A similar paradigm could be used to study the effect on interview responses of rudeness, dialect, skin color, gender, and putatively rapport-evoking behaviors such as leaning forward, smiling etc. An identical set of questions asked by identical virtual humans, with just one single

difference – gender or age or race or posture shifts – could give valuable information about the interviewing process.

In this paper we have argued that researchers must walk between Scylla and Charybdis in using technology for survey interviews. On the one hand, technologies must be natural, intuitive, and easy to use –they must resemble human conversation to the extent possible so as to open their use to the broadest segment of the population. On the other hand, the more these technologies resemble human conversation, the more likely they are to evoke similar kinds of interviewee social desirability effects, and to be subject to similar constraints as human interviewing. For this reason, in parallel with the introduction of new technology, and attempts to understand the effects of these technologies on interviewing, there must be no slow down in the attempt to understand human conversation, and how human interaction affects our use of technology.

References

Argyle, M. (1988). *Bodily Communication*. New York: Methuen & Co. Ltd.

Athey, E. R., Coleman, J. E., Reitmans, A. P., & Lang, J. (1960). Two Experiments Showing the Effect of the Interviewer's Racial Background in Responses to Questions Concerning Racial Issues. *Journal of Applied Psychology*, 381-385.

Baylor, A., Rosenberg-Kima, R., & Plant, A. (2006). Interface Agents as Social Models: The Impact of Appearance on Females' Attitude Toward Engineering. *Proceedings of CHI 2006 (Conference on Human Factors in Computing Systems*. Montreal, Canada.

Benney, M., Riesman, D., & Star, S. A. (1956). Age and Sex in the Interview. *American Journal of Sociology, 62,* 143-152.

Bickmore, T., & Cassell, J. (1999, November 5-7). Small Talk and Conversational Storytelling in Embodied Conversational Characters. *Proceedings of American Association for Artificial Intelligence Fall Symposium on Narrative Intelligence* (pp. 87-92). Cape Cod, MA: AAAI Press.

Bickmore, T., & Cassell, J. (2005). Social Dialogue with Embodied Conversational Agents. In J. v. Kuppevelt, L. Dybkjaer & N. Bernsen (Eds.), *Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems*. New York: Kluwer Academic.

Bischoping, K., & Schuman, H. (1992). Pens and Polls in Nicaragua: An Analysis of the 1990 Pre-election Surveys. *American Journal of Political Science, 36*(2), 331-350.

Cannell, C. F., Fowler, F. J., & Marquis, K. H. (1968). The Influence of Interviewer and Respondent Psychological and Behavioral Variables on Reporting in Household Interviews. *Vital and Health Statistics, Series 2., 26*, 1-65.

Cassell, J., & Bickmore, T. (2002). Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents. *User Modeling and Adaptive Interfaces, 12,* 1-44.

Cassell, J., Nakano, Y., Bickmore, T., Sidner, C., & Rich, C. (2001, July 17-19). Non-Verbal Cues for Discourse Structure. *Proceedings of Thirty-ninth Annual Meeting of the Association of Computational Linguistics* (pp. 106-115). Toulouse, France: Association for Computational Linguistics.

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (2000). *Embodied Conversational Agents*. Cambridge: MIT Press.

Cicourel, A. U. (1964). *Method and Measurement in Sociology*. New York: Free Press.

Currivan, D. B., Nyman, A. L., Turner, C. F., & Biener, L. (2004). Does Telephone Audio Computer-Assisted Self Interviewing Improve the Accuracy of Prevalence Estimates of Youth Smoking? *Public Opinion Quarterly, 68*(4), 542-564.

Dohrenwend, B. S., Colombotos, J., & Dohrenwend, B. P. (1968). Social Distance and Interviewer Effects. *Public Opinion Quarterly, 32*(3), 410-422.

Ehrlich, J. S., & Riesman, D. (1961). Age and Authority in the Interview. *Public Opinion Quarterly, 25*(1), 39-56.

Goudy, W. J., & Potter, H. R. (1975). Interview Rapport: Demise of a Concept. *Public Opinion Quarterly, 39*(4), 529-543.

Gustafson, J., Lindberg, N., & Lundeberg, M. (1999). The August Spoken Dialogue System. *Proceedings of Eurospeech*.

Hatchett, S., & Schuman, H. (1975). White Respondents and Race of Interviewer Effects. *Public Opinion Quarterly, 39*(4), 523-528.

Henson, R., Cannell, C. F., & Lawson, S. (1973). *Effects of Interviewer Style and Question Form on Reporting of Automobile Accidents.* Ann Arbor, Michigan: Survey Research Center, University of Michigan.

Hochstim, J. R. (1967). A Critical Comparison of Three Strategies for Collecting Data from Households. *Journal of the American Statistical Association, 62*(319), 976-989.

Hyman, H. (1954). *Interviewing in Social Research*. University of Chicago Press.

Katz, D. (1942). Do Interviewers Bias Poll Results? *Public Opinion Quarterly, 6*(2), 248-268.

LaFrance, M. (1982). Posture Mirroring and Rapport. In M. Davis (Ed.), *Interaction Rhythms: Periodicity in Communicative Behavior* (pp. 279-298). New York: Human Sciences Press, Inc.

Lakin, J. L., Jefferis, V. E., Cheng, C. M., & Chartrand, T. (2003). The Chameleon Effect as Social Glue: Evidence for the Evolutionary Significance of Nonconscious Mimicry. *Journal of Nonverbal Behavior*, *27*(3), 145-162.

Lenski, G. E., & Leggett, J. C. (1960). Caste, Class and Deference in the Research Interview. *American Journal of Sociology*, *65*(5), 463-467.

Link, M. W., Armsby, P. P., Hubal, R. C., & Guinn, C. I. (2006). Accessibility and acceptance of responsive virtual human technology as a survey interviewer training tool. *Computers in human behavior*, *22*(3), 15.

Marquis, K. H., & Cannell, C. F. (1969). *A Study of Interviewer-Respondent Interaction in the Urban Employment Survey.* Ann Arbor, Michigan: Survey Research Center, University of Michigan.

Miller, P. V. (1984). Alternative Question Wording for Attitude Scale Questions in Telephone Interviews. *Public Opinion Quarterly*, *48*(4), 766-778.

Moskowitz, J. (2004). Assessment of Cigarette Smoking and Smoking Susceptibility Among Youth. *Public Opinion Quarterly*, *68*(4), 565-587.

Nakano, Y. I., Reinstein, G., Stocky, T., & Cassell, J. (2003, July 7-12). Towards a Model of Face-to-Face Grounding. *Proceedings of Annual Meeting of the Association for Computational Linguistics* (pp. 553–561). Sapporo, Japan: Association for Computational Linguistics

Nass, C., Moon, Y., & Carney, P. (1999). Are People Polite to Computers? Responses to Computer-Based Interviewing Systems. *Journal of applied social psychology*, *29*(5), 1093.

Oksenberg, L., Coleman, L., & Cannell, C. F. (1986). Interviewers' Voices and Refusal Rates in Telephone Surveys. *Public Opinion Quarterly*, *50*(1), 97-111.

Oviatt, S., MacEachern, M., & Levow, G.-A. (1998, May). Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication*, *24*(2), 87-110.

Prevost, S. A. (1996). Modeling Contrast in the Generation and Synthesis of Spoken Language. *Proceedings of International Conference on Spoken Language Processing*. Philadelphia, PA: Association for Computational Linguistics.

Reeves, B., & Nass, C. (1996). *The Media Equation: how people treat computers, televisions and new media like real people and places*. Cambridge: Cambridge University Press.

Riesman, D., & Glazer, N. (1948). The Meaning of Opinion. *Public Opinion Quarterly*, *12*(4), 633-648.

Robinson, D., & Rohde, S. (1945). A Public Opinion Study of Anti-Semitism in New York City. *American Sociological Review*, *10*(4), 511-515.

Schober, M. F., & Bloom, J. E. (2004, 3). Discourse Cues That Respondents Have Misunderstood Survey Questions. *Discourse Processes*, *38*, 287-308.

Schuman, H., & Converse, J. M. (1971). The Effects of Black and White Interviewers on White Respondents in 1968. *Public Opinion Quarterly*, *35*(1), 44-68.

Sheatsley, P. B. (1950). An Analysis of Interviewer Characteristics in Relationship to Performance, Parts I and II. *International Journal of Public Opinion Research, 4*.

Sheatsley, P. B. (1951). An Analysis of Interviewer Characteristics and Their Relationship to Performance, Part III. *International Journal of Public Opinion Research*, *5*, 193-197.

Sundar, S. S., & Nass, C. (2000, Dec). Source orientation in human-computer interaction: Programmer, networker or independent social actor? *Communication Research*, *27*(6), 683-703.

Tickle-Degnen, L., & Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, *1*(4), 285-293.

Torres, O. E., Cassell, J., & Prevost, S. (1997, July 14-16). Modeling Gaze Behavior as a Function of Discourse Structure. *Proceedings of First International Workshop on Human-Computer Conversation*. Bellagio, Italy: Intelligent Research, Ltd.

Tourangeau, R., Couper, M. P., & Steiger, D. M. (2003). Humanizing self-administered surveys: experiments on social presence in web and IVR surveys. *Computers in human behavior*, *19*(1), 24.

Van der Zouwen, J., Dijkstra, W., & Smit, J. H. (1991). Studying Interviewer Respondent Interaction: The Relationship Between Interviewing Style, Interviewer Behavior and Response Behavior. In P. Biemer, et. al. (Ed.), *Measurement Errors in Surveys*. New York: Wiley.

Villarroel, M. A., Turner, C. F., Eggleston, E., Al-Tayyib, A., Rogers, S. M., Roman, A. M., et al. (2006). Same Gender Sex in the United States: Impact of T-ACASI on Prevalence Estimates. *Public Opinion Quarterly*, *70*(2), 166-196.

Weiss, C. H. (1968). Validity of Welfare Mothers' Interview Responses. *Public Opinion Quarterly*, *32*(4), 622-633.

Weiss, C. H. (1970). Interaction in the Research Interview: The Effects of Rapport on Response. *Proceedings, Social Statistics Section. American Statistical Association*, 18-19.

Williams, J. A. (1964). Interviewer-Respondent Interaction: A Study of Bias in the Information Interview. *Sociometry*, *27*, 338-352.

Williams, J. A. (1968). Interviewer Role Performance: A Further Note on Bias in the Information Interview. *Public Opinion Quarterly*, *32*(2), 287-294.

Yan, H. (2000). *Paired Speech and Gesture Generation in Embodied Conversational Agents.* Unpublished Masters of Science, MIT, Cambridge, MA.