Journal of Official Statistics, Vol. 20, No. 3, 2004, pp. 551–570

Does Voice Matter? An Interactive Voice Response (IVR) Experiment

Mick P. Couper[1], *Eleanor Singer*[2], *and Roger Tourangeau*[3]

Audio-CASI and its telephone counterpart, interactive voice response (IVR, also called telephone ACASI), have been shown to increase the reporting of socially undesirable behaviors relative to interviewer-administered surveys. But the development of the recorded voice files is a costly and time-consuming undertaking, and may in fact reintroduce some social presence, with respondents reacting to characteristics of the voice such as gender. One potential solution to both these problems may be the use of computer-generated voices (text-to-speech systems). We conducted an experiment to explore these issues, using an IVR survey on sensitive topics. We contrasted live interviewers (CATI) and recorded human voices with two different text-to-speech (TTS) systems, one sounding more human-like, the other more machine-like. We crossed this with gender of the voice, yielding a 4*2 experiment. Equal numbers of male and female subjects were recruited by telephone from list-based samples of Michigan residents and randomly assigned to mode, yielding almost 1,400 completes. We examined the effect of gender and "humanness" of voice on the reporting of socially desirable and undesirable behaviors. We also examined respondents' reactions to the different voices and compared break-off rates across the different conditions to explore whether TTS systems could be a reasonable alternative to recorded human voices for audio-CASI and IVR applications.

Key words: Interactive voice response; IVR; audio-CASI; sensitive questions; voice.

1. Introduction

The last several years have seen widespread adoption of audio-computer assisted self-interviewing (audio-CASI or ACASI) methods, particularly for the administration of sensitive questions as part of a face-to-face survey. We are also seeing renewed interest in the use of automated telephone surveys using interactive voice response (IVR, also called telephone ACASI) for the administration of questions on a variety of sensitive topics.

[1] University of Michigan and Joint Program in Survey Methodology, P.O. Box 1248, Ann Arbor, MI 48106, U.S.A. Email: mcouper@umich.edu
[2] University of Michigan, P.O. Box 1248, Ann Arbor, MI 48106, U.S.A. Email: esinger@isr.umich.edu
[3] University of Michigan and Joint Program in Survey Methodology, 1218 LeFrak Hall, College Park, MD 20742, U.S.A. Email: rtourang@survey.umd.edu

Acknowledgments: This study was supported by a grant from NICHD (R01 HD38714-02) to Couper and Singer. Several people were instrumental in the successful execution of the study. At the Survey Research Center, Theresa Marvin was responsible for evaluating the TTS systems and producing the voice files, and assisted with data management and analysis; Jenefer Willem assisted with the preparation of the instrument, acquisition of the samples, and data management; Andrey Peytchev assisted with the analysis. At Gallup, Darby Miller Steiger and Margrethe Montgomery were responsible for overall management of the data collection, and Shelly Fowler assisted with the development and testing of the IVR voice files. We also thank the Associate Editor and the anonymous reviewers for their helpful comments.

The main motivation for this trend is the accumulation of research evidence that self-administration yields more honest reporting of socially sensitive information.

Despite the increasing use of voice-based interfaces for the automated administration of survey questions, there has been relatively little systematic research on the effect of voice on responses to sensitive questions. While the recent survey research literature suggests that voice characteristics may not be important, early studies focused on interviewer voice qualities as a factor in nonresponse and measurement error (Oksenberg, Coleman, and Cannell 1986; Oksenberg and Cannell 1988). Furthermore, there is a growing body of experimental findings in the field of human-computer interaction (HCI) and communication research suggesting that voice may indeed make a difference in surveys of sensitive questions.

This article addresses whether the type of voice makes a difference in responses to sensitive questions in audio-based computer assisted surveys. Specifically, we administer questions using live interviewers, recordings of human voices, or computer-generated voices. We also vary gender of voice. This design allows us to examine in a controlled manner whether "humanness" and gender of the voice affect responses. Furthermore, if we find no differences between recorded and computer-generated voices, this would suggest that the latter may be a cheaper alternative to the recorded voices typically used in IVR and ACASI applications.

2. Background

Since the development of audio-CASI in the early 1990s (O'Reilly et al. 1994; Johnston and Walton 1995), and with several early studies demonstrating the positive effect of audio-CASI for eliciting highly sensitive behaviors and attitudes (e.g., Turner et al. 1998a, 1998b), the use of audio-CASI has grown explosively (Gribble et al. 1999). More recent findings suggest that the benefits of audio-CASI may not be as large as first imagined (Tourangeau and Smith 1996, 1998), or, more correctly, that the major advantage lies with self-administration, and that the addition of audio output may add only marginal gains in terms of data quality (Couper, Singer and Tourangeau 2003; but see Turner et al. 1998a, who find larger gains for adolescents). Despite this, the interest in audio-CASI for surveys of sensitive behaviors has not waned.

The telephone equivalent of audio-CASI goes by various names, including interactive voice response (IVR, most often used in the market research literature), touchtone data entry (TDE, favored by the establishment survey field), and, more recently, telephone audio-CASI (a term coined by those extending audio-CASI to the telephone). We prefer to use the more common term IVR. A key distinction can be made between inbound IVR, in which the respondent initiates the call, and outbound IVR, in which an interviewer makes the initial recruitment call, then switches the respondent to the automated system after some initial questions have been asked (much like audio-CASI in the face-to-face interview). While inbound IVR surveys have been around for several decades, particularly in the world of market research (e.g., Blyth and Piper 1994) and establishment surveys (Werking, Tupek, and Clayton 1988; Phipps and Tupek 1991), there is renewed interest in the application of outbound IVR surveys for sensitive topics (Cooley et al. 2000; Gribble et al. 2000). For example, Gribble and colleagues (1999, p. 23) report that "Pilot studies indicate that most respondents prefer T-ACASI (that is, outbound IVR) to a human telephone interviewer on several dimensions. Preliminary results also indicate that participants report significantly higher levels of stigmatized or illicit activities and lower levels of normative behaviors when interviewed using T-ACASI technology." Our focus is on this latter type of IVR.

In almost every application of audio-CASI (and IVR), to our knowledge, a single female voice has been used – typically that of an experienced interviewer. This is true of several large-scale data collection efforts such as the National Household Survey on Drug Abuse (NHSDA), the National Survey of Family Growth (NSFG) and even the National Survey of Adolescent Males (NSAM). Other studies do not even mention the gender of the voice used (e.g., Gribble et al. 2000). In a CATI survey on sexual behavior, Catania et al. (1996) gave a random subset of respondents a choice of interviewer gender; 18% of female and 28% of male respondents did not make a choice. Of those who did choose, 94% of female respondents selected a female interviewer, while 55% of male respondents selected a female interviewer. We know of no IVR or ACASI studies that have offered respondents a choice.

In an early test of ACASI voice in a pilot study in Baltimore County, Maryland, Rogers et al. (1996) used both a male and a female voice. They randomly assigned respondents to one or the other voice. Rogers et al. (1996) conclude: "Our preliminary analysis indicates the voice gender does not appear to have much of an effect on responses."

The relatively more recent technology of outbound IVR for sensitive topics has similarly paid little heed to issues of voice characteristics. In their review of IVR studies, Corkrey and Parkinson (2002a) found only two studies that addressed voice. In an early application of inbound IVR for depression screening, Baer et al. (1995) alternated a male and female voice to maintain interest across the 20-item instrument, but offered no evaluation of the effectiveness of this approach. Phipps and Tupek (1991) conducted a qualitative evaluation of the digitized voice used for an establishment survey application of TDE (touchtone data entry), and found no problems with the comprehension of the voice used. In their own IVR study (which they call RVS, or recorded voice system), Corkrey and Parkinson (2002b) used voice recordings by a single female staff member. Tourangeau, Couper, and Steiger (2003) experimented with male, female, and mixed-voice (a male reads the questions and a female reads the answer categories) conditions in their IVR study. They found no effect of gender of the IVR voice on gender attitudes or the degree of disclosure on several sensitive items.

Despite the relative paucity of research on the issue of voice, Turner and colleagues (1998b), p. 486) concluded that, "even in sex surveys, the gender of the voice is unimportant." This conclusion stands in sharp contrast to the experimental literature on "computers as social actors" (CASA), to use a term coined by Reeves and Nass (1997). The main thrust of this work is that users (respondents) imbue computers with personality, and even subtle cues such as the gender of the computer voice trigger responses from users akin to what one would expect if they were interacting with a human actor. For example, in an early review of voice interfaces, Tucker and Jones (1991), p. 148) wrote: "speech introduces anthropomorphism. If speech input or output are employed, users tend to overestimate the capabilities of the machine, and may be tempted to treat the device as another person."

Several of these studies involve tasks in which the subjects interact with a computer, following which they evaluate the computer's performance. For example, Nass, Moon, and Green (1997) exposed subjects ($n = 40$) to either a male- or female-voiced computer in a tutoring task on one computer followed by an evaluation task on another computer. They used two male and two female prerecorded voices, so no subject heard the same voice in the tutor and evaluator computer. The tutor computer was rated significantly ($p < .05$) higher on friendliness and competence when the evaluator voice was male rather than female. In addition, the female-voiced computers were perceived to be more informative about "feminine" topics (love and relationships), while the male-voiced computers were seen as more informative about "masculine" topics (computers). Nass, Moon, and Green (1997, p. 874) conclude: "It thus appears that the tendency to gender stereotype is deeply ingrained in human psychology, extending even to machines. . . when voice technology is embedded in a machine interface, voice selection is highly consequential. Indeed, by choosing (or casting) a particular voice, a designer or engineer may trigger in the user's mind a whole set of expectations associated with that voice's gender."

Lee, Nass, and Brave (2000) varied the gender of a synthesized (text-to-speech or TTS) voice in a series of social dilemma situations, with 48 subjects in a 2 (respondent gender) by 2 (TTS gender) design. Their conclusions mirror those of Nass, Moon, and Green: "The male-voiced computer exerted greater influence on the user's decision than the female-voiced computer and was perceived to be more socially attractive and trustworthy. More strikingly, gendered synthesized speech triggered social identification processes, such that female subjects conformed more to the female-voiced computer, while males conformed more to the male-voiced computer" (Lee, Nass, and Brave 2000).

Some of these studies have involved the elicitation of sensitive information. For example, Nass et al. (2003) administered a 63-item survey on sexual behavior. Using a telephone interface, 100 subjects were randomly assigned to five conditions: synthesized versus recorded speech, crossed with male versus female voices, with an extra cell assigned to a standard GUI text interface. They found that synthetic speech participants refused to answer significantly more than did recorded speech participants. Furthermore, there was a significant interaction between type of voice and gender of voice ($F(1,72) = 3.96$, $p < .05$). For female voices, participants were more willing to admit inappropriate behaviors to synthesized speech as compared to recorded speech, while for male voices, participants were more willing to admit to recorded speech than synthesized speech. "In discussing the implications of these and other findings, Nass et al. (2003) note that if designers choose to implement voice interfaces, they should take special care in casting the voice. In other words, the type of voice does matter."

These studies all suggest that both gender of voice and type of voice (synthesized versus recorded) *are* relevant when designing interfaces for computer-human interaction. As Nass et al. (1997, p. 154) note in their review of this research, "choosing a computer voice's gender is one of the most important design decisions that can be made." These claims stand in sharp contrast to Turner et al.'s (1998b, p. 486) claim that gender of voice is unimportant, even in sex surveys. This apparent contradiction between the two sets of findings serves as the departure point for the present study.

The stated advantages of audio-CASI are two-fold: 1) the added sense of privacy afforded by the use of audio output and computer-based administration, and 2) possible benefits in terms of those with low levels of literacy, for whom text-based administration may not be ideal. On the other hand, audio-CASI is expensive to develop, especially given the large number of voice files that have to be recorded for the complex and customized instruments often used in such surveys. If audio-CASI is indeed a superior method for eliciting more honest responses on a wide range of sensitive topics, then efforts to reduce the costs associated with the technique would be beneficial.

In addition to the reduction of social desirability effects, IVR also offers the advantage of cost-savings over interviewer-administered telephone surveys, in that the interviewer does not stay on the line while the respondent is completing the self-administered items. But the cost savings may only be realized if the costs of recording and preparing the voice files are offset by a large sample.

We chose to focus on IVR for several reasons. First, the cost of telephone administration being significantly lower than that of face-to-face administration meant we could afford a much larger sample using IVR than ACASI. Second, doing the experiment over the telephone allows us to isolate the effects of voice. In audio-CASI, the audio may be confounded with text – the respondent is typically exposed to both, and there is some evidence that more attention is given to the text (see Caspar and Couper 1997; Couper et al. 2003). Thus, using an audio-only mode allows us to better detect any possible effects of voice on survey responses.

IVR studies are not without drawbacks, of course. Key among them are breakoffs. Gribble et al. (2000) report a breakoff rate of 24% in their T-ACASI study, compared to only 2% for those interviewed by a human. Tourangeau, Steiger, and Wilson (2002) report breakoff rates as high as 31% for various IVR studies they reviewed. Therefore, we are also interested in the effect of the voice on respondents' willingness to complete the questionnaire.

In addition to the possible effects of gender of IVR voice on survey responses, we are interested in exploring potential cost-savings associated with the use of synthesized, rather than recorded, voices. The quality of automated voice generation systems has now reached such a level that they are in widespread use in commercial applications. Whereas most ACASI and IVR survey applications use digitized voices, synthesized voices are becoming much more common. The digitized voices are recordings of live human voices, converted to digital format for use in an IVR system. The voice may sound more or less like the original depending on the quality of recording devices, the sample rate used, and other factors. In contrast, a synthesized voice is computer-generated, typically using a text-to-speech (TTS) system. Using TTS systems for survey applications may generate substantial cost savings over the laborious recording of human voices.

This study thus allows us to address two key questions: 1) do characteristics of the IVR voice (in particular, gender) affect the answers provided to sensitive questions, and 2) is a synthesized (TTS) voice a cost-effective alternative to a digitized (recorded) voice for such applications? These two questions are interrelated, and have implications for practice. For example, if we find that neither gender nor type of voice affect data quality or completion, the task of preparing voice files for IVR and ACASI could be made easier. On the other hand, if voice characteristics (such as gender) do matter, but similar effects are found for synthesized and digitized voices, we can use the more efficient TTS systems to generate voices either to match to respondent characteristics (to reduce errors) or to systematically vary other characteristics (to measure the effects of other characteristics). If both gender and source of voice (synthesized/digitized) affect the answers provided in IVR surveys, more research will be needed to minimize the potential impact on survey data quality and costs.

We expand on our hypotheses more fully in the results section, after describing details of the study design.

3. Study Design and Implementation

The experiment was designed to explore the effects of gender of voice and type of voice used in an IVR survey. Details of the design are spelled out below.

3.1. Experimental design and IVR voices

A key goal was to compare recorded and synthesized IVR voices. Given the "computers as social actors" findings, it was clear that the quality of the synthesized voice may affect the degree of social presence it conveys. For this reason we decided to test voices of different quality – more human-sounding text-to-speech (TTS) voices compared with more mechanical-sounding TTS voices. A second goal was to assess the effect of voice gender. We also wanted to contrast the various IVR voices to a CATI control group of live interviewers. This led to a 4*2 design with four voice types (live interviewer, recorded interviewer, human-like TTS, and machine-like TTS) crossed with two voice genders (male, female). In addition, we assigned roughly equal numbers of male and female respondents to each experimental condition.

After reviewing a number of TTS generators on the World Wide Web and testing a variety of TTS voices, we settled on a more human-sounding TTS system from AT&T (http://www.naturalvoices.att.com/) and a more mechanical-sounding system from Bell Labs (http://www.bell-labs.com/project/tts/). Sound files were generated on the World Wide Web using demonstration versions of each software system. Windows wave format (.wav) files were produced at a sampling rate of 22kHz. These were then converted to the system used by Gallup, Audio Works Station from BitWorks, Inc. (www.bitworks.org), which was also used to develop the recorded voice files. These were recorded at 8kHz. All the TTS voice files were also converted to a sampling rate of 8kHz for use in Gallup's IVR system.

3.2. Sample

None of the recent studies examining the effect of alternative modes of data collection on answers to sensitive question have access to validation data, and are thus forced to use the assumption that increased reporting of socially undesirable behavior (and decreased reporting of desirable behavior) implies more accurate measurement. We made similar assumptions, but also sought to gain access to publicly accessible data to strengthen the assumption. We thus used samples from three separate sources, described below. As the first two of these were available online on a state-by-state basis, for efficiency reasons we restricted our sample to residents of Michigan.

3.2.1. Nonvoter sample

Aristotle International, Inc. (www.aristotle.com) has a list of voters and nonvoters available for purchase online. An error was made in the selection of this sample. While we intended to select a list of those who had not voted in the 2000 general election, we instead obtained a list of those who had not voted in the earlier primary election. While we also had their voting record in the general election, we obviously ended up with many more voters than we had intended. The sample was purchased from Aristotle's website, www.VoterListsOnline.com. We purchased a total of 12,000 names of registered Michigan voters who did not vote in the 2000 presidential primary. The file contained name, address, phone number, gender, voting history, and other demographic information. We restricted the search to exclude records that did not include a telephone number and to exclude duplicate selections from the same address.

3.2.2. Bankruptcy sample

We purchased this sample from the Public Access to Court Electronic Records (PACER) website, http://pacer.psc.uscourts.gov. We obtained names of 4,095 persons who had declared personal bankruptcy between October 26th and December 1st, 2001 in the state of Michigan. The bankruptcy datafile included name, address, date filed, filing district, and type of bankruptcy declared. The records did not include telephone numbers, so a separate lookup was conducted. After we removed duplicate addresses and records with no residential address, we submitted 3,116 records to The Allant Group (www.allantgroup.com). Using their Telefind and PrimeFind services, they found telephone numbers for 62% of the records.

3.2.3. List sample

These two samples were supplemented with a listed sample of Michigan telephone numbers, provided by Survey Sampling, Inc. (www.surveysampling.com). Separate samples of 600 male adult residents and 600 female adult residents of Michigan were obtained.

Our goal was to have 200 completed cases (100 male respondents and 100 female respondents) in each of the 8 cells of the design (2 voice genders by 4 voice types), yielding a total of 1,600 cases. In addition, we aimed to have approximately 600 of the cases come from each of the bankruptcy and nonvoter samples, with the balance of 400 from the general list sample. The sample was combined into separate replicates and released in such a way that interviewers were unaware of the frame from which we had selected each number. They were simply told that they would be calling a sample of Michigan residents.

3.3. Survey instrument

Our experience with IVR studies suggests that respondents do not tolerate lengthy surveys in this mode. We thus identified a subset of items used in previous studies on social desirability (e.g., Couper, Singer, and Tourangeau 2003; Tourangeau, Couper, and Steiger 2003). Based on evidence from Tourangeau, Couper, and Steiger (2003) that asking a few demographic items first reduces breakoffs, CATI interviewers asked about education, age, gender, Hispanic origin and race before switching the respondent over to the IVR system. A total of 34 questions were administered by the automated system, covering the following topics:

- Gender attitudes: Five items from Kane and Macauley's (1993) study regarding the roles of men and women (e.g., "Men have more of the top jobs because they are born with more drive and ambition than women." "Men benefit from the fact that there are more women in certain kinds of jobs, such as nurses and secretaries.").
- Socially undesirable behaviors: Four items on alcohol consumption and marijuana use, four items on sexual activity, one item on personal bankruptcy, and two items on exercise and weight.
- Socially desirable behaviors: Items on voting and church attendance.
- Self-reported social desirability: A random subset of ten items from the 20-item Balanced Inventory of Desirable Responding (BIDR) Impression Management (IM) scale (Paulhus 1984).
- Debriefing questions: Six items to evaluate the interview experience (e.g., "How much was this interview like an ordinary conversation?" "How threatening were some of the questions on this survey?").

The CATI interviews used the same instrument, except for the debriefing items. For those cases assigned to CATI, the interviewer simply continued to administer the remaining items to the respondent instead of switching them to the IVR system.

Before switching the respondent to the IVR system, the interviewer explained how the system worked. Respondents were instructed to press the star (*) key to repeat a question and the pound (#) key to skip a question. This instruction was repeated in the IVR instrument. All other aspects of the survey instrument were the same across modes.

3.4. Data collection

All sample persons were sent an advance postcard in an effort to increase cooperation. A 7 + 7 call design was used: 7 attempts were made to contact the number, and 7 further attempts were made to gain cooperation from a contacted number. Data collection was done by trained Gallup interviewers and the field period lasted from March 18 to June 28, 2002. Equal numbers of male and female interviewers were employed on the study, and cases were randomly assigned to interviewers by gender. At no time was interviewer gender switched during the interview. In other words, for sample persons assigned to one of the male IVR conditions, a male interviewer did the initial calls and recruiting. Similarly, those cases assigned to a female-voice IVR condition were recruited by a female interviewer. We did not attempt to match the gender of the interviewer and the respondent.

Our goal was not generalization to a population but rather analysis of differences between experimental treatments; in Kish's (1987) terms, our focus was on randomization rather than representation. While we attempted to obtain a reasonable response rate in our data collection efforts, our primary focus was ensuring that we enrolled sufficient numbers of cases in each cell of the design. The sample was released in several replicates until the desired number of completes was obtained. The results of the calling effort are summarized in Tables 1 and 2.

Table 1. Data collection results and response rate information

Mode	A Numbers fielded	B Working residential numbers	C WRN rate (B/A)	D Completed	E Refusal	F IVR break offs	G Other non-response	H Res-ponse rate (D/B)
CATI	1,822	1,376	75.5%	400	312		664	29.1%
IVR	5,450	4,079	74.8%	996	1,061	308	2,022	24.4%
Total	7,272	5,455	75.0%	1,396	1,373	308	2,701	25.6%

Table 2. Number of completed cases per cell*

Voice gender	CATI	IVR			Total
		Recorded IVR voice	Human-like TTS voice	Machine-like TTS voice	
Male	201	173	169	166	709
Female	199	157	172	159	687
Total	400	330	341	325	1,396

*This does not include the 308 cases switched to IVR but who did not complete the instrument.

The "other nonresponse" category in Table 1 includes noncontacts, no such person at that number, language problems, and other sources of nonresponse. Overall, the response rate for the bankruptcy sample was 29.1%, compared to 23.7% for the nonvoter sample and 24.3% for the general list sample.

4. Results

The substantive analyses below are based on the set of 1,396 completed cases (see Table 2). The drop-out analysis adds the 308 respondents who were transferred to the IVR system after answering several demographic items, but failed to complete the survey.

We first address some of the operational questions – in particular, respondents' behavior on the IVR system and their reactions to the voices – before turning to an examination of the effects of the experimental manipulations on the answers provided to the key survey questions.

4.1. Drop-out analysis

One of the first questions to address is whether the TTS voices (especially the machine-like ones) increase breakoffs relative to the recorded voices. As we have already noted, breakoffs are not uncommon in IVR studies, and a large proportion of these occur at the time of the switch (see Gribble et al. 2000; Tourangeau, Steiger, and Wilson 2002). Our study is no exception. Of the 1,304 sample cases transferred to the IVR system, 308 or

23.6% did not complete the items. Of these 308, almost half (44.5%) broke off during the transfer to the IVR system, without even answering the first IVR question.

Contrary to expectation, however, we find no differences in breakoff rates across the three IVR voice types ($\chi^2 = 1.5, \text{d.f.} = 2, p > .10$): 23.1% in the recorded voice condition broke off, compared to 22.5% in the human-like TTS and 25.8% in the machine-like TTS conditions. Restricting the analysis to those who actually heard the IVR voice does not change this finding. Using survival models to explore the pattern of breakoffs after the first IVR question, we find no significant trends or differences by IVR voice type. We also find no effect of the gender of the voices used, either as main effects or in interactions with voice type.

4.2. *Missing data rates*

Another general issue relates to differential rates of missing data across the various voice types. With no human present to prompt the respondent to provide an answer and with the possibility of technical errors (e.g., pressing an out-of-range key), we expect missing data rates to be higher for the IVR conditions than for CATI. Among those who completed the survey, the mean numbers of missing items (including explicit refusals, keying errors, and time-outs in IVR), on a base on 25 items asked of all respondents, are presented in Table 3.

Table 3. Mean number of missing items (out of 25), by mode and voice type

Missing data rate	CATI	IVR		
		Recorded IVR voice	Human-like TTS voice	Machine-like TTS voice
Mean	0.50	2.15	2.24	2.34
(s.e.)	(0.052)	(0.204)	(0.199)	(0.221)

In a gender of voice by voice type ANOVA model, the only significant effect is the contrast between the CATI condition and the three IVR groups ($p < .001$). We also tested a linear contrast across the IVR conditions, and it does not reach statistical significance. Although IVR has higher rates of missing data than CATI (more than 2 percent of the answers are missing in the three IVR conditions versus about 0.5 percent in the CATI condition), we find no effect of the type of voice or the gender of voice on IVR missing data rates.

4.3. *Reactions to IVR voices*

A related question of interest is how respondents reacted to the IVR voices, in particular the computer-generated TTS voices. This is of interest for two reasons: First, to evaluate the feasibility of TTS systems for IVR and ACASI applications and, second, to serve as a manipulation check. In other words, did respondents distinguish between what we called the more human-like TTS voices and the more machine-like TTS voices?

The final question posed to the IVR respondents was "How human did my voice sound?" Response options ranged from not at all human (1) to very human (5). Only those respondents who made it this far in the survey answered the debriefing questions, although we find no evidence of differential drop-out by voice type.

The mean ratings (with missing values coded at the midpoint) and standard errors are as follows for each of the IVR voice types respectively: recorded voice, 3.93 (s.e. = 0.064), human-like TTS, 2.97 (s.e. = 0.061), machine-like TTS, 2.37 (s.e. = 0.065). (Missing data rates do not differ significantly among voice types for this item.) We are thus assured that respondents clearly differentiated between the two TTS voices, as we had intended. A two-way ANOVA revealed a significant difference among the types of IVR voice but no effect for the voice gender.

We included four additional debriefing items for the IVR conditions. The items are reproduced in Table 4. We ran three-way ANOVAs on each of these items, with voice type and gender and respondent sex as the three factors. Interestingly, we find no differences in the responses to these items by type of voice or voice gender. This suggests, at least in the case of those who made it this far in the instrument, that while there were perceived differences in the humanness of the voices used, these appear to have little effect on their rating of the interaction – respondents reacted in the same way to the computer-generated TTS voices as to the recorded IVR voices, as measured by these four items.

Table 4. Mean responses to IVR debriefing items, by voice type (standard errors in parentheses)

Debriefing questions (1 = not at all, 5 = very much)	IVR Voice type		
	Recorded IVR voice	Human-like TTS voice	Machine-like TTS voice
How much was completing this survey like taking part in an ordinary conversation?	2.62 (0.068)	2.58 (0.068)	2.57 (0.068)
How much was completing this survey like talking to an acquaintance?	2.25 (0.068)	2.40 (0.070)	2.29 (0.064)
How much was completing this survey like dealing with a machine?	3.68 (0.070)	3.72 (0.070)	3.75 (0.076)
How much was completing this survey like interacting with a computer?	3.61 (0.071)	3.76 (0.066)	3.86 (0.069)

Combined with the results showing no differential drop-out by voice type, these findings suggest that, in terms of respondent reactions, TTS voices could be considered as a reasonable alternative to recorded IVR (and by extension also audio-CASI) voices. We next turn to an examination of the substantive responses.

4.4. Substantive differences

Given the diverse set of topics (gender attitudes, both socially desirable and undesirable behaviors, and impression management) in the instrument, and the range of hypotheses we can test, we summarize the key findings here, presenting examples where appropriate.

4.4.1. Hypotheses

Our analyses tested three main hypotheses. First, we expected more disclosure in all the IVR conditions than the CATI condition. For this test, we pooled across all IVR conditions. This prediction is consistent with the literature on interviewer-administration versus self-administration effects. Second, we expected that the more human-sounding the recorded voice, the more socially desirable the responses. The social presence or computers as social actors literature would suggest that the more human-sounding voice should have effects closer to a live interviewer, while the more machine-sounding voice should have less social presence, and therefore fewer social desirability biases. We thus tested a linear effect across the IVR conditions, with the recorded human voice expected to produce the most socially desirable responses, followed by the human-like TTS voice, and then the machine-like TTS voice. Finally, we expected gender of voice effects both for live interviewers (CATI) and for the IVR voices, though we expected larger effects for the live interviewers. We expected detectable gender-of-voice effects mainly for the gender-related items, with both male and female respondents giving more pro-feminist responses in the female-voice conditions than in the male-voice conditions. We also examined possible effects of voice gender for sensitive items with possible gender implications (e.g., sexual behavior, weight, and exercise), but did not expect such effects for other items (bankruptcy, voting, alcohol and drug use, etc.).

In addition to the main effects we tested for several possible interactions. As already noted, we expected the gender-of-voice effects to be larger for CATI than IVR, and explicitly tested this interaction. In addition, we suspected that gender of voice might interact with the sex of the respondent (cf. Lee, Nass, and Brave 2000); specifically, we posited that matched genders (e.g., male respondent with male voice) would produce more disclosure of socially sensitive information.

We thus ran a series of ANOVAs (for each of the continuous items or scale scores) or logistic regressions (for each of the binary outcomes). In addition to testing the main and interaction effects of the experimental conditions and respondent gender, we also tested the following two contrasts in the models as described above: 1) CATI versus IVR, and 2) linear effects of IVR voice (recorded, human-like TTS and machine-like TTS). This was done using the PROC GLM and PROC CATMOD procedures in SAS 8.0.

4.4.2. Gender attitudes

We begin with an examination of the gender attitude items. We used a subset of items from Kane and Macauley (1993). These were found to produce gender-of-interviewer effects with telephone interviewers in their study. We modified the items to create a common agree-disagree format more suitable to IVR administration. We then combined the five items into a scale of egalitarian attitudes. None of the experimental conditions – mode of data collection (CATI or IVR), the type of voice, or the gender of the voice – had any significant effect ($p > .10$) on responses to these questions, either tested singly or in a combined scale. Furthermore, we see no discernible pattern in the responses to the five items by voice type or gender. The only effect we did find, as expected, was a large main effect of respondent gender ($F = 15.56, \text{d.f.} = 1, 1170, p < .0001$). This suggests that while the gender items did distinguish between men and women respondents, the voice manipulation appeared to have no effect on their attitudes (that is, we were unable to replicate the Kane and Macauley (1993) finding, even on the CATI cases). These results parallel the findings of Tourangeau, Couper, and Steiger (2003).

4.4.3. Responses to sensitive questions

Turning to the sensitive items – both socially undesirable and socially desirable behaviors – Table 5 presents a summary of the key findings. Several key things stand out. First, we find no differences across the three types of IVR voice (recorded human voices, human-like text-to-speech voices, and machine-like TTS voices). In terms of the quality of

Table 5. Summary of significance tests

Dependent variables	Significance test from ANOVA or logistic regression			
	CATI vs IVR	Recorded IVR vs human-like TTS vs machine-like TTS	Gender of voice	Interactions
Alcohol consumption (1 = Drink several times a week or more, 0 = other)	$p = .039$	n.s.	n.s.	None significant
Marijuana use (1 = Ever smoked marijuana, 0 = other)	$p = .030$	n.s.	n.s.	None significant
Exercise (1 = Exercise less than once per week, 0 = other)	n.s.	n.s.	n.s.	None significant
Weight (1 = 20 pounds or more overweight, 0-other)	n.s.	n.s.	n.s.	None significant
Video porn (1 = bought or rented x-rated movies/ videos in past year, 0 = other)	$p = .018$	n.s.	n.s.	None significant
Printed porn (1 = bought any sexually explicit magazines or books in past year, 0 = other)	n.s.	n.s.	n.s.	None significant
Vote (1 = Did not vote in 2000, 0 = other)	n.s.	n.s.	n.s.	None significant
Bankruptcy (1 = Ever declared personal bankruptcy, 0 = other)	n.s.	n.s.	n.s.	None significant
Church attendance (1 = Did not attend religious service in past week, 0 = other)	n.s.	n.s	n.s.	None significant
Impression management (mean)	$p < .001$	n.s.	n.s.	None significant
12-month sex partners (log(# of partners + 0.5))	$p = .015$	n.s.	n.s.	None significant
Lifetime sex partners (log(# partners + .05))	$p = .037$	n.s.	n.s.	None significant

responses obtained, TTS and recorded voices appear to be equally effective. Second, gender of interviewer or IVR voice appears to have no effect on the answers given to sensitive questions, either as a main effect or in interaction with the respondent's gender. Gender matching appears neither to hurt nor to help in eliciting sensitive information in a survey setting (but compare Catania et al. 1996). Third, none of the interactions we tested reached statistical significance. Furthermore, our inspection of the individual models revealed no consistent pattern across the variables.

The most consistent effects we found are for differences between the responses to a live CATI interviewer and to the automated IVR system. In each case where we found a difference in reporting for the sensitive items, there was greater disclosure of sensitive information in IVR than in CATI. Table 6 shows some examples. These examples also illustrate the general lack of differences among the IVR conditions reported above. While relatively modest, the differences between CATI and IVR are consistent with the literature on interviewer versus self-administration of sensitive questions.

Table 6. Percentages of respondents admitting to selected behaviors, by mode and voice type

	CATI	IVR		
		Recorded IVR voice	Human-like TTS voice	Machine-like TTS voice
Alcohol consumption (% drink several times a week or more)	13.5	16.4	18.5	17.8
Marijuana use (% ever smoked marijuana)	26.1	31.2	32.0	31.9
Video porn (% bought or rented x-rated movies/videos in past year)	4.3	7.6	7.9	9.8

For the two questions on number of sex partners (past 12 months and lifetime) we were particularly interested in interactions with respondent gender. (We truncated the raw reports at 97 and took the log of the number (plus .5) prior to analyzing these data.) As Tourangeau and Smith (1996, 1998) note, females tend to underreport and males tend to overreport the number of sex partners in the presence of an interviewer, and we would expect self-administration to close the gap between the two. We find no evidence for this in our study – both female and male respondents report higher numbers of sex partners in the IVR conditions than in CATI. We also examined whether this effect depended on whether the gender of the voice and the gender of the respondent were matched. Again we find no discernible patterns in the models.

4.5. Validation items

Finally, we turn our attention to the two questions for which we have validation data for at least a subset of the sample.

One part of our sample consisted of persons who had declared personal bankruptcy in the state of Michigan in the months preceding data collection. All respondents were asked, "Have you ever declared personal bankruptcy?" Overall, 75.7% of respondents from this sample admitted to having declared bankruptcy. The responses to this question by voice type and mode are presented in Table 7. First, we note that the percentage of respondents who did not provide an answer to this question is significantly ($p < .01$) higher for the IVR conditions than the CATI condition; however, we find no differences in the breakoff rates among the three sample sources (bankruptcy, nonvoter, and list). If we remove the cases with missing data from the analysis, we find significant differences in the rate of reporting bankruptcies, with CATI respondents significantly lower than IVR respondents (79.9% of CATI respondents who gave a substantive answer reported a bankruptcy, compared to 90.0% across the three IVR conditions), which suggests support for our hypothesis. However, when we combine the missing cases with the "no" responses (i.e., looking at the third row of numbers in Table 7), the effect of CATI versus IVR disappears. We suspect that the higher rate of "no" responses in CATI is a function of the pressure to respond, while many more IVR respondents simply avoid answering the question. But if they answer the question at all, the bankrupt IVR respondents were more likely to answer it truthfully.

Table 7. Bankruptcy sample responses to bankruptcy question, by mode and voice type

Response	CATI	IVR		
		Recorded IVR voice	Human-like TTS voice	Machine-like TTS voice
Missing, not answered	7.3	15.0	13.5	17.0
No, never declared personal bankruptcy	18.7	9.2	7.1	9.3
Yes, have declared personal bankruptcy	74.0	75.8	79.4	73.7
Total	100.0	100.0	100.0	100.0
(*n*)	(150)	(120)	(126)	(118)

The other source of validation data we have is from voting records for the 2000 election. As already noted, our file consisted of those who did not vote in the 2000 primary election. However, we also had access to voting record information for the general election. Of those sampled from the frame, 53.6% actually voted in the general election, slightly lower than the actual statewide turnout of 58%. There was a 17% turnout in the 2000 primary election, and a 58% turnout in the general election. If we assume that *all* primary voters also voted in the general election, we should expect a turnout of 49% among the primary nonvoters. The 54% rate in the sampling frame suggests some coverage error. In addition, among those who were respondents to our survey, the validated voting rate in the general election was 67.4%. This suggests the possibility of nonresponse bias, with those being more likely to vote also being more likely to participate in the survey (see Couper 1997).

Overall, 81.2% of those who answered the voting question *said* they voted, compared to 68.3% who answered the voting question and actually *did* vote according to the records. This suggests a level of vote overreporting consistent with the literature (Belli, Traugott, and Beckmann 2001). Our concern here, however, is less with the overall rate than with differences among the voice conditions. We have already found no difference among voice conditions in the level of reporting in the full sample (see Table 5). Here we focus specifically on the subset of 535 respondents for whom we have validated vote information.

Examining a cross-tabulation of the actual and reported voting status of these 535 respondents, we find that 66.4% reported voting and actually did according to the records; a further 15.3% reported not voting and the record reflects this. This represents an agreement rate of 81.7% between the reports and the records. A further 16.3% of respondents reported having voted when the records indicated they had not. Finally, 2.1% said they did not vote, but the record indicated they had. Of course, the records are not necessarily infallible, but for the sake of these analyses we assume that the voting records represent "truth." There are many ways to look at the numbers, but the one of most interest is the rate of misreporting, and whether this differs by mode and voice type. Overall, 51.5% of those whose records indicate they are nonvoters responded positively to the vote question in our survey. Table 8 shows the misreporting rates for nonvoters and voters respectively.

Table 8. Percent misreporting vote, by mode and voice type

Status according to records	CATI	IVR		
		Recorded IVR voice	Human-like TTS voice	Machine-like TTS voice
Nonvoter (*n*)	53.1% (49)	40.5% (37)	51.4% (35)	58.3% (48)
Voter (*n*)	2.0% (98)	2.2% (91)	4.3% (93)	3.6% (84)

Given the small number of voters who misreported that they did not vote, we combined these two into a disagreement rate (report $\neq$ record) and ran a logistic regression model. As can be expected given the estimates in Table 8, we find no significant differences between CATI and IVR; however, we do find a significant ($p = .015$) linear effect across the three IVR conditions for the disagreement rate. However, as can be seen from the percentages presented in Table 8, the trend is in a direction contrary to that expected, with misreporting *increasing* as the IVR voices become more machine-like. In fact, it appears that the CATI condition occupies a middle position along this continuum.

We also find an interaction between the voice type and the gender of the voice, with the stronger effects in the same direction for the female voice conditions than the male voice conditions. Given that these effects are hard to interpret, and are not found for any of the other variables we tested, we are inclined to view this as an anomaly. We do not find a similar effect when looking at the self-reported voting rate in the overall sample, nor do we find it for any of the other socially desirable variables. While this puzzling anomaly is worth further investigation, for now we reaffirm our earlier conclusions: In general, self-administration using IVR yields better quality data for sensitive items; and there are few discernible differences among the different types or gender of voice used for IVR.

5. Discussion and Conclusions

The experimental work of Nass and his colleagues at Stanford University, along with similar findings from research conducted at Carnegie Mellon University (e.g., Sproull et al. 1996; Parise et al. 1999), has raised concerns about the possible reintroduction of social desirability effects as survey designers add voice and other humanizing cues to CASI interfaces. This, together with the increasing adoption of audio-CASI and IVR techniques for surveys on a wide variety of sensitive topics, and the increased costs associated with such trends, prompted us to explore these issues in an IVR study. Our goal was two-fold:

(1) to evaluate the effect of varying voice quality and voice gender in the elicitation of socially sensitive information; and
(2) to explore the feasibility of text-to-speech (TTS) systems for creating voice files for automated interviewing applications such as IVR and audio-CASI.

We discuss the implications of our findings in terms of each of these goals in turn.

In terms of the first goal, we do find consistent differences between the responses elicited by live interviewers (CATI) and those elicited by an automated system (IVR). These effects are consistent with the already large literature on the advantages of self-administration over interviewer administration for sensitive questions (see Tourangeau and Smith 1998). That we get such differences suggests that the IVR experience (even with the recorded voice of a human interviewer) is a qualitatively different experience for respondents than interacting with a real interviewer. However, the fact that IVR surveys appear to produce higher breakoff rates highlights an important trade-off in the use of this method, and suggests the need for further research attention.

Furthermore, the fact that we find no consistent effects of the gender of the voices used, either on a series of gender-related attitudes or on a series of sensitive items involving gender (such as the number of sex partners and the purchase or use of pornographic materials), suggests that under normal interviewing conditions such as these, respondents appear to be relatively immune to these features of the interface. Why such strong effects of humanizing cues are produced in laboratory studies but not in the field is an issue for further investigation. Our findings provide support for the claim by Turner et al. (1998b), p. 486) that "even in sex surveys, the gender of the (ACASI) voice is unimportant." These findings parallel those by Tourangeau, Couper, and Steiger (2003) regarding the addition of visual social presence cues on the Web, and research by Couper, Singer, and Tourangeau (2003) on audio-CASI in a laboratory setting. Across these studies, little evidence is found to support the "computers as social actors" thesis, at least insofar as it is operationalized in a survey setting.

The findings related to the feasibility of using synthesized voices to deliver questions have intriguing implications for survey research. The lack of significant differences across the three types of IVR voice (recorded human voice, human-like TTS, and machine-like TTS) suggests potential savings of time and money in the development of audio-CASI and IVR applications. In response to debriefing questions, respondents could clearly distinguish among the different types of voices; however, voice type appears to have no effect on respondents' willingness to complete the survey (the breakoff rate) or to answer specific questions (the item missing data rate), or on the answers they do provide to such questions (substantive differences).

While this test was conducted using interactive voice response, it has clear implications for audio-CASI. Indeed, the drawbacks of IVR in terms of high breakoff and missing data rates relative to CATI are mitigated in the audio-CASI environment. We have demonstrated the feasibility of using computer-generated or text-to-speech voice files for such automated survey applications. Given computer-assisted data collection (either IVR or CASI), the raw materials for producing these TTS files already exist in digital form, reducing the marginal cost of creating such files relative to digital recording of live interviewers. The finding that using TTS does not appear to vitiate the quality of respondents' performance on the system suggests a potentially wider application of such technology in surveys. Ongoing improvements in the quality of TTS voice systems, and the wider application of such systems in the commercial world, make this an increasingly feasible and acceptable option for voice-based self-administered surveys.

6. References

Baer, L. et al. (1995). Automated Telephone Screening Survey for Depression. Journal of the American Medical Association, 273, 1943–1944.

Belli, R., Traugott, M., and Beckmann, M. (2001). What Leads to Voting Overreports? Contrasts of Overreporters to Validated Voters and Admitted Nonvoters in the American National Election Studies. Journal of Official Statistics, 17, 479–498.

Blyth, W.G. and Piper, H. (1994). Speech Recognition – New Dimension in Survey Research. Journal of the Market Research Society, 36, 183–203.

Caspar, R.A. and Couper, M.P. (1997). Using Keystroke Files to Assess Respondent Difficulties with an ACASI Instrument. Proceedings of the American Statistical Association, Section on Survey Research Methods. Alexandria: ASA, 239–244.

Catania, J.A., Binson, D., Canchola, J., Pollack, L.M., Hauck, W., and Coates, T.J. (1996). Effects of Interviewer Gender, Interviewer Choice, and Item Wording on Responses to Questions Concerning Sexual Behavior. Public Opinion Quarterly, 60, 345–375.

Cooley, P.C., Miller, H.G., Gribble, J.N., and Turner, C.F. (2000). Automating Telephone Surveys: Using T-ACASI to Obtain Data on Sensitive Topics. Computers in Human Behavior, 16, 1–11.

Corkrey, R. and Parkinson, L. (2002a). Interactive Voice Response: Review of Studies 1989–2000. Behavior Research Methods, Instruments, and Computers, 34, 342–353.

Corkrey, R. and Parkinson, L. (2002b). A Comparison of Four Computer-Based Telephone Interviewing Methods: Getting Answers to Sensitive Questions. Behavior Research Methods, Instruments, and Computers, 34, 354–363.

Couper, M.P. (1997). Survey Introductions and Data Quality. Public Opinion Quarterly, 61, 317–338.

Couper, M.P., Singer, E., and Tourangeau, R. (2003). Understanding the Effects of Audio-CASI on Self-Reports of Sensitive Behavior. Public Opinion Quarterly, 67, 385–395.

Gribble, J.N., Miller, H.G., Rogers, S.M., and Turner, C.F. (1999). Interview Mode and Measurement of Sexual Behaviors: Methodological Issues. Journal of Sex Research, 36, 16–24.

Gribble, J.N., Miller, H.G., Cooley, P.C., Catania, J.A., Pollack, L., and Turner, C.F. (2000). The Impact of T-ACASI Interviewing on Reported Drug Use among Men Who Have Sex with Men. Substance Use and Misuse, 35, 869–890.

Johnston, J. and Walton, C. (1995). Reducing Response Effects for Sensitive Questions: A Computer-Assisted Self Interview with Audio. Social Science Computer Review, 13, 304–309.

Kane, E.W. and Macauley, L.J. (1993). Interviewer Gender and Gender Attitudes. Public Opinion Quarterly, 57, 1–28.

Kish, L. (1987). Statistical Design for Research. New York: Wiley.

Lee, E.-J., Nass, C., and Brave, S. (2000). Can Computer-Generated Speech Have Gender? An Experimental Test of Gender Stereotype. CHI 2000 Extended Abstracts. New York: ACM, 289–290.

Nass, C., Moon, Y., and Green, N. (1997). Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers with Voices. Journal of Applied Social Psychology, 27, 864–876.

Nass, C., Moon, Y., Morkes, J., Kim, E.-Y., and Fogg, B.J. (1997). Computers Are Social Actors: A Review of Current Research. In Human Values and the Design of Computer Technology, B. Friedman (ed.). Stanford, CA: CSLI Press.

Nass, C., Robles, E., Bienenstock, H., Treinen, M., and Heenan, C. (2003). Voice-Based Disclosure Systems: Effects of Modality, Gender of Prompt, and Gender of User. International Journal of Speech Technology, 6, 113–121.

Oksenberg, L. and Cannell, C.F. (1988). Effects of Interviewer Vocal Characteristics on Nonresponse. In Telephone Survey Methodology, R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls, II, and J. Waksberg (eds). New York: Wiley, 257–269.

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

O'Reilly, J.M., Hubbard, M., Lessler, J., Biemer, P.P., and Turner, C.F. (1994). Audio and Video Computer Assisted Self-Interviewing: Preliminary Tests of New Technologies for Data Collection. Journal of Official Statistics, 10, 197–214.

Parise, S., Kiesler, S., Sproull, L., and Waters, K. (1999). Cooperating with Life-Like Interface Agents. Computers in Human Behavior, 15, 123–142.

Paulhus, D.L. (1984). Two-Component Models of Socially Desirable Responding. Journal of Personality and Social Psychology, 46, 598–609.

Phipps, P.A. and Tupek, A.R. (1991). Assessing Measurement Errors in a Touchtone Recognition Survey. Survey Methodology, 17, 15–26.

Reeves, B. and Nass, C. (1997). The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge: CSLI and Cambridge University Press.

Rogers, S.M., Miller, H.G., Forsyth, B.H., Smith, T.K., and Turner, C.F. (1996). Audio-CASI: The Impact of Operational Characteristics on Data Quality. Proceedings of the

American Statistical Association, Section on Survey Research Methods. Alexandria, VA: 1042–1047.

Sproull, L., Subramani, M., Kiesler, S., Walker, J., and Waters, K. (1996). When the Interface Is a Face. Human-Computer Interaction, 11, 97–124.

Tourangeau, R. and Smith, T.W. (1996). Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context. Public Opinion Quarterly, 60, 275–304.

Tourangeau, R. and Smith, T.W. (1998). Collecting Sensitive Information with Different Modes of Data Collection. In Computer Assisted Survey Information Collection, M.P. Couper, R.P. Baker, J. Bethlehem, C.Z.F. Clark, J. Martin, W.L. Nicholls, II, and J. O'Reilly (eds). New York: Wiley.

Tourangeau, R., Couper, M.P., and Steiger, D.M. (2003). Humanizing Self-Administered Surveys: Experiments on Social Presence in Web and IVR Surveys. Computers in Human Behavior, 19, 1–24.

Tourangeau, R., Steiger, D.M., and Wilson, D. (2002). Self-Administered Questions by Telephone: Evaluating Interactive Voice Response. Public Opinion Quarterly, 66, 265–278.

Tucker, P. and Jones, D.M. (1991). Voice as Interface: An Overview. International Journal of Human-Computer Interaction, 3, 145–170.

Turner, C.F., Ku, L., Rogers, S.M., Lindberg, L.D., Pleck, J.H., and Sonenstein, F.L. (1998a). Adolescent Sexual Behavior, Drug Use and Violence: Increased Reporting with Computer Survey Technology. Science, 280, 867–873.

Turner, C.F., Forsyth, B.H., O'Reilly, J.M., Cooley, P.C., Smith, T.K., Rogers, S.M., and Miller, H.G. (1998b). Automated Self-Interviewing and the Survey Measurement of Sensitive Behaviors. In Computer Assisted Survey Information Collection, M.P. Couper, R.P. Baker, J. Bethlehem, C.Z.F. Clark, J. Martin, W.L. Nicholls, II, and J. O'Reilly (eds). New York: Wiley.

Werking, G., Tupek, A., and Clayton, R.L. (1988). CATI and Touchtone Self-Response Applications for Establishment Surveys. Journal of Official Statistics, 4, 349–362.

Received February 2003
Revised March 2004