Behavior Research Methods, Instruments, & Computers
2002, 34 (3), 354-363

A comparison of four computer-based telephone interviewing methods: Getting answers to sensitive questions

ROSS CORKREY and LYNNE PARKINSON
University of Newcastle, New South Wales, Australia

Interactive voice response (IVR) technology presents a new and promising approach by which to collect accurate data on sensitive topics by telephone interviews. In a national survey of 2,880 households of alcohol and drug consumption, we compared computer-assisted telephone interviewing (CATI) and IVR with two hybrid methods that combine IVR with CATI. The principal hypothesis was that the self-report rates of sensitive behaviors would be higher for the hybrid and IVR methods owing to greater perceived confidentiality than with CATI. All the methods obtained similar sample demographic compositions. Response rates did not differ significantly between the CATI and the hybrid methods; however, the response rate with IVR was significantly lower. The hybrid and IVR methods obtained significantly higher self-report consumption rates for alcohol and marijuana and significantly higher hazardous drinking scores, as measured by the Alcohol Use Disorders Identification Test (AUDIT).

Many of the behaviors of interest to survey researchers are among those that are most difficult to gauge. These are sensitive and personal behaviors for which self-report is subject to, at least, a strong social desirability bias and, sometimes, legal consequences. Respondents may underreport behavior that they consider socially undesirable or illegal. However, these behaviors are commonly difficult to assess other than by self-report.

A question is sensitive if the respondent becomes concerned about disapproval or the possibility of an undesired consequence, such as prosecution for illegal activities. Those people with the most sensitive information are often least likely to disclose it and may, therefore, fail to respond or give a biased response (Tourangeau & Smith, 1996; Turner et al., 1998). Questions about alcohol and illicit drug use are likely to be sensitive, because they are associated with a strong social desirability response bias (Duffy & Waterton, 1984). Previously, many different survey methods have been used to measure sensitive behaviors, such as self-administered paper questionnaires (Wright, Aquilino, & Supple, 1998), face-to-face interviews (Konings, Bantebya, Caraël, Bagenda, & Mertens, 1995), computer-assisted personal interviews (Tourangeau & Smith, 1996), computer-assisted telephone interviews (CATIs; Aquilino & Lo Sciuto, 1990), computer-assisted self-interviews (Bonevski, Sanson-Fisher, Campbell, & Ireland, 1997; Millstein & Irwin, 1983), and audio computer-assisted self-interviews (Tourangeau & Smith, 1996; Turner et al., 1998). Those methods that involve self-administration and greater anonymity have been found to produce higher reporting rates on sensitive issues (Kobak et al., 1997; Tourangeau & Smith, 1996).

INTERACTIVE VOICE RESPONSE

One way to ameliorate many of the problems associated with asking sensitive questions is to improve the technology used for data collection. Interactive voice response (IVR) is a relatively new and promising technique for phone interviews that enables the rapid, accurate, and timely collection of data while maintaining respondents' privacy.

Most previous IVR studies have been nonsurvey applications in health areas, and its acceptability in household surveys is unknown (Corkrey & Parkinson, 2002). Where IVR has been used in surveys (Havice, 1989, 1990a, 1990b; Havice & Banks, 1991), it has suffered from low response rates, although it appears to have been successful in institutional surveys (e.g., McKay, Robison, & Malik, 1994; Nicholls & Appel, 1994; O'Connell, Rosen, & Clayton, 1996; Phipps & Tupek, 1991; Rosen, Clayton, & Pivetz, 1994; Werking, Tupek, & Clayton, 1988).

The hypothesis addressed in this study was that the self-report rate for a sensitive behavior would be higher with survey methods that were perceived as being more confidential. Our objectives were to compare two methods, Hybrid I and Hybrid II, of conducting CATIs on sensitive topics (alcohol and drug abuse) with the conventional CATI method and with IVR. Specific aims were to compare the following aspects between methods: (1) sample demographic profiles, (2) contact, response, cooperation, and refusal rates, (3) interview duration, (4) number of calls needed to contact respondents and to complete interviews,

The project was funded by a grant from Hunter Medical Research Institute and forms part of the doctoral studies of R.C., which is supported by an Australian Postgraduate Scholarship. We thank Maria Rees for volunteering her voice and time. Correspondence concerning this article should be addressed to R. Corkrey, P. O. Box 491, Wallsend, NSW 2287, Australia (e-mail: corkrey@optusnet.com.au).

(5) item nonresponse rates, (6) costs, (7) self-report rates, and (8) acceptability.

METHOD

Design

An Australia-wide telephone survey of households was conducted in 2000 using four different telephone interviewing methods. A follow-up CATI survey assessed the acceptability of each interviewing method.

Setting

The Australian population numbers 17,892,423, housed within 7,175,237 households (Australian Bureau of Statistics [ABS], 1996), and is distributed across six states and two territories. In 1998, the fixed telephone coverage was high, ranging from 96.0% to 97.6% between states and territories, except for the Northern Territory, with a relatively low coverage of 91.4%. The proportion of households paying bills or transferring funds by telephone ranges from 27.4% to 44.0% between states and territories (ABS, 1998), indicating that Australians are increasingly familiar with IVR technology.

Sample

A total of 2,880 households with fixed telephone connections were selected using simple random sampling from an electronic version of the Telstra White Pages (Desktop Marketing System, 2000) covering all the states of Australia.

Procedure

Interviewer training. Five experienced telephone interviewers attended a 5-h briefing session covering survey aims, interviewing standards, script familiarization, software, and practice interviews.

Interview methods. Four computer-based telephone interviewing methods were used, herein called: CATI, IVR, Hybrid I, and Hybrid II. For the CATI, an interviewer conducted telephone interviews and entered responses directly into the computer interface. We refer to the IVR hardware as the *recorded voice system* (RVS). For the IVR, the RVS rang households directly and conducted the interview. For the hybrid methods, an interviewer initiated the interviews, but the remainder of the interview was conducted by either an interviewer or RVS. In Hybrid I, only questions involving alcohol and drug items were asked by the RVS, after which the call was transferred back to the interviewer. In Hybrid II, questions involving demographic items were also asked by the RVS, and the call was then terminated.

Figure 1 shows the stages involved in each interview method and the questionnaire modules used by either an interviewer or RVS.

The same questionnaire was presented for each method, but specific domains were asked by either an interviewer or RVS. Respondents with rotary telephones or improperly configured touchphones were automatically reassigned to the CATI method.

To deter respondents from hanging up during Hybrid I, interviewers indicated that they would immediately ring back if the respondent was inadvertently disconnected from the RVS. This was also done if they were disconnected for some other reason than a hang-up.

It was expected that all the questions during the CATI and those asked by the interviewer with the hybrid methods would not be perceived to be as confidential as the questions asked by the RVS.

Respondent instruction. Information letters printed on letterhead stationery and addressed to "The Household" were posted using DL-sized envelopes, with the institution details plainly marked, 1 week before the first call. The letters stated that the interview would be voluntary and would involve questions about alcohol and drug use. No remuneration was offered. Hybrid and IVR letters explained that a recorded voice would be used. The term *recorded voice* was used to avoid the words *technology* or *computers*.

A simple instruction sheet showed a picture of a typical touchphone keypad. It indicated that pressing the star key (*) would repeat a question and pressing the hash key (#) returned to the previous question. The instructions were repeated at the start of the IVR interview and by the interviewers before the start of the hybrid methods. They were given again before the drug items in the IVR and hybrid methods and before the demographic items in IVR and Hybrid II.

Respondent recruitment. Assignment of respondents to each method was unknown to interviewers until an interview had begun. In an attempt to minimize within-household selection bias, the *last birthday* method (Lavrakas, Bauman, & Merkle, 1993) was used for selecting respondents within a household. This consisted of asking for the person 18 years or older who had the most recent birthday. No proxies were used. Interviewers recorded businesses as out of scope. For households, if the eligible person was unavailable, they made an appointment to ring back. In the IVR method, the RVS distinguished answering machines from genuine individuals by the length of the salutation ("Hallo?" vs. "Hallo. We're not in at present . . ."). If con-

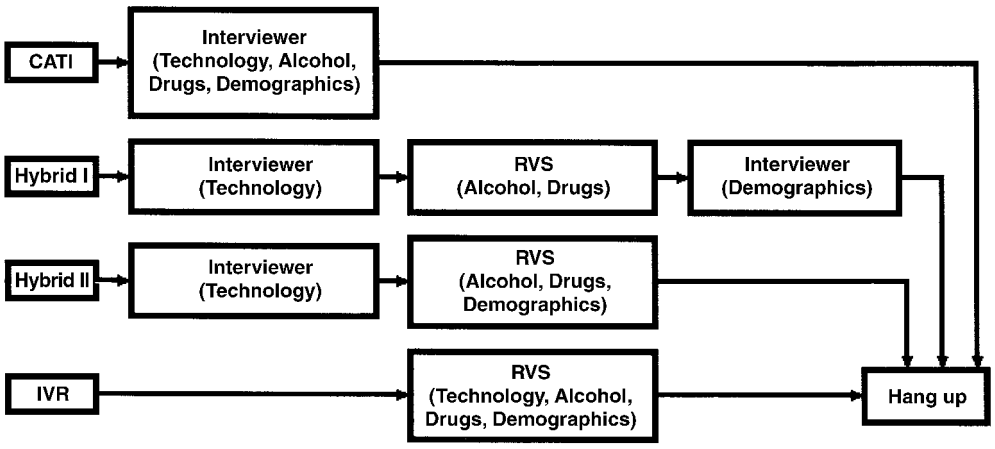

Figure 1. Interview procedure by method. CATI , computer-assisted telephone interview; IVR, interactive voice response; RVS, recorded voice system.

nected, it asked whether the telephone number belonged to a business or a residence. Business numbers were recorded as out of scope. For households, the system asked to speak to the eligible person. If the eligible person was unavailable, it offered to call back at a date and time convenient to the household. From this point on, if the call was terminated by the respondent, the interview outcome was recorded as a refusal.

Call scheduling. Higher contact rates can be obtained by scheduling calls at appropriate times and by efficient rescheduling for noncontacts (Bennett & Steel, 2000; Kulka & Weeks, 1988). Initial calls were made on weekday afternoons or evenings. Interviewers and the RVS rang back noncontacted numbers at regular intervals. Interviewers made at least seven call attempts and usually called back noncontacts within a few hours or, if there was still no contact, at least a day later. In the IVR method, noncontacted numbers were reattempted at alternating 30-min and 18-h intervals. There was no calling limit for the IVR method.

Apparatus

Equipment. The interviewing stations and RVS were Pentium II computers. A four line Dialogic D/41H voice card was installed in the RVS. Interviewers used a mouse or a keyboard to enter responses. The RVS played sound files, and respondents answered questions by pressing keys on their telephones. Seven telephone lines were used, three for the interviewers and four for the RVS. In the hybrid methods, interviewers transferred calls to vacant RVS lines. Voice recordings were made in 16-bit mono 11-kHz format with an Optimus 33-3104 omnidirectional microphone and a Creative Sound Blaster Vibra 128 sound card. Voice recordings were made in English by a single female staff member selected using a voice assessment method based on Oksenberg, Coleman, and Cannell (1986).

Software. All the methods and the follow-up CATI were implemented with a single software, Generalised Electronic Interviewing System (GEIS), written by the first author using SAS 6.12 (SAS Institute, 1991). Data were accumulated on a central computer using SAS/SHARE software (SAS Institute, 1991). GEIS automatically stored control information, including interview duration, number of call attempts, call outcomes, and so on.

In all the methods, answers to questions were provided by selecting one of a set of options, entering a number or date, or entering an open-ended response. Answers to open-ended questions were entered verbatim by interviewers, whereas the RVS allowed respondents to record a short spoken sentence. For numeric answers, absolute and reasonable limits prevented range errors. If the respondent entered an invalid response to the RVS, the interface would advise the respondent of the error and would repeat the question. In the CATI method, an incorrect entry triggered an appropriate message to be displayed by the graphical interface. After three repetitions of an item, a nonresponse caused the RVS to hang up and record a refusal. In all the methods, respondents could refuse to answer a particular item and could return to earlier questions and modify their answers.

Measures. Script items were kept as similar as possible between methods; however, some RVS questions were broken into a series of shorter questions, because of the more limited data entry capabilities of a telephone keypad (Schumacher, Hardzinski, & Schwartz, 1995).

Questions were grouped into domains (technology, alcohol, drugs, and demographics). The technology domain is not reported in detail here. The alcohol domain included consumption questions and the five-item Alcohol Use Disorders Identification Test (AUDIT; Piccinelli et al., 1997; Saunders, Aasland, Babor, de la Fuente, & Grant, 1993). The drugs domain covered amphetamines, marijuana, and heroin, which are the more commonly used drugs in Australia (Commonwealth Department of Health and Family Services, 1995); respondents were asked for the age at which they first consumed the drug, use in the previous 12 months, and frequency of use. The items were always presented in what was considered to be the order of increasing sensitivity (alcohol, marijuana, amphetamines, heroin). This avoided disproportionate break-offs in the RVS interviews by ensuring that they always began with the least sensitive questions. The demographic domain included date of birth, education, marital status, sex, country of birth, and employment status.

To assess acceptability, a follow-up CATI was automatically scheduled after an interview was terminated. The follow-up sample consisted of a random selection of one third of those who completed at least part of an interview or hung up but excluded those who had explicitly refused when talking to an interviewer. The follow-up CATI used a single interviewer who had not participated in the main interview. A follow-up CATI was used, rather than measuring acceptability at the end of each interview. If acceptability had been measured within the main interview, it would have been confounded with interview method. To minimize concerns about perceived anonymity, respondents were not told that the follow-up would take place but were advised that a further contact might occur for quality control reasons. Respondents rated the previous interview for ease, enjoyableness, stressfulness, and likeableness, using a 5-point Likert scale (Maxim, 1999, p. 233) and standard questions from Bonevski et al. (1997). In addition, CATI and hybrid respondents were asked whether they would have preferred a recorded voice to ask about alcohol and drug use, whereas IVR respondents were asked if they would have preferred a human interviewer. All were asked whether they thought people would be more honest with a human interviewer than with a recorded voice.

RESULTS AND DISCUSSION

The Hybrid I and II methods performed as well as the traditional CATI method on most criteria. By comparison, the usefulness of IVR in household surveys appears to be more restricted by its low response rate.

Sample

Analyses were conducted with SAS/STAT software (SAS Institute, 1999), except where otherwise stated. The sample demographic statistics consisted of tabulations of proportions. Therefore, chi-square goodness-of-fit tests (Snedecor & Cochran, 1980) were used to test equality of sample demographic statistics by nominal method, as well as by equivalence to the ABS Census (1996).

Sample age distributions from Hybrid I, Hybrid II, and IVR did not differ significantly from CATI, as is shown in Table 1. However, in comparison with the 1996 census, CATI marginally overrepresented the oldest age group (58+), whereas Hybrids I and II overrepresented the 31–42 age group. All the methods produced similar education, marital status, Australian-born, and employed distributions, but in comparison with the ABS census (1996) Hybrid I overrepresented the married/de-facto category and underrepresented employed persons, CATI, Hybrid II, and IVR overrepresented Australian-born respondents, and all the methods overrepresented females. None of the conclusions was altered if the oldest age group (58+) was deleted from the analyses.

Table 1 details the demographic characteristics of the sample for each method.

Chi-square analysis was used to examine independence of touchphone ownership by each of the demographic vari-

Table 1
Sample Demographic Composition by Method Compared With the Australian Population

		CATI		Hybrid I		Hybrid II		IVR		ABS*
Variable	Levels	*n*	%	*n*	%	*n*	%	*n*	%	%
Sex	male	116	38	108	34	106	42	28	34	49*
p values	ABS $\chi^2(p)$†	16.0 (<.01)		26.0 (< .01)		5.1 (.02)		9.1 (<.01)		
	CATI $\chi^2(p)$‡			0.6 (.43)		1.1 (.20)		0.7 (.41)		
Age groups	18–30	63	20	50	20	50	20	20	23	26
	31–42	75	24	99	24	79	31	18	21	25
	43–57	79	26	82	26	70	27	31	36	25
	58+	92	30	83	30	58	23	17	20	24
p values	ABS $\chi^2(p)$†	8.6 (.04)		18.1 (<.01)		9.4 (.02)		6.1 (.11)		
	CATI $\chi^2(p)$‡			5.3 (.15)		5.1 (.17)		5.5 (.14)		
Education	school or none	181	58	175	55	137	50	47	55	54
	vocational	69	22	77	24	63	23	16	19	25
	university	65	21	69	22	74	27	23	27	22
p values	ABS $\chi^2(p)$†	1.8 (0.41)		0.072 (0.97)		4.9 (0.09)		2.4 (0.31)		
	CATI $\chi^2(p)$‡			0.60 (0.74)		4.1 (0.13)		1.6 (0.45)		
Marital Status	married / de facto	187	61	200	64	168	66	58	68	58
	divorced/separated	36	13	36	12	28	11	9	11	11
	widowed	25	8	32	10	13	5	2	2	7
	never married	57	19	46	15	44	17	16	18	24
p values	ABS $\chi^2(p)$†	6.4 (.10)		17.5 (<.01)		8.5 (.04)		5.1 (.17)		
	CATI $\chi^2(p)$‡			2.6 (.45)		3.1 (.38)		4.1 (.24)		
Birth place	Australia	248	79	239	75	218	79	73	83	73
p values	ABS $\chi^2(p)$†	4.7 (.03)		0.30 (.58)		4.1 (.04)		4.2 (.04)		
	CATI $\chi^2(p)$‡			1.5 (.23)		0.01 (.99)		0.75 (.38)		
Employment	in labor force	182	58	177	55	175	64	54	63	62
p values	ABS $\chi^2(p)$†	2.0 (.16)		5.8 (.02)		0.86 (.35)		0.033 (.86)		
	CATI $\chi^2(p)$‡			0.45 (.50)		2.49 (.12)		0.65 (.42)		

Note—CATI, computer-assisted telephone interview; IVR, interactive voice response. *Population at least 18 years old (Australian Bureau of Statistics, 1996). †Chi-square goodness-of-fit tests comparing each method with the ABS census (1996), two-sided *p* values. ‡Chi-square analyses comparing Hybrid I, Hybrid II, and IVR with CATI, two-sided *p* values.

ables. Ownership of touchphones did not depend on sex or being Australian-born but was lower for those in the oldest age group (58+), who were widowed, without postschool qualifications, or were not in the labor force. However, if the oldest age group was omitted, then none of the comparisons remained significant.

Table 2 summarizes demographic characteristics by touchphone ownership.

All samples had an excess of female respondents, which is commonly found in telephone surveys (Romuald & Haggard, 1994). Hybrid I, Hybrid II, and IVR did not differ significantly from CATI in any demographic characteristic, but differences were found from the ABS census (1996). It appears that relative sampling biases between methods were small, as compared with the bias introduced by telephone sampling itself.

Apart from those over 58 years of age, touchphone ownership did not depend on demographic profile. This indicates that bias in IVR surveys introduced by touchphone ownership is minor, unless the survey includes older respondents.

Contact, Response, Cooperation, and Refusal Rates

The response, refusal, and contact rates were defined as the number of interviews, refusals, and contacted cases, respectively, each divided by the total number of interviews, refusals, break-offs, ineligible cases, cases with unknown eligibility, and noncontacts. The cooperation rate was the number of interviews divided by the total number of interviews, refusals, and break-offs. All the rates were compared using contingency table analysis (Everitt, 1992).

All the methods had similar contact rates, but IVR obtained a response rate significantly less than that for CATI. Hybrid II had a significantly lower cooperation rate and a higher refusal rate than did CATI.

Table 3 summarizes the sample sizes and response rates.

All the methods contacted similar proportions of their respective samples, but IVR respondents were less cooperative, with fewer responding and more refusing. We obtained a CATI response rate of 61.2%, which although moderate, may occur with sensitive telephone surveys (Krebs, 1994). It was considerably higher than the response rate of postal surveys (Fox, Crask, & Jooghoon, 1988) and most e-mail and Web surveys (Couper, Traugott, & Lamias, 2001; Schaefer & Dillman, 1998; Sheehan & Hoy, 1999; Smith, 1997).

The response rate of the methods may have been reduced by addressing letters to "The Household," making them appear to be junk mail and be thrown out without reading or delivered late by the postal service. The hybrid and IVR letters may have also dissuaded respondents by stating that

Table 2
Sample Demographic Composition by Touchphone Ownership

Variable	Levels	Touchphone Ownership *n*	Touchphone Ownership %	All Age Groups (χ^2)*	Ages 18–57 (χ^2)*
Sex	male	226	92	$\chi^2(1, n = 660) = 1.0, p = .31$	$\chi^2(1, n = 502) = 0.04, p = .85$
	female	373	90		
Age groups	18–30	118	95	$\chi^2(3, n = 660) = 30.8, p < .01$	$\chi^2(1, n = 502) = 1.2, p = .56$
	31–42	181	93		
	43–57	174	95		
	58+	126	80		
Education	school or none	304	88	$\chi^2(1, n = 657) = 5.8, p = .02$	$\chi^2(1, n = 499) = 1.2, p = .28$
	postschool	292	94		
Marital status	married / de facto	393	92	$\chi^2(3, n = 656) = 25.3, p < .01$	$\chi^2(3, n = 498) = 0.6, p = .91$
	divorced/separated	68	93		
	widowed	33	70		
	never married	101	92		
Birth place	Australia	459	91	$\chi^2(1, n = 654) = 0.004, p = .95$	$\chi^2(1, n = 496) = 0.01, p = .91$
	elsewhere	134	91		
Employment	In labor force	394	94	$\chi^2(1, n = 648) = 16.5, p < .01$	$\chi^2(1, n = 493) = 0.89, p = .35$
	not in labor force	193	84		

*Chi-square analyses, two-sided *p* values.

the RVS would be used. The low IVR response rate, which is consistent with earlier work (Havice, 1989, 1990a, 1990b), suggests that IVR requires motivated respondents or that it should be restricted to institutional surveys to which respondents may be obliged to respond. Household survey respondents usually have no reason to cooperate beyond altruism.

Interview Duration

In Hybrids I and II, there were two intervals corresponding to whether the respondent interacted with an interviewer or RVS, which were summed to obtain total interview duration. The statistical distribution of duration was distinctly nonnormal, and therefore, nonparametric tests were used.

Equality of duration between methods was tested with Kruskal–Wallis tests (Daniel, 1978). The median total duration differed significantly for all the methods [$\chi^2(3, n = 1{,}006) = 615.9, p < .01$], with CATI being the briefest and Hybrid I the longest. The median interviewer duration also differed significantly [$\chi^2(2, n = 917) = 502.1, p < .01$] between methods, with Hybrid II being the briefest and Hybrid I the longest. Equality of variability of interviewer duration was compared between the CATI, Hybrid I, and Hybrid II methods with the Ansari–Bradley test (Ansari & Bradley, 1960) after subtracting a Hodges–Lehmann estimate (Fligner, 1988), using StatXact3 (Cytel Software Corporation). The variability of the alcohol domain duration did not differ significantly between CATI and Hybrid I ($Z = 1.1, n = 656, p = .13$), but CATI was more variable than Hybrid II ($Z = 3.1, n = 658, p < .01$) or IVR ($Z = 5.9, n = 478, p < .01$).

Figure 2 shows the interview duration for each method, representing parts of the interview (interviewer and RVS) and the whole interview. The central bar of each box plot indicates the median interview duration, the central box indicates the 25th and 75th percentiles, and the whiskers indicate the 10th and 90th percentiles.

Hybrid I had the longest overall duration since it involved the respondent's being transferred twice and additional RVS explanatory messages. The Hybrid II interviewer duration was least, because most questions were asked by the RVS. The RVS duration was less variable than the interviewer duration, because the RVS were always read at the same pace, suggesting that this method may produce more consistent data.

Number of Calls

Since the number of calls required to contact respondents and to complete interviews consisted of count data, a Pois-

Table 3
Sample Size and Response Rates by Nominal Method

Rates	CATI ($n = 661$) *n*	CATI %	Hybrid I ($n = 706$) *n*	Hybrid I %	Hybrid I $\chi^2(p)$	Hybrid II ($n = 697$) *n*	Hybrid II %	Hybrid II $\chi^2(p)$	IVR ($n = 816$) *n*	IVR %	IVR $\chi^2(p)$
Contact	432	84	478	83	0.04 (0.84)	482	85	0.20 (0.66)	663	86	0.91 (0.34)
Response	316	61	322	56	2.96 (0.09)	319	56	2.99 (0.08)	90	12	353.3 (<0.01)
Cooperation	316	73	322	67	3.62 (0.06)	203	66	5.21 (0.02)	90	14	397.9 (<0.01)
Refusal	109	21	143	25	2.20 (0.14)	153	27	4.91 (0.03)	573	74	347.8 (<0.01)
Noncontacts	91	18	109	19	0.33 (0.58)	97	17	0.07 (0.81)	111	14	2.5 (0.12)
Out of scope	145	22	132	19	2.2 (0.14)	128	18	2.7 (0.10)	42	5	93.1 (<0.01)

Note—The nominal method does not include the transfer of some cases to CATI. Chi-square analyses compared rates for each method with that for CATI, two-sided *p* values. CATI, computer-assisted telephone interview; IVR, interactive voice response.

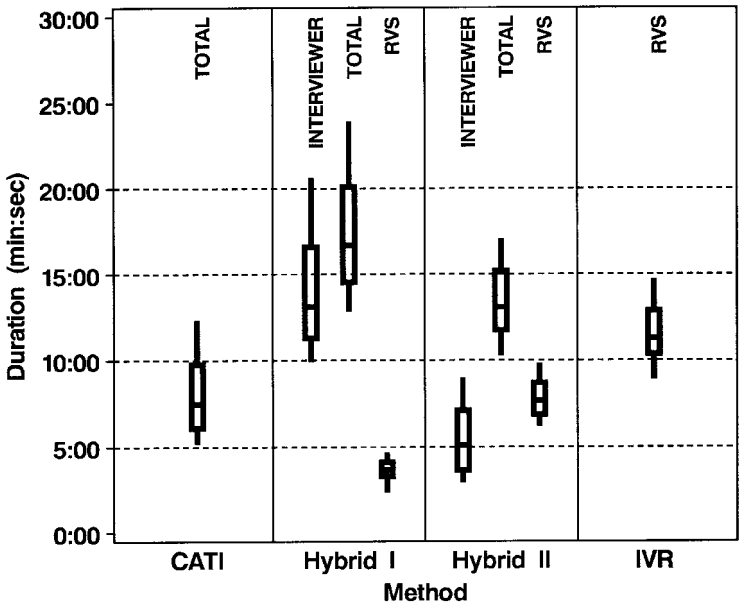

Figure 2. Interview duration by method. CATI , computer-assisted telephone interview; IVR, interactive voice response; RVS, recorded voice system.

son regression with adjustment for overdispersion (Snedecor & Cochran, 1980) was used to test for equality between methods. The number of calls required to complete interviews did not differ significantly between methods, but the number of calls required to contact respondents was significantly higher for IVR. Bootstrap simulations (Beran, 1986) using 1,000 replicates indicated that the former test had low empirical power ($p = .11$) to detect the small observed differences, whereas the latter test had good power ($p = .62$). Although the methods could not be distinguished by the number of calls required to complete interviews, IVR obtained a greater number of contact calls. This was because the IVR method did not have an upper calling limit and could continue indefinitely if needed.

Item Nonresponse Rates

Item nonresponse rates were calculated as the proportion of refused items per domain from complete interviews. Since these are proportions, contingency table analysis (Everitt, 1992) was used to test equality between methods. In comparison with CATI, IVR had higher rates for the drugs domain, whereas Hybrid II and IVR had higher rates for the demographics domain, mostly owing to refusals to respond to age and income items. Break-offs that occurred in Hybrids I and II were due to hang-ups. Most occurred in the alcohol domain in Hybrids I and II and in the drugs and demographics domains for Hybrid II. In a minority of these cases, the respondents hung up after one or more failures to enter data correctly, but the rest hung up without attempting to answer the question last asked. Very few break-offs were the result of technical difficulties.

Most break-offs occurred in the alcohol domain for Hybrids I and II and in the drugs and demographics domains for Hybrid II. Hybrid I respondents knew that they would be transferred back to an interviewer, which probably reduced the perceived anonymity of the method and increased item nonresponse rates, whereas Hybrid II respondents may have interpreted its confidentiality as an opportunity to hang up. The lower item nonresponse and lower self-report rates in the alcohol and drugs domains for CATI may be due to respondents' providing socially acceptable answers, rather than refusing to answer. It seems unlikely that nonresponses to the Hybrid II demographic items were due to a lack of human contact, since IVR obtained a similar result. Instead, the higher item nonresponse rates for Hybrid II and IVR in the demographic domain may have been due to a reluctance to provide information apparently unrelated to the survey topic, alcohol and drugs, or a resistance to providing age and income to a computer.

Costs

Costs were calculated by summing salary cost and accumulated telephone call charges. Salary cost was calculated by multiplying the total of interview duration, including time spent waiting on hold in Hybrid I, and noncontact attempt duration by the interviewer's salary rate. The mean completed interview cost and total survey cost per completed interview both differed significantly between methods, using an analysis of variance. IVR showed the minimum interview cost, Hybrid I had the largest interview cost and survey cost per completed interview, and Hybrid II had the minimum survey cost per completed interview.

Table 4 details the number of calls and break-offs, item nonresponse rates, and costs by method.

Hybrid I had the greatest interview cost and survey cost per completed interview, because interviewers had to wait until the RVS interview was completed. IVR obtained the

Table 4
Mean Number of Calls, Break-Offs, RVS Break-Off Cause, Item Nonresponse Rate, and Mean Cost by Method

Variable	Level	CATI	Hybrid I	Hybrid II	IVR	Analysis
Mean calls	to contact	2.12	2.14	1.94	2.59	$F(3,2305) = 9.85, p < .01$*
	to complete	2.86	2.82	2.85	2.65	$F(3,1045) = 0.25, p = .86$*
Break-offs	technology	0.0%	0.0%	0.0%	0.0%	-
	alcohol	0.5%	2.5%	6.7%	1.1%	$\chi^2(3, n = 1{,}054) = 25.2, p < .01$†
	drugs	0.0%	1.1%	2.4%	0.0%	$\chi^2(3, n = 1{,}054) = 11.0, p = .01$†
	demographics	0.0%	0.4%	10.5%	0.0%	$\chi^2(3, n = 1{,}054) = 77.4, p < .01$†
RVS break-offs	hang-up,					
	no data entered	–	8	43	1	
	hang-up after data entered	–	3	6	0	
	technical error	–	0	8	0	
Item nonresponse	technology	0.0%	0.4%	0.4%	0.0%	$\chi^2(3, n = 6{,}006) = 9.6, p = .02$†
	alcohol	0.0%	0.0%	0.0%	0.0%	–
	drugs	0.0%	0.3%	0.5%	1.1%	$\chi^2(3, n = 4{,}041) = 13.5, p < .01$†
	demographics	1.3%	0.8%	7.6%	7.4%	$\chi^2(3, n = 11{,}065) = 322, p < .01$†
Mean cost	completed	\$4.97	\$9.46	\$3.64	\$2.27	$F(3,1043) = 310.9, p < .01$‡
	total	\$6.03	\$11.44	\$5.27	\$7.92	$F(3,1043) = 515.9, p < .01$‡

Note—Break-offs indicate the percentages of individuals terminating interviews. Mean cost includes the mean completed interview costs and the mean total survey cost per completed interview (Australian dollars). CATI, computer-assisted telephone interview; IVR, interactive voice response. *Poisson regression. †Chi-square test for equal proportions. ‡Analysis of variance.

least mean interview cost, since it lacked a salary component. Hybrid II obtained the lowest total survey cost per completed interview, reflecting the method's efficiency: Interviewers concentrated on contacting and persuading respondents, rather than conducting interviews. If enough interviewers are available, the cost of the Hybrid I method could be minimized by having the interviewer move on the next call while a different interviewer picked up the call transferred from the RVS. In addition, the Hybrid II method may be further improved by having the software automatically schedule a return call in the case of a break-off.

Self-Report Rates

Since self-reported alcohol and drug use measures were proportions, contingency table analyses (Everitt, 1992) were used to test for equality between all the methods and between CATI and the other methods. Median age of first consumption was compared using the Wilcoxon–Mann–Whitney tests (Stokes, Davis, & Koch, 2000), owing to its nonnormal distribution.

As is shown in Table 5, a significantly higher proportion of respondents reported consuming alcohol or marijuana, and consuming hazardous levels of alcohol, with the non-CATI methods than with CATI. The proportion of respondents using marijuana at least monthly differed significantly between methods, with Hybrid II and CATI being the highest, whereas the proportion using amphetamines in the previous 12 months differed significantly, with Hybrid II and IVR being the highest. The ages at which respondents first tried alcohol or a drug did not vary significantly.

The higher reporting rates for consuming alcohol and marijuana for the non-CATI methods support the hypothesis of higher reporting rates with more anonymous methods for sensitive behaviors. In other studies, higher self-report rates were found for alcohol and drug consumption with self-administered questionnaires than with CATI (Aquilino, 1994), for alcohol consumption with a computer than with face-to-face interviews (Duffy & Waterton, 1984), for drug consumption and sex partners with self-administered methods than with personal interviewing (Jobe, Pratt, Tourangeau, Baldwin, & Rasinski, 1997; Tourangeau, Jobe, Pratt, & Rasinski, 1994; Tourangeau & Smith, 1996), and for sensitive personal health characteristics with e-mail than with a postal survey (Kiesler & Sproull, 1986).

Acceptability

The median time until the follow-up CATI that measured acceptability was 3 days, 6 h. Equality of acceptability scores were compared between methods by using Kruskal–Wallis tests and, when significant, by using Wilcoxon tests and Bonferroni adjustments (Meddis, 1984) to compare CATI scores with those for the other methods. After ordering the methods according to the degree of automation (CATI, Hybrid I, Hybrid II, IVR), a trend in the proportion of respondents preferring a human interviewer or agreeing with the statement that people are more likely to be honest about their alcohol use to a person than to a recorded voice was tested with a Cochran–Armitage Trend test with modified ridit scores (Margolin, 1988). This is a test for trends in binomial proportions.

The methods were not distinguished by stressfulness or likeableness, but differences were found for ease and enjoyableness. After Bonferroni corrections, Hybrid II was rated significantly less easy than CATI [$T(n = 157) = 5{,}490, p < .01$], and Hybrids I and II were rated less enjoyable than CATI [$T(n = 368) = 15{,}444, p < .01$; $T(n = 157) = 5{,}538, p < .01$]. With increasing automation (CATI to IVR), we found significant decreasing trends for preference for a human interviewer and for the belief that peo-

Table 5
Alcohol, Marijuana, Amphetamines, and Heroin Statistics by Method

	CATI		Hybrid I		Hybrid II		IVR		Comparisons	
Topic	*n*	%	*n*	%	*n*	%	*n*	%	Any Difference	CATI Versus Rest
Alcohol										
Full glass	331	91.2	249	96.9	258	94.2	84	96.6	$\chi^2 = 9.9, p = .02$	$\chi^2 = 8.0, p < .01$†
Median age first tried	17		17		17		17		$\chi^2 = 5.4, p = .14$	$Z = 0.1, p = 94$‡
Audit score*	65	21.2	60	28.6	73	32.4	21	30.4	$\chi^2 = 9.4, p = .03$	$\chi^2 = 8.5, p < .01$†
Marijuana										
Ever tried	114	29.3	92	34.6	102	37.8	32	36.8	$\chi^2 = 5.8, p = .12$	$\chi^2 = 5.2, p = .02$†
Median age first tried	18		18		17		20		$\chi^2 = 7.7, p = .05$	$Z = 0.6, p = .52$‡
Used last 12 months	30	7.7	20	7.4	27	9.2	11	12.2	$\chi^2 = 2.6, p = .46$	$\chi^2 = 0.41, p = .52$†
Use at least monthly	14	1.8	5	1.8	17	5.8	2	2.2	$\chi^2 = 6.8, p = .07$	$\chi^2 = .02, p = .96$†
Amphetamines										
Ever tried	24	6.2	20	7.6	22	8.2	4	4.6	$\chi^2 = 2.0, p = .58$	$\chi^2 = 0.6, p = .45$†
Median age first tried	18		19		18.5		16		$\chi^2 = 3.7, p = .29$	$Z = 0.7, p = .50$‡
Used last 12 months	4	1.0	3	1.1	10	3.4	4	4.4	$\chi^2 = 8.6, p = .03$	$\chi^2 = 3.0, p = .08$†
Use at least monthly	0	0	1	0.4	2	0.7	2	2.2	$\chi^2 = 7.8, p = .05$	$\chi^2 = 3.0, p = .08$†
Heroin										
Ever tried	5	1.3	4	1.5	8	3.0	2	2.3	$\chi^2 = 2.8, p = .42$	$\chi^2 = 1.2, p = .27$†
Median age first tried	21		21.5		22.5		28		$\chi^2 = 0.03, p = .99$	$Z = 0.00, p = 1.00$‡
Used last 12 months	1	0.3	0	0	4	1.4	2	2.2	$\chi^2 = 8.2, p = .04$	$\chi^2 = 1.6, p = .20$†
Use at least monthly	0	0	0	0	1	0.03	1	1.1	$\chi^2 = 5.6, p = .13$	$\chi^2 = 1.2, p = .28$†

Note—CATI, computer-assisted telephone interview; IVR, interactive voice response. *Frequency and percentage with audit score indicating hazardous drinking. †Chi-square test, two-sided *p* values. ‡Kruskal–Wallis and Wilcoxon–Mann–Whitney tests, two-sided *p* values.

ple would be more likely to be honest with a human interviewer. The acceptability of each method is summarized in Table 6.

Although all the methods were rated as acceptable, the hybrid methods were the least enjoyable, and Hybrid II was the hardest. However, the marked decreasing trend in preference for a human interviewer with increasing automation (CATI to IVR) suggests that attitudes to the RVS may depend on exposure. This was supported by a parallel decreasing trend in the belief that people would be more likely to be honest with a human interviewer. These results indicate that use of the RVS is acceptable and becomes more so with exposure.

CONCLUSION

It appears that Hybrids I and II can provide more accurate telephone survey data on sensitive topics than can CATI, whereas Hybrid II also does so at a lesser cost. IVR is probably of greatest use in business and staff surveys.

Although IVR and the hybrid methods both used the RVS to ask questions, only the hybrid methods had acceptable response rates. This suggests that the initial human contact is important in persuading respondents to cooperate. Human contact also helps to keep the respondent on the line, as evidenced by the lower cooperation rate for Hybrid II than for Hybrid I. Once the RVS interview has begun, questions can be asked in a more consistent fashion than with an interviewer. The hybrid response rates also compared well with postal, Web, and e-mail surveys.

To keep the cost below that of CATI, the interviewer must not wait for the call to transfer back before ringing another number. Costs can be minimized by reducing the questions the interviewer asks, but without eliminating the human contact. This allows interviewers to concentrate on persuading respondents.

Future development should concentrate on the dynamics of the interview process. Specifically, the break-off rate for the Hybrid II method may be reduced by the soft-

Table 6
Acceptability of Interview Methods

	CATI			Hybrid I			Hybrid II			IVR			
Measure	Score	*n*	%	Score	*n*	%	Score	*n*	%	Score	*n*	%	Test
Ease*	1			1			2			1			$\chi^2(3, n = 468) = 26.4, p < .01$†
Enjoyable*	2			3			3			2			$\chi^2(3, n = 468) = 20.4, p < .01$†
Stressful*	2			2			2			2			$\chi^2(3, n = 468) = 5.3, p = .15$†
Likeable*	3			2			2			3			$\chi^2(3, n = 468) = 2.1, p = .56$†
Preference		81	94		127	47		26	43		13	33	$Z = -7.3, p < .01$‡
Honesty		49	57		69	26		11	18		2	5	$Z = -6.6, p < .01$‡

Note—Preference indicates preference for a human interviewer. Honesty indicates the percentage of respondents agreeing with the statement that people are more likely to be honest about their alcohol use to a person than to a recorded voice. CATI, computer-assisted telephone interview; IVR, interactive voice response. *Median Likert scale score, from 1 (*strongly agree*) to 5 (*strongly disagree*). †Kruskal–Wallis tests. ‡Cochran–Armitage test with modified ridit scores, two-side *p* values.

ware's automatically scheduling a return call in the interviewer task list if the RVS interview is terminated prematurely.

REFERENCES

Ansari, A. R., & Bradley, R. A. (1960). Rank-sum tests for dispersion. *Annals of Mathematical Statistics*, **31**, 1174-1189.

Aquilino, W. S. (1994). Interview mode effects in surveys of drug and alcohol use: A field experiment. *Public Opinion Quarterly*, **58**, 210-240.

Aquilino, W. S., & Lo Sciuto, L. A. (1990). Effects of interview mode on self-reported drug use. *Public Opinion Quarterly*, **54**, 362-395.

Australian Bureau of Statistics (1996). *Census of population and housing*. Canberra: Australian Bureau of Statistics.

Australian Bureau of Statistics (1998). *Household use of information technology*. Canberra: Australian Bureau of Statistics.

Bennett, D. J., & Steel, D. (2000). An evaluation of a large-scale CATI household survey using random digit dialing. *Australian & New Zealand Journal of Statistics*, **42**, 255-270.

Beran, R. (1986). Simulated power functions. *Annals of Statistics*, **14**, 151-173.

Bonevski, B., Sanson-Fisher, R. W., Campbell, E. M., & Ireland, M. C. (1997). Do general practice patients find computer health risk surveys acceptable? A comparison with pen-and-paper method. *Health Promotion Journal of Australia*, **7**, 100-106.

Commonwealth Department of Health and Family Services (1995). *National Drug Strategy: Household survey. Survey report 1995*. Canberra: Commonwealth Department of Health & Family Services.

Corkrey, R., & Parkinson, L. (2002). Interactive voice response: Review of studies 1989–2000. *Behavior Research Methods, Instruments, & Computers*, **34**, 342-353.

Couper, M. P., Traugott, M. W., & Lamias, M. J. (2001). Web survey design and administration. *Public Opinion Quarterly*, **65**, 230-253.

Daniel, W. W. (1978). *Biostatistics: A foundation for analysis in the health sciences*. New York: Wiley.

Desktop Marketing System (2000). *Marketing Pro* (*July 2000*). Blackburn: Author.

Duffy, J. C., & Waterton, J. J. (1984). Under-reporting of alcohol consumption in sample surveys: The effect of computer interviewing in fieldwork. *British Journal of Addiction*, **79**, 303-308.

Everitt, B. S. (1992). *The analysis of contingency tables*. London: Chapman & Hall.

Fligner, M. A. (1988). Scale tests. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 8, pp. 271-278). New York: Wiley.

Fox, R. J., Crask, M. R., & Jooghoon, K. (1988). Mail survey response rate: Meta-analysis of selected techniques for inducing response. *Public Opinion Quarterly*, **52**, 467-491.

Havice, M. (1989). How response rates compare for human and digitized phone surveys. *Journalism Quarterly*, **66**, 137-142.

Havice, M. (1990a). Measuring nonresponse and refusals to an electronic telephone survey. *Journalism Quarterly*, **67**, 521-530.

Havice, M. (1990b). Touch-tone polling in a university setting. *College & University*, **65**, 227-234.

Havice, M., & Banks, M. J. (1991). Live and automated telephone surveys: A comparison of human interviewers and an automated technique. *Journal of the Market Research Society*, **33**, 91-102.

Jobe, J. B., Pratt, W. F., Tourangeau, R., Baldwin, A. K., & Rasinski, K. A. (1997). Effects of interview mode on sensitive questions in a fertility survey. In L. E. Lyberg, P. P. Beimer, M. Collins, E. de Leewu, C. S. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 311-329). New York: Wiley.

Kiesler, S., & Sproull, L. S. (1986). Response effects in the electronic survey. *Public Opinion Quarterly*, **50**, 402-413.

Kobak, K. A., Taylor, L. v., Dottl, S. L., Greist, J. H., Jefferson, J. W., Burroughs, D., Mantle, J. M., Katzelnick, D. J., Norton, R., Henk, H. J., & Serlin, R. C. (1997). A computer-administered telephone interview to identify mental disorders. *Journal of the American Medical Association*, **278**, 905-910.

Konings, E., Bantebya, G., Caraël, M., Bagenda, D., & Mertens, T. (1995). Validating population surveys for the measurement of HIV/STD prevention indicators. *AIDS*, **9**, 375-382.

Krebs, D. (1994). Non-response to sensitive questions: Nationalism in Germany. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association* (pp. 1199-1203). Toronto: American Statistical Association.

Kulka, R. A., & Weeks, M. F. (1988). Towards the development of optimal calling protocols for telephone surveys: A conditional probabilities approach. *Journal of Official Statistics*, **4**, 319-332.

Lavrakas, P. J., Bauman, S. L., & Merkle, D. M. (1993). The last-birthday selection method and within-unit coverage problems. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association.* (pp. 1107-1112). Toronto: American Statistical Association.

Margolin, B. H. (1988). Test for trend in proportions. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 9, pp. 334-336). New York: Wiley.

Maxim, P. S. (1999). *Quantitative research methods in the social sciences*. Oxford: Oxford University Press.

McKay, R. B., Robison, E. L., & Malik, A. B. (1994, August). Touchtone data entry for household surveys: Research findings and possible applications. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association* (pp. 509-511). Toronto: American Statistical Association.

Meddis, R. (1984). *Statistics using ranks: A unified approach*. Oxford: Blackwell.

Millstein, S. G., & Irwin, C. E. (1983). Acceptability of computer-acquired sexual histories in adolescent girls. *Journal of Pediatrics*, **103**, 815-819.

Nicholls, W. L., II, & Appel, M. V. (1994, August). New CASIC technologies at the U.S. Bureau of the Census. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association* (pp. 757-762). Toronto: American Statistical Association.

O'Connell, D., Rosen, R. J., & Clayton, R. L. (1996). Long-term results of touchtone data entry in the current employment statistics survey program. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association* (pp. 764-767). Alexandria: American Statistical Association.

Oksenberg, L., Coleman, L., & Cannell, C. F. (1986). Interviewers' voices and refusal rates in telephone surveys. *Public Opinion Quarterly*, **50**, 97-111.

Phipps, P. A., & Tupek, A. R. (1991). Assessing measurement errors in a touchtone recognition survey. *Survey Methodology*, **17**, 15-26.

Piccinelli, M., Tessari, E., Bortolomasi, M., Piasere, O., Semenzin, M., Garzotto, N., & Tansella, M. (1997). Efficacy of the alcohol use disorders identification test as a screening tool for hazardous alcohol intake and related disorders in primary care: A validity study. *British Medical Journal*, **314**, 420-424.

Romuald, K. S., & Haggard, L. M. (1994). The effect of varying the respondent selection script on respondent self-selection in RDD telephone surveys. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association* (pp. 1299-1304). Toronto: American Statistical Association.

Rosen, R. J., Clayton, R. L., & Pivetz, L. L. (1994). Converting mail reporters to touchtone data entry. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association* (pp. 763-768). Toronto: American Statistical Association.

SAS Institute (1991). *SAS/SHARE software: Usage and reference* (1st ed.). Cary, NC: Author.

SAS Institute (1999). SAS (Version 8). Cary: SAS Institute Inc.

Saunders, J. B., Aasland, O. G., Babor, T. F., de la Fuente, J. R., & Grant, M. (1993). Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption. II. *Addiction*, **88**, 791-804.

Schaefer, D. R., & Dillman, D. A. (1998). Development of a standard e-mail methodology. Results of an experiment. *Public Opinion Quarterly*, **62**, 378-397.

Schumacher, R. M., Hardzinski, M. L., & Schwartz, A. L. (1995). In-

creasing the usability of interactive voice response systems: Research and guidelines for phone-based interfaces. *Human Factors*, **37**, 251-264.

SHEEHAN, K. B., & HOY, M. G. (1999). Using e-mail to survey Internet users in the United States: Methodology and assessment. *Journal of Computer Mediated Communication* [On line], 4(3). Retrieved from http://www.ascusc.org/jcmc.

SMITH, C. B. (1997). Casting the net: Surveying an Internet population. *Journal of Computer-Mediated Communication* [On-line], 3(1), Retrieved from http://www.ascusc.org/jcmc.

SNEDECOR, G. W., & COCHRAN, W. G. (1980). *Statistical methods* (7th ed.). Ames: Iowa State University Press.

STOKES, M. E., DAVIS, C. S., & KOCH, G. G. (2000). *Categorical data analysis using the SAS system* (2nd ed.). Cary, NC: SAS Institute.

TOURANGEAU, R., JOBE, J. B., PRATT, W. F., & RASINSKI, K. (1994, August). Design and results of the women's health study. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association* (pp. 49-58). Toronto: American Statistical Association.

TOURANGEAU, R., & SMITH, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, **60**, 275-304.

TURNER, C. F., KU, L., ROGERS, S. M., LINDBERG, L. D., PLECK, J. H., & SONENSTEIN, F. L. (1998). Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. *Science*, **280**, 867-873.

WERKING, G., TUPEK, A., & CLAYTON, R. (1988). CATI and touchtone self-response applications for establishment surveys. *Journal of Official Statistics*, **4**, 349-362.

WRIGHT, D. L., AQUILINO, W. S., & SUPPLE, A. J. (1998). A comparison of computer-assisted and paper-and-pencil self-administered questionnaires in a survey on smoking, alcohol, and drug use. *Public Opinion Quarterly*, **62**, 331-353.

(Manuscript received May 7, 2001;
revision accepted for publication March 18, 2002.)