INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI[®]

USABILITY OF INTERACTIVE VOICE RESPONSE SYSTEMS
IN REAL-LIFE TASKS PERFORMED
BY OLD AND YOUNG WOMEN

by Louise Dulude

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfilment of the requirements for the degree of
Master of Arts

December 2000

© copyright Louise Dulude

Carleton University

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-57703-1

Canada

The undersigned recommend to the
Faculty of Graduate Studies and Research

Acceptance of the thesis:

**USABILITY OF INTERACTIVE VOICE RESPONSE SYSTEMS
IN REAL-LIFE TASKS PERFORMED BY OLD AND YOUNG WOMEN**

Submitted by

Louise Dulude

in partial fulfilment of the requirements for the degree of

Master of Arts

Chair

External Examiner

Thesis Supervisor

Carleton University

December 2000

Abstract

Surveys about interactive voice response systems (IVRs) report high levels of dissatisfaction, especially among older users. To identify the problems IVR users experience, 22 senior women and 22 female students were monitored while they performed six IVR tasks. As predicted, old age had a negative impact on performance, but seniors were polarized into very poor and very good performers. Seniors gave lower usability ratings than young people; only young users gave high ratings to the one voice-activated system. No consistent associations were found between performance, ratings, prior experience of technology, and scores on an attention/memory test. Users' problems were mostly caused by design flaws, the most frequent being unclear choices/instructions and too-rapid automated voices. Most young users overcame all difficulties, but the majority of seniors failed because of age-related losses in their capacities. The solution is not special products for the old, but universal design that will make IVRs and other technological products easier for everyone.

Acknowledgements

Most of all, I owe sincere thanks to my thesis advisor, Dr. Warren Thorngate. When no other teacher was willing to supervise my thesis because my interests did not match their areas of specialty, he generously took me on and enlisted other faculty members who provided the required expertise. When the gaps in my unusual education led me to ask him many more questions than professors normally have to bear, he formulated and reformulated answers until I was finally able to understand. Thanks to his guidance, I have learned to ask even more questions than before.

I also want to thank Dr. Jo Wood, who supervised my studies during a qualifying year and the two first years of my M.A. program while I took time to decide what areas of psychology I wanted to pursue. Her unfailing support and wise advice made it possible for me to enter the M.A. program and to feel that I always had at least one faithful supporter at Carleton. I will always be grateful to Professor Elinor Burwell, now deceased, for introducing me to Dr. Wood.

Finally, I thank the other members of my thesis committee for their time and their useful suggestions and I thank my husband, Leonard Shifrin, for his unflagging support and encouragement.

TABLE OF CONTENTS

ACCEPTANCE FORM.....	ii
ABSTRACT.....	iii
ACKNOWLEDGEMENTS.....	iv
TABLE OF CONTENTS.....	v
LIST OF TABLES.....	vii
LIST OF FIGURES.....	viii
LIST OF APPENDICES.....	ix
INTRODUCTION.....	1
Literature Review.....	8
Attitudes toward new technologies and IVRs.....	8
Abilities required to use IVRs and age-related changes.....	14
Guidelines in designing features of IVR systems.....	21
METHOD.....	28
Participants.....	28
Procedure.....	28
Materials and Equipment.....	29
Prior Experience Questionnaire.....	30
Test of Attention and Short-Term Memory.....	30
IVR Tasks.....	31
Usability Questionnaire.....	33
Performance Measures.....	34
Number of successful tasks.....	34
Performance scores.....	34
Length of time required to succeed.....	34
Number of keystrokes/choices required to succeed.....	35
Hypotheses.....	35
Design and Data Analysis.....	36

RESULTS.....	38
Age Differences in Successes.....	38
Age Differences in Performance Scores.....	40
Age Differences in Usability Ratings.....	45
Age Differences in the Length of Time Required to Succeed.....	49
Age Differences in the Keystrokes/Choices Required to Succeed.....	51
Age Differences in Usability Ratings for Successful Tasks.....	54
Association Between Age, Attention/Memory, Technological Experience and IVR Results.....	56
Differences Between the IVR Systems in Performance Scores.....	60
Differences Between the IVR Systems in Usability Ratings.....	64
Case Studies - Features of IVR Systems and Their Effects.....	68
Income Security (System 3).....	69
Regional Municipality of Ottawa-Carleton (System 1).....	71
AccuRate (System 5).....	73
United Airlines (System 6).....	75
Statistics Canada (System 4).....	78
Canadian Airlines (System 2).....	79
Typology of Difficulties Encountered in Using IVR Systems.....	81
DISCUSSION.....	85
REFERENCES.....	94

LIST OF TABLES

1.	Medians, Means and SDs of Performance Scores by Age and System.....	42
2.	Results of Mann-Whitney U Tests and OPA of Age Differences in Performance Scores.....	42
3.	Medians, Means and SDs of Usability Ratings by Age and System.....	45
4.	Results of Pairwise Comparisons (after Bonferroni's Correction) and Mann- Whitney U Tests of Age Differences in Usability Ratings.....	48
5.	Medians, Means and SDs of Time to Succeed, by Age and System.....	50
6.	Medians, Means and SDs of Keystrokes/Choices to Succeed, by Age and System.....	52
7.	Correlations Between Time to Succeed and Keystrokes/Choices to Succeed.....	53
8.	Medians, Means and SDs of Usability Ratings for Successful Tasks, by Age and System.....	55
9.	Intercorrelations Between Age, Technological Experience, Memory Scores, Sums of Performance Scores and Sums of Usability Ratings.....	57
10.	Intercorrelations Between Age, Technological Experience, Memory Scores, Sums of Performance Scores and Sums of Usability Ratings, by Age.....	59
11.	Ranking of IVR Systems by Sums of Performance Scores, Old and Young Participants.....	62
12.	Results of Wilcoxon Tests of Differences Between the IVR Systems in Performance Scores, for Old and Young Participants.....	63
13.	Ranking of IVR Systems by Sums of Usability Ratings, Old and Young Participants.....	65
14.	Results of Pairwise Comparisons and Wilcoxon Tests of Differences Between the IVR Systems in Usability Ratings, by Age.....	67
15.	Details of Answers to Usability Questionnaires.....	70
16.	Main Problems with IVR Systems Reported by Participants, in Order of Frequency.....	82

LIST OF FIGURES

1.	Histogram of Successes by Age Group.....	39
2.	Graph of Successes by Age and System.....	40
3.	Histograms of Performance Scores by Age and System.....	41
4.	Graph of Performance Scores by Age and System.....	44
5.	Graph of Mean Usability Ratings by Age and System.....	47
6.	Scatterplots of Age with Technological Experience and Memory Scores.....	57

LIST OF APPENDICES

A.	Informed Consent Form.....	102
B.	Information/Debriefing Form.....	104
C.	Questionnaire on Prior Experience with Technological Products.....	105
D.	Word Span Test of Attention and Short-Term Memory.....	106
E.	Structure of IVR Systems.....	107
	System 1 - Regional Municipality of Ottawa-Carleton.....	107
	System 2 - Canadian Airlines.....	108
	System 3 - Government of Canada Income Security.....	109
	System 4 - Statistics Canada.....	110
	System 5 - AccuRate.....	111
	System 6 - Canadian Airlines.....	112
F.	Instructions for the IVR Tasks.....	113
G.	Usability Questionnaire.....	114

INTRODUCTION

Automated telephone attendants, known in the trade as interactive voice response systems (IVRs), have come a long way since their introduction in the 1980s (McAteer, 1997; Ollari, 1997). IVR technology was subsequently overshadowed by the Internet and other technological advances, until developments in computer-telephone integration recently brought IVRs back to the forefront by making them "portals" for Web-enabled "smart" mobile phones that give access to information from the Net "anytime, anyplace" (Lindquist, 2000; Li-Ron, 2000; Lockey, 2000).

Within a few years of their introduction, IVRs quietly became ubiquitous to the point where, by 1993, nearly 97% of large American corporations used them to greet incoming telephone calls (Communications International, 1993). Today, members of the U.S. and Canadian public are often surprised when their calls are answered by live operators because they have come to expect recorded voices reciting messages such as, "Welcome to XYZ Corp. If you know the extension number, enter it now. If you do not know the extension, you may choose from the following menu of options . . ." Callers usually answer prompts by entering their responses on a touchtone keypad; responses can also be verbal with more advanced IVRs that include speech recognition.

IVR systems are found in fields as varied as manufacturing and health care. In addition to handling the traditional operator's function of routing calls, IVRs are also used for simple tasks such as obtaining weather forecasts and hours of operation, and for more complex tasks such as renewing library books, performing banking transactions, registering for classes, enrolling in benefit programs, making airline reservations, and

listening to computerized voices that read callers' e-mail messages (Lockey, 2000; Schumacher, Hardzinski, & Schwartz, 1995; Settle, Dillon, & Alreck, 1999). The proliferation of IVR systems has been explained as follows:

The simple reason Interactive Voice Response is so prevalent is that people can easily relate to the technology. There is still no comparison to the mass acceptance, familiarity and functionality of the telephone. The television has a limited one-way communication path, and the Internet is still only accessed by about 20% of the U.S. population. But the telephone . . . refuses to relinquish its role as the greatest communications vehicle ever invented. It is difficult to find an individual who is not comfortable and accustomed to the telephone . . . (Ollari, 1997, p. 2).

The principal reason why corporations and other organizations are adopting IVRs is that they are cheaper than live operators. IVRs reduce the number of employees, never call in sick, require no fringe benefits, and handle the more routine aspects of directing calls and handing out information, leaving the remaining staff free to concentrate on "higher-value" activities and problem calls (Cook, 2000; Lennig, Bielby, & Massicotte, 1995; Whittaker & Attwater, 1996). From the point of view of customers, IVRs answer calls on the first ring, increase privacy (no operator knows who called whom), avoid embarrassment (about banking transactions, unemployment benefits, school grades, etc.), make services available 24 hours a day, 365 days a year, and handle overflow calls at peak periods, thereby reducing busy signals, call times, and times on hold (British Telecommunications, 1998; Stentiford & Popay, 1999).

Until the mid-1990s, the only articles that appeared on IVRs were in technical and trade publications in which the producers of IVR systems promoted their own products by extolling the benefits of these systems for consumers as well as organizations (Stentiford & Popay, 1999; Katz, Aspden, & Reich, 1997). When Meeks (1992) delivered a paper at a U.S. conference on voice-based technology in 1992, outlining a seven-step process that organizations should follow in selecting an IVR system (e.g., review the organization's needs, schedule demonstrations by vendors, etc.), none of the seven steps involved customer reactions.

Another reason for the lack of publications in this field is the long-standing neglect of the telephone by academic researchers (Beswick & Reinsch, 1987). A notable exception was a 1977 series of papers presented at AT&T-supported seminars at the Massachusetts Institute of Technology in celebration of the centennial of the telephone (de Sola Pool, 1977). Research done for this project confirmed that the telephone had been the least studied of the technologies, even in the 1930s when sociologists engaged in lively debates on technology and social change. This neglect was attributed in part to the fact that "The telephone is so much part of our lives that use of it is habitual rather than conscious. In the history of invention and technology, no other device can be used (in safety) with such total disregard of the thing itself" (Boettinger, 1977, p. 201).

Another explanation was that the social impact of the telephone is extremely complex, "polymorphous though indubitably large" (de Sola Pool, 1977, p. 4). For example, the phone can invade our privacy with its ring, but can also protect it by allowing us to do business from our homes. Phones allow decentralization of authority, but also

closer supervision from the centre. Phones disseminate more information, but leave no written documentation. Overall,

The phone, in short, adds to human freedom, but those who gain freedom can use it however they choose . . . the study of the telephone's social impact belongs to the important and subtle class of problems in the social sciences which demands a logic more complex than that of simple causality - a logic that allows for purposive behavior as an element in the analysis.

The theory of games and economic behavior is an example; economists do not predict a uniform consequence from a single input into the economic process; the specific consequences of the new input depend on purposive human calculations (de Sola Pool, p. 4).

One factor upon which researchers appeared to agree in 1977 was that the telephone was a great social equalizer, allowing anyone to "call powerful persons with little anxiety, even though both would be uncomfortable with a visit in person. Thus, the element of awe in power relationships is subtly and progressively undermined" (Boettinger, p. 203). With the new telephone technologies, this may no longer be the case. Although technologies such as answering machines, call waiting, and automated attendants have long been the subjects of private complaints and jokes, it was only in 1994, when Marx called for the adoption of a new code of manners in using these technologies, that public discussion started on the fact that these new systems offer increased potential for doing social harm:

New opportunities and temptations for deception and rudeness are provided by technologies that offer remote access and anonymity . . . In limiting and channelling response possibilities, the format of the new technologies may alter communication patterns by lessening reciprocal social skills and weakening the ability to express nuance and complexity . . . Emotional needs may go unfulfilled given the impersonality, standardization and automatization of the technology . . . Talking out loud to another person who is not there, which in the past might have been seen as a sign of mental illness, is gradually becoming an expected social grace . . . There is increased communications inequality such that one party may have much more powerful tools than the other. The role of the telephone as an egalitarian device permitting anyone to communicate with anyone else, regardless of social location, is changing . . . Recent innovations offer clear advantages to the sophisticated user able to afford them. But with this comes increased inequality and risks of misuse. The playing field is less level. This provides new opportunities for the abuse of power . . .

(Marx, 1994, pp. 541-2).

Marx's view that technologies like IVRs may have impoverished people's lives instead of enriching them appears to be supported by the results of surveys of public attitudes toward automated telephone systems (Katz et al., 1997; Settle et al., 1999; Society of Consumer Affairs Professionals in Business [SOCAP], 1998). The gist of these findings, which are described in greater detail below, is that while almost all American

adults are now exposed to IVRs on a regular basis and have learned to navigate them without too much difficulty, the vast majority consider these systems inferior to the traditional live-operator method and a substantial minority (about a quarter) find IVRs dehumanizing and offensive. The other main conclusion of all three surveys was that age is the factor most strongly related to attitudes toward IVRs: older respondents were more likely than younger respondents to rate their experiences with automated attendants as negative, inconvenient and frustrating.

Although these results are interesting, they raise the question of whether seniors and younger people who are asked, outside of any specific context, if they would rather do business with a real person than with an automated attendant, are imagining the same type of live-operator service. Older respondents may be thinking of the now-almost-obsolete system they experienced during most of their lives, in which live operators were expected to answer calls promptly and cheerfully. Younger respondents may be more likely to think of live operators as harassed people who finally come on the line after a frustratingly long wait listening to bad music.

General questionnaires on attitudes toward automated telephone response systems also wrongly assume that these systems are all essentially identical, though some might be better than others. Research reviewed below suggests that different features can have different effects on users, especially on older users. To cite one example, research on cognitive aging established that seniors have particular problems with tasks that place heavy demands on short-term memory; experts in computer-user interface design developed guidelines on the number of menus and menu items that should be used in IVR

systems in order to minimize the burden on short-term memory. If these guidelines are effective, IVR systems that follow them should be less frustrating for older users than systems that do not follow them, and should produce smaller differences in attitudes toward automated telephone attendants between younger and older people.

The purpose of my research was to examine old and young people while they are using IVR systems in order to identify the problems they encounter and to try to determine their cause. This was done by analysing the performance scores, usability ratings and comments of 22 old and 22 young participants who accomplished the same six real-life IVR tasks in semi-controlled conditions. All these participants were women, in part because the two populations from which the subjects for this experiment were drawn are mostly female, consisting of psychology students in their 20s for the younger group and members of the community aged 65 and over for the older group.

Another reason for choosing only female subjects is related to Marx's view, cited above, that new telephone technologies can increase inequalities for some people. Women are poorer than men at all ages and elderly women are among the poorest people in Canada. Senior women, unlike senior men, are usually widowed and live alone without a spouse who can help them with their daily activities (National Council of Welfare, 1990). As a result, older women have less help in making telephone calls and are more dependent upon social services that are reached via IVR systems. These factors make women particularly vulnerable in the light of sociological research which sees technology as a tool of elites to exploit weaker groups (Dutton & Meadow, 1987).

Literature Review

The crucial factor in this study is performance on IVR tasks. Performance is affected by attitude, by individual capacities and by the characteristics of the tasks. Relevant research on these topics includes: 1) research on the attitudes of women and men of different ages toward new technologies, and particularly toward new computer and computerized-telephone technologies; 2) research on the cognitive and motor skills required to perform IVR tasks and on the ways in which these skills are affected by age; and 3) research on the guidelines to follow in designing the best possible IVR systems.

Attitudes toward new technologies and IVRs. It is generally accepted that attitudes significantly guide behaviour (Kraus, 1995). As a result, one of the main ways in which researchers have assessed attitudes toward new technologies has been by measuring the proportion of people who use them and measuring the growth in their use. By these measures, many researchers concluded that North Americans have positive attitudes toward computers, automatic bank teller machines (ATM), and video cassette recorders (VCR), and that these technologies have been successful in gaining wide acceptance and use (Rogers, Cabrera, Walker, Gilbert, & Fisk, 1996; Schwartz, 1988; Zeithaml & Gilly, 1987). The same studies found that people over age 65 were much less likely to use these products than younger people.

On the other hand, older people who were provided explanations of the benefits of ATMs and who were introduced to them in an interactive and accessible manner were more likely to use them than older people who had not benefited from these types of introduction (Edwards & Engelhardt, 1989). Also, seniors who were reticent to accept

ATMs willingly accepted other forms of technology such as grocery checkout scanners (Czaja & Sharit, 1998). Women aged 50 to 95 who participated in a study of e-mail services found it valuable to have a computer in their home and said they would be willing to use it to communicate and pay bills (Czaja, Guerrier, Nair, & Landauer, 1993). And studies of ATMs found no gender differences in their use (Rogers, Cabrera, et al, 1996).

Related research in which younger and older subjects performed specific computer tasks under controlled conditions concludes that, contrary to popular stereotypes, being older and being female are not in themselves related to negative attitudes toward new technologies. When prior computer experience was held constant, it was repeatedly found that computer anxiety or phobia did not vary by age or by gender (Henderson, Deane, Barrelle, & Mahar, 1995; Czaja & Sharit, 1998). Other studies indicated that seniors' attitudes toward computers change with the nature of their experiences: positive experiences in using computers for games or e-mail messages over periods of at least a few days significantly improved attitudes (Danowski & Sacks, 1980; Dyck & Smither, 1994; Jay & Willis, 1992; Zandri & Charness, 1989). Positive changes in attitudes were linked to longer use and to performing meaningful tasks that were interesting and personally relevant to the participants (Morrell & Echt, 1996).

The first of the three public surveys of attitudes toward IVRs was done by sending a questionnaire to 5,000 people whose names and addresses were supplied by a commercial list service (Katz et al, 1997). The response rate was 21%, and the respondents included 44.5% who were men and 55.5% who were women. The authors admit to possible sample bias because of the low response rate and write that "we are

modest in our claims of population generalizability since we see this only as exploratory research" (p. 127). Their main findings were that although 60% of the 913 respondents reported having accomplished their goals when they had used IVR systems, only 40% had been satisfied with these services and 30% had been dissatisfied. The factor most strongly related to attitudes toward IVRs was the age of the respondents: a smaller proportion of those over age 65 (55%) than of those over age 18 to 49 (68%) had succeeded in getting through to those they wished to reach, and respondents over age 65 were more frustrated and less satisfied than those aged 18 to 49. There was also a small but significant gender difference, with women reporting fewer negative experiences than men.

On the positive side, more than half of all respondents disagreed that IVR services were difficult to use; half agreed that IVRs allowed them to call at any time and that they liked to get information like schedules and hours of operation from a recording. Less positively, 75% agreed (with 7.5% disagreeing) that IVRs make life easier for the companies that have them; 47% agreed (with 25% disagreeing) that these companies valued their employees' time more than the callers' time; 25% agreed (49% disagreeing) that IVRs make life easier for the callers; and 20% agreed (53% disagreeing) that IVRs put the customer in control. Overall, respondents were equally divided (one-third each) between those who agreed, disagreed, or were neutral on whether IVRs help companies serve their customers more efficiently. Specific problems were reported: 70% of respondents found it difficult to communicate their needs, 70% disliked listening to irrelevant options, 52% worried that nobody would get the messages they had left, 42% encountered choices leading nowhere, and 20% found the systems difficult to use.

The second survey of consumer attitudes toward IVRs, sponsored by SOCAP (1998), consisted of interviews with more than 6,000 consumers who called the consumer affairs departments of Fortune 500 companies over a period of 10 months in 1997-1998. The only publicly available information about this study was in a press release posted on SOCAP's website. The main conclusion was that consumer frustration with IVRs can have costly effects: 53% of the participants said their impression of a company's IVR system affected their decisions on whether to buy that company's products or services.

Another SOCAP conclusion was that most consumers still prefer to speak to a real person: 38% said that given the choice between an IVR menu and a live operator, they would choose the operator 9 to 10 times out of 10; 33% said they would choose the operator at least half of the time; and 29% answered 0 to 3 times out of 10. Age made an important difference: 52% of callers aged 65 and older said they preferred a live agent, compared with 14% of callers under age 18. On the positive side, 80% said they preferred IVRs to busy signals and 52% said they would be either "very or somewhat satisfied" getting an answer to a question through an IVR system (SOCAP, 1998). When asked what options they considered essential to any IVR system, 92% answered the option to speak to a live agent and 72% the ability to make a selection at any time.

When the SOCAP survey participants were asked what they liked about IVRs, 29% said they were fast, 18% said they were convenient, 17% that they got them to the right person, and 13% said 24-hour access. A "significant" number answered that they liked "nothing" about IVRs, and many had major dislikes: 22% wanted a real person, 16% thought IVRs had too many options, 13% found them impersonal, 10% were kept holding

too long, 8% did not find the options they needed, 7% had difficulty reaching an operator, and 7% found IVRs confusing (SCAP, 1998). Finally, the SOCAP study found that consumers can remember and easily use no more than three or four menu options; above four options a greater proportion of callers opted to speak to a live operator.

The third survey of public attitudes toward IVRs used student field workers to recruit 800 adult subjects in shopping malls and other public areas and give them self-administered questionnaires (Settle et al., 1999). Compared with the general population, the sample overrepresented younger people and those with higher education and higher socioeconomic status. Subjects completed a 28-adjective favourability questionnaire asking them how well they thought each adjective (e.g., modern, impersonal, high-tech, frustrating, etc.) described IVR systems. Scores ranged from 1 for "Not at all" to 5 for "Perfectly." Scores for negative items were reversed to calculate total scores ranging from 28 to 140, with a lower score representing a more positive attitude. Age was the only variable that produced significantly different scores, with average scores of 76 for respondents in their twenties and thirties, 81 for respondents in their forties and fifties, and 87 for respondents aged 60 and over.

The report of the study provides no details on the features of IVR systems which were perceived more negatively by older participants. About two-thirds of all respondents agreed that IVRs are desirable for certain kinds of organizations or businesses, that they are accessible outside office hours, and that they are fairly simple and easy to use. Forty-six percent agreed that IVRs often save time for the caller, and 26% thought they are a faster, easier way to do business than phones answered by a person. On the negative side,

69% of respondents said they would rather do business with people who answer in person, 71% complained of having to listen to a long list of things they did not want, 50% thought IVRs were too mechanical and impersonal, and 12% answered that they get totally confused by automated attendants.

To summarize, past research on attitudes has shown that older people of both sexes tend to have more negative attitudes toward new technologies than younger people, but that these differences might be due less to age than to unfamiliarity with these new products. Older women and men who are provided convincing explanations about the benefits of new technologies such as ATMs and computers, and who receive age-sensitive training in their operation at their own pace in real-life contexts that demonstrate to them that these technologies can be a useful or agreeable addition to their lives, acquire more positive attitudes toward these products. The literature does not specify whether the attitudes of older people who benefit from such age-sensitive introductions to new technologies become as positive as those of younger people.

On the other hand, it is far from certain that more familiarity with IVR systems will breed more contentment in older people who use them. For one thing, prior experience with the technology is not an issue in the case of IVRs because all adults have used telephones all their lives, and seniors more than anyone else. Surveys also report that almost all non-institutionalized seniors, like almost all younger people, have had frequent encounters with IVR systems because they are ubiquitous and cannot be avoided (Katz et al., 1997). If older people oppose IVRs more strongly than younger people, as the three surveys indicate, it could be because seniors are more likely to have judged this new

technology to be neither useful nor agreeable. The other reason why older people are more likely to dislike IVRs, as demonstrated in the Katz et al. survey, is that seniors are less able to successfully complete IVR tasks than younger people, perhaps because using IVRs, by necessity or due to poor design or both, requires skills that deteriorate with age.

Abilities required to use IVRs and age-related changes. A focus group analysis of functional limitations to daily living tasks in healthy, active seniors aged 65 to 88 found that they were experiencing particular difficulties with many new technologies, including VCRs, telephone menus, multiple-line telephones, home security systems, credit card scanners, answering machines, entertainment centers, copy machines, cameras, microwave ovens, and fax machines. (Rogers, Meyer, Walker, & Fisk, 1998). In discussions about ways of coping with these problems, it emerged that telephone systems with menus were more troublesome to seniors than most other technologies because they could not be ignored since they were imposed by offices and stores.

These active seniors reported different types of difficulties with new technologies in general and with IVR systems in particular. The main type involved cognitive limitations: the seniors found it harder to learn new tasks than when they were young, and had memory problems that made it difficult for them to negotiate telephone menus. A small proportion reported problems using telephones, for unspecified reasons which may have been hearing limitations, visual limitations (difficulty in seeing the phone dial or keypad clearly), or motor limitations (arthritis) in the use of their hands making it difficult to enter digits on telephone keypads (Charness & Bosman, 1994). When the seniors were asked to make suggestions in order to reduce their problems with technological systems

such as IVRs, they proposed a combination of redesign of the products and of training in the tasks involved (Rogers, Meyer, et al., 1998).

Many studies have established that hearing impairments increase with age, especially after the age of 50 (Kline & Scialfa, 1997). Hearing in older people is characterized by presbycusis, which is a decreased capacity to perceive high-frequency sounds. Presbycusis can affect the ability to detect and identify speech, and it is usually more severe in men than in women. According to Feldman and Reger (1967), by age 80 people can miss as much as 25% of the words in a conversation. The problem is worse with female voices than male ones because of their higher pitch. Older people also process sequences of high-frequency sounds more slowly than younger people, making it more difficult to recall long sentences. A study of volunteer subjects found that the proportion of people who had some hearing disorders was 17% under the age of 40, 20% for ages 40 to 59, and 54% at ages 60-79 (Slawinski, Hartel, & Kline, 1993).

Other studies examined older people's greater difficulty in perceiving speech under less than ideal conditions. Seniors hear speech less well than younger people when there are distracting noises (Nabelek, Letowski, & Tucker, 1989). They also find it harder than younger people to understand distorted speech such as time-compressed (speeded-up) speech and synthetic/reconstituted speech, both of which are used in some IVR systems (Fitzgibbons & Gordon-Salant, 1999; Smither, 1993). Synthetic speech is used when the number of possible responses is too large to record them all. The most common example is telephone directory systems; numbers are recorded separately and are assembled by a computer, producing telephone numbers recited in robotic-sounding voices.

As well as having more hearing problems, older listeners are worse than younger ones at maintaining auditory attention, which is the mental ability to concentrate on what we hear in spite of distractions (Barr & Giambra, 1990; Kline & Scialfa, 1997). Greater distractibility can make an important difference in using IVR systems because "Listening requires *attention*" (Schmandt, 1994, p. 103). This makes IVR systems more demanding of attention than traditional computers with video monitors because, "Although the user can perform other tasks while executing a long text-based keyword search, this is not possible while listening" (Schmandt, p. 103). Older adults find it more difficult than young people to maintain concentration over long periods of time and tire particularly quickly when performing tasks that require continuous, uninterrupted attention, such as is required to successfully negotiate most IVR systems (Vercruyssen, 1997).

Speech becomes less clear with age, perhaps because of reduced control of tongue and lips, or due to impaired ability to hear and correct oneself, or because some elderly people have speech impairments caused by strokes (Hawthorn, 2000). Older people's speech is also slower, more hesitant and more likely to contain interruptions, filled pauses ("um" and "er") and repetitions (Balota & Duchek, 1988). IVR systems that use speech recognition are typically programmed to capture the speech patterns of younger people and may have difficulty recognizing older voices. Elderly people are also more likely than young ones to suffer from respiratory infections that make them breathe audibly, which can trigger unwanted automatic speech recognition responses.

Good vision is less essential than hearing in using automated telephone systems, but it must generally meet at least minimal standards in order to enable callers to punch in

the appropriate responses on a normal touchtone phone in response to automated prompts about menu choices. Population surveys indicate that vision problems and blindness are mostly found among older people. In the United States, about 46% of those who are legally blind and approximately 68% of those with severe visual impairments are aged 65 and over. The proportion of people with serious visual impairments was 9% for those aged 55 to 64, 18% for ages 65 to 74, and 35% for ages 75 to 79 (Kline & Scialfa, 1997).

Movement control and speed of movements, which are required to punch the appropriate keys on the telephone keypad in using IVRs, also diminish with age. Although the participants in the focus group analysis by Rogers, Meyer, et al. (1998) were healthy and active, many nevertheless made remarks such as "I am slowing down," or "Everything I do is just about a struggle and it takes longer" (p. 117). Thirty-eight percent reported motor limitations, mostly related to gross movements involving the knees or the back. Fine movement problems, which can make telephone handling difficult, were much less common and mostly due to arthritis.

Slower physical responses, meaning the time required to react and respond, are among the most characteristic signs of aging (Vercruyssen, 1997). Several hypotheses have been formulated to explain slower movement in healthy older adults. One is that there is more "noise" in the visual processing system of older than of younger adults, which leads to poorer movement performance by impeding the visually-guided feedback mechanism involved in most movements (Walker, Philbin, & Spruell, 1996). Another hypothesis concerns "neural noise" in the motor system that increases with age, leading older adults to produce slower movements "because they must move more slowly in

order to maintain the same level of movement accuracy as younger adults" (Walker et al., p. 47). A third hypothesis is that older adults plan slower movement because they are more "error averse" or cautious than younger adults. Greater cautiousness may be the reason why older people slow down more following an error than do young people. It may also explain why the speed gap between old and young subjects becomes wider when there are more choices (Vercruyssen, 1997).

A final hypothesis put forward to explain older people's slower physical responses is that all underlying processes (cognitive, perceptual, and motor) become less efficient or slower with age, resulting in slower performance (Salthouse, 1993, 1994). As this explanation suggests, cognitive abilities also become slower with age, with age differences in speed and accuracy on intelligence and information-processing tests increasing with the speed and complexity of the tasks (Kausler, 1994; Myerson & Hale, 1993; Verhaeghen, Vandenbroucke, & Dierckx, 1998). The fact that older minds are slower than younger ones may explain the differences we saw earlier between younger and older people in accomplishing their goals when they called automated telephone response systems (Katz et al., 1997), because many of these systems require quick choices between ambiguous items often described in bureaucratic jargon.

Although the finding that seniors are slower at information processing is among the least disputed and most reliable in the psychology-of-aging literature, it is also controversial because experts disagree on what it implies (Lemme, 1995). According to some, the fact that older adults are slower does not mean that they are less competent, but that it simply takes them longer to learn and to perform the same tasks. This seems to be

confirmed by many studies demonstrating that older adults perform much better in situations in which extended or self-paced schedules are used (Czaja, 1996). On the other hand, this appears to be contradicted by findings that some age differences remain even when time limits are removed (Salthouse, 1993, 1994).

Along with speed of information processing, the cognitive process most often associated with inferior performance by older people on computer tasks is working memory (Baddeley, 1994; Czaja, 1996; Morrell & Echt, 1996). Research on this subject found that age differences are small for short-term memory tasks requiring that people remember modest amounts of information for short periods, but that they are larger when people must simultaneously process or manipulate the information in their minds. A classic example of working memory consists of calculating a 15% tip on a restaurant cheque. The usual way to do this is to calculate 10%, keeping that amount in mind while dividing it in half to obtain 5%, then mentally adding together the 10% and the 5% to obtain the total tip amount (Howard & Howard, 1997). Many studies found a substantial decrease in working memory capacity with advancing age. The more complex the tasks, the greater the differences between younger and older people (Craig & Jennings, 1992; Gick, Craig, & Morris, 1988; Morris, Gick, & Craig, 1988; Salthouse, Mitchell, Skovronek, & Babcock, 1989; Wingfield, Stine, Lahar, & Aberdeen, 1988;).

The link between these findings on working memory and the use of IVR systems is easily made. For example, a person who telephones an organization to obtain information about product or program X may have to listen to and remember five or more menu choices and their corresponding menu numbers while she or he simultaneously

assesses each newly offered choice to decide whether it sounds more likely than the others to include information about product or program X. Then, the person would choose the most promising option and press its corresponding number on the keypad, perhaps at the same time committing to memory a number to press to return to the first menu in case it turns out that the choice that was made was a wrong one. The next step would likely involve listening to a second-level menu containing several items that might have nothing to do with product or program X, in which case the person might press the appropriate key to return to the first menu, etc.

Automated telephone response systems impose much greater burdens on memory than ordinary computers with video monitors:

A stream of voice conveys only one word at a time; we must wait for the first word to be spoken before we can hear the second. If a menu is displayed on a [video] terminal, all the items appear very quickly and the user then visually scans them until a choice is made. With speech each item must be recited in turn; once spoken it is gone unless the user remembers it. If the user misses a portion of the message, it must be repeated in its entirety . . . this serial nature of speech places additional memory requirements on the user (Schmandt, 1994, p. 102).

In light of all these results on the negative effects of advancing age on physical and mental capacities, various attempts were made to qualify them by attributing age differences to factors other than aging, such as education, health, physical fitness, or motivation. The results are still inconclusive, with some researchers denying that any of

these factors make much difference (Smith, 1996), and others insisting that they do (Bunce, Morris, & Barrowclough, 1996). Another relevant line of argument concerns debates between researchers such as Salthouse (1993, 1994), whose findings are based on cross-sectional studies and who sees almost only declines with age, and researchers such as Schaie (1994, 1996), who are dedicated to longitudinal studies and who suggest that major decline in intellectual abilities mostly occurs in late old age and mostly affects abilities which were not central to the person's life experience. Schaie also points out the huge individual differences in age-related changes and the success of interventions such as training in reducing cognitive declines in aging populations.

To summarize research on the skills required to use IVR systems, it appears that for healthy community-dwelling people, the most essential skills are cognitive and concern: 1) attention or concentration; 2) speed of information-processing; and 3) memory, including short-term memory and working memory for tasks that include simultaneous memory and processing elements. All these cognitive skills show significant decrements with age, with differences in performance between younger and older people increasing with the speed and complexity of the tasks. Vision, hearing and motor control play a minor role in the use of IVR by healthy people in the community, although they can be major barriers for people who are physically disabled.

Guidelines in designing features of IVR systems. The features of computerized systems that allow them to communicate with the people who use them are called "computer-user interfaces" (Kloosterman, 1994). In the case of interactive voice response systems, the interface is a series of exchanges in which the system provides audio

instructions to the caller, who directs the system by pressing a key or sequence of keys on the telephone keypad or by making a verbal response (Blanchard & Lewis, 1999; McTear, 2000). Interfaces can be good or bad. A good automated telephone interface "makes the users feel comfortable and in control, guides them to give predictable responses, and ensures they know how to get help or exit at any time. It must be clear, simple and consistent to accommodate all users, regardless of their experience or degree of computer literacy" (Stentiford & Popay, 1999, p. 142).

Because of the dramatic increase in the use of IVRs in the past 15 years, many of these systems were developed in haste with little consideration of the callers' needs or satisfaction. This was partially blamed on "the proliferation of Computer Telephony (CT) application toolkits" which "do an excellent job of enabling the rapid development of CT applications" but "allow IVR applications to be quickly and easily created that break virtually every IVR rule in the book" (Bajorek, 2000). Callers are particularly outraged when they find themselves in "IVR hell," where "no amount of button-pushing can get you out of some information loop" (Cross, 1995). A legendary comment on IVRs attributed to the London Daily Mail invites us to "Press one for confusion, two for frustration and any other number to go completely mad" (Turner, 2000).

Perhaps because this type of publicity is harmful for the whole communications industry, major producers of IVR systems such as British Telecommunications took the lead in trying to convince IVR producers and their corporate clients that better IVR interface design is not a luxury, but an essential element of IVR production whose neglect is bad for business in both the short and the long term:

Experience has shown that there are major benefits to be derived from good human interface design and that this can have a significant impact on the commercial success of automated telecommunications. Good design . . . reduces overall development costs and the time to roll-out by eliminating problems early in the design cycle. Risks of customer dissatisfaction are also reduced, as well as the likelihood of callers moving to competing products. Streams of revenue can only be sustained if end users are able to achieve their goals and return again and again to use the service (Stentiford & Popay, 1999, p. 142).

Another reason for bad IVR design is that although a great deal of research had been done on these systems since they were introduced in the 1980s, almost none of it was made public until the last few years because most of this research was done by telecommunications companies that had no desire to share their findings with their competitors. As the need for standardization of the different IVR systems and improved customer service became recognized, industry guidelines were gradually developed on many specific features of IVRs. These guidelines remain extremely flexible, however, because it is generally accepted that all rules must be subordinated to the most basic objective of all, which is to tailor each IVR system to the particular requirements of its owners and customers (Gardner-Bonneau, 1999).

The most prominent guidelines concern menus. It was recognized that many IVRs failed in the 1990s because: 1) callers kept getting lost in systems whose multi-level menus were too deep; or 2) callers became exasperated by having to waste time listening to

numerous items in systems whose menus were too broad; or 3) callers could not handle the memory burdens imposed by menu items that were too long (Gardner-Bonneau, 1999). Many experts agreed that, as a general rule, menus should have four or fewer items; longer lists should be split up, for example by presenting the first four options, followed by an option (number 5) to hear more options (Schumacher et al., 1995). Qualifiers were added to recommend a maximum of four unrelated items or six related ones (British Telecommunications, 1998), and to allow as many as nine items when they are objects such as movies and names.

Consensus was not achieved on these rules, and some experts still maintain that splitting a top-level list of more than four items into two menus is unnatural and makes the system harder to learn and use (Virzi & Huitema, 1997). It is generally agreed that menus should be constructed with items in descending order of popularity or other logical grouping, and that the items should be numbered in ascending order starting with 1 (British Telecommunications, 1998; Gardner-Bonneau, 1992). An overall limit on the menu is recommended, such as having no more than three levels in all with no more than four choices per level (Schumacher et al., 1995).

Other guidelines were formulated about the recorded "attendant" voices. Everyone agrees that automated attendant voices should be agreeable, assertive, courteous, friendly and helpful, and that both the voices and the recordings should be of very good professional quality because poor-quality speech, whether synthetic or recorded, quickly becomes annoying. Because natural voices are preferred by callers and best understood, synthetic or reconstituted (also called concatenated or recombinant)

speech should not be used unless there is no option. A single speaker should be used rather than a combination of two or more, especially with reconstituted speech. The choice of using male or female voices, or light-hearted or serious voices, should depend on the nature of the service provided. Generally, female contralto voices are preferred because they are perceived as more helpful and less authoritarian than male voices (Noonan, 1998). Background tones or music should not be used during prompts because they can hamper comprehension (Schmandt, 1994; Schumacher et al., 1995).

IVR dialogue and scripts are often accused of being irritating and ill-thought-out. This was attributed to the fact that IVR designers mostly concentrate on the technical aspects of these systems, while the scripting is often just "left to the customer," that is to the company or institution that is buying the system (Christell, 2000). Experts recommend that scripts include the following: 1) a carefully-worded, welcoming and informative introductory message informing callers that they are connected to a machine, and briefly explaining the facilities offered by the service; 2) signposts and feedback that inform callers about where they are in the system (for example, by incorporating the name of the item just chosen in the title of the subsequent message, or by confirming, "You entered 10 dollars"); 3) comfort messages to explain problems such as delays ("Please hold while we process your request"); 4) recorded information the caller has phoned there to obtain, such as weather forecasts, schedules, hours of operation, bank balance. Because speech disappears so quickly, repetition of all messages and prompts must be possible to reduce the burden on attention and memory (Cross, Hayhow, Postle, & Bloomfield, 1994).

The language used in all messages and prompts should be simple, clear, unambiguous, concise, and free of technical jargon unless the users are expected to be familiar with it (Schumacher et al., 1995). Because speech is slow, voice messages and prompts should almost always be interruptible, meaning that if a passage is irrelevant or redundant, callers should be able to skip over it. This includes the capacity to "dial through" ("talk-through" for voice recognition systems), that is to press the number of the item they want (or say the name of the item they want) as soon as they hear it without having to listen to the rest of the choices. Users should also be able to "dial-ahead" (or "talk-ahead") by entering the option number (or naming the option) even before the system explicitly starts to recite a menu. These navigational features put experienced users in charge and allow them to move very quickly through the system. Studies have found that some callers are unaware that they can respond before their option is mentioned; they were frustrated and felt that the system was very slow. Introductory messages to the first menu should always specify to "Make your selection at any time" (Schumacher et al., 1995; British Telecommunications, 1998).

According to the guidelines, keypad keys should have standardized names and meanings; for example, 0 should be called "zero," never "oh," and should be usable at any time to stop the current action and take the user to human or automated help within the system (Schumacher et al., 1995). Equivalent guidelines for IVRs with speech recognition include universal commands such as "Help" to obtain context-sensitive assistance, "repeat" to hear the prompt just played, and "transfer to agent" (Cohen, 1999). The system should avoid as much as possible asking callers to input information (such as credit card number)

by touchtone and should never ask redundant or annoying questions, such as requiring customers to identify themselves more than once in the same call (Gartner Group, 1998; Stentiford & Popay, 1999).

Error-handling messages should be graceful, polite, and informative, and should not blame the caller. Messages like "You gave the wrong answer" or "You made an invalid response" can upset callers and increase errors. Instead, the system should use soothing messages like "Sorry, I didn't understand that" or "Sorry, there appears to be a problem" (British Telecommunications, 1998) and should tell callers what the valid options are. Callers who are stuck or lost - who often remain silent while trying to figure out what to do - should be given the means to recover, by hearing a repetition of the menu or prompt, or by being clearly told what they are expected to do next. Studies found that people who are given no help in such circumstances often say loud words (perhaps swear words?) and hang up (Stentiford & Popay, 1999). Exit routes should always be provided:

Callers should always know where they are, what they can do, and how to get help. Context-sensitive help should be available at each interactive stage, and, if possible, callers experiencing problems should be connected with a human operator or given a helpline number (Stentiford & Popay, p.147).

By identifying the most common problems old and young people encounter when they use IVR systems, this study could inspire improvements that would make these systems less unpleasant and easier to use, thus improving the quality of life of large numbers of people.

METHOD

Participants

As explained in the introduction, all the subjects in this study were women. They included 22 young women (8 aged 18 or 19, 12 aged 20 to 24, 2 aged 25 to 27) and 22 women aged 65 or over (8 aged 65 to 69, 4 aged 70 to 74, 6 aged 75 to 79, 4 aged 80 to 86). The younger subjects were psychology students at Carleton University who were recompensed for their participation by receiving course credits. The older subjects were members of the Good Companions Seniors' Centre in Ottawa who participated on a purely volunteer basis; many of these seniors said they had agreed to participate in the hope that this study would bring about improvements to IVR systems. All participants had at least a high-school education. They were screened for basic telephone competency by being asked to dial 745-1576 - the National Research Council line that provides the time of day - and to report on the subject they heard about on the line. Everyone passed this test.

Procedure

I personally carried out all the experiments in January and February 2000. Young participants were recruited on the Carleton University campus and were tested individually in a psychology lab. Older participants were recruited by setting up an information table at lunch times at the entrance to the cafeteria of the Good Companions Seniors' Centre; those who volunteered were tested individually in an adjoining room provided by the administration. Except for one senior participant who became tired and finished the next day, testing was done in single sessions lasting from an hour to an hour and a half.

Once participants had given their written informed consents (see Informed Consent Form in Appendix A) and provided their birth dates (and their permanent addresses for those who wanted to be informed of the study's results, which was everyone), the following procedures were administered: the basic telephone competency task described above; a questionnaire on prior experience with technological products; a word-span test of attention and short-term memory; and six consecutive IVR tasks, each of which was immediately followed by a questionnaire on the usability of the IVR system just used. To end, participants were given an information/debriefing form (see Appendix B).

Materials and Equipment

The equipment consisted of an ordinary touchtone telephone, a Sony tape recorder (cassette-corder model TC-67) with earphones, a Radio Shack telephone recording device (cost around \$30), one 6" by 4" index card displaying instructions for the basic telephone competency task, six 8" by 5" index cards with instructions for each of the six IVR tasks, a Radio Shack stopwatch (Easy-Set Count Up/Down Timer, around \$20), and blank sheets of paper and pencils. Participants used the telephone to make the basic competency call and the six IVR calls; the telephone recording device and tape recorder were used together to record the IVR tasks with the knowledge and consent of the participants; the earphones allowed me to monitor the calls and take notes on them; the tape recorder was also used to present the recorded word sets for the attention/memory test. The stopwatch was used after the tests to establish, by listening to the tape recordings, the length of time participants took to accomplish each IVR task.

Prior Experience Questionnaire

The questionnaire about participants' prior experience with IVR systems and other technological products (microwave oven, ATM, debit card scanner, home telephone answering machine, VCR, video camera, photocopy machine, fax machine, compact disk player, computer, home security system) is shown in Appendix C. For each product, participants were asked whether they had used it in the past month (with Yes producing a score of 1 and No a score of 0), how many times they had used it in the past month (the score being the number of times used, up to a maximum of 5) and, if they had ever used the product, whether the experience had been unsuccessful (score=0), successful (score=2) or mixed (score=1). Total possible scores ranged from 0 for no experience of any of the 12 products to 96 for maximum successful use of all 12 products.

Test of Attention and Short-Term Memory

This test was based on the simple word-span task developed by Lehto (1996). The words used, which were all 5-to-7-letter words of two syllables, are shown in Appendix D. Each word was used only once. The words were presented from a tape recording (using an unaccented male voice) at the rate of approximately one word per second. They were presented in sets, the first two sets consisting of three words, the next two sets of four words, the next sets of five words, and so on. The largest set size was nine words, for a total of 84 words. After each set was played, the tape was stopped and participants took the time they wanted to note down the words they recalled, in any order. Testing was discontinued after a participant failed to correctly recall two sets of the same length. The score was the total number of correctly recalled words (maximum 84).

This test of attention and short-term memory was retained after a test of working memory was initially chosen and then discarded on the basis of pilot tests with five people aged between 23 and 60. The initial working-memory test was based on the procedure developed by Daneman and Carpenter (1980) and required people to read aloud blocks of unrelated short sentences, after which they had to recall the final words of all the sentences in the block. The first blocks included two sentences, which were followed by blocks of three sentences, then four, and so on up to and including six sentences. In the pilot tests, two participants who did poorly at this task complained that it was too difficult and made them feel humiliated. The easier word-span test was substituted to avoid this problem and also because (a) the IVR tasks used in the study made only minimal demands on working memory, and (b) the word-span test, like the IVR tasks, was an auditory test requiring participants to recall words spoken aloud by someone they could not see.

IVR Tasks

Participants were asked to place six calls to different IVR systems to obtain specific pieces of information. They were instructed to use only automated menu options and to avoid live operators. The six IVR systems were those of the Regional Municipality of Ottawa-Carleton (called System 1), Canadian Airlines (System 2), Government of Canada Income Security (System 3), Statistics Canada (System 4), Accu-Rate Foreign Exchange (System 5), and Canadian Airlines (System 6). These systems are described in Appendix E with charts of the options offered during the period when testing took place. The United Airlines system differed from the others by featuring voice recognition, so that callers generally responded to prompts verbally instead of with keystrokes. The

instructions to participants for the six tasks are shown in Appendix F, and the tasks are described in detail in the Case Studies section below. No time limit was imposed to complete a task, but re-dialing was not permitted except when a line was busy or a wrong initial number was reached, so that recoveries from error had to take place within the same call. A test ended if a live operator came on the line. To avoid practice effects, counterbalancing of the tasks was approximated by means of a Latin square arrangement. Full counterbalancing, which would have required a number of subjects that was a multiple of six (such as 24 old and 24 young participants), was not possible due to difficulties with obtaining a larger number of subjects.

The six IVR systems were chosen out of an initial selection of 24 systems that were identified by questioning approximately 15 Ottawa people in their fifties and sixties about their own, their friends' and their parents' experiences with IVRs. This was supplemented by consulting telephone operators and directories. The criteria for retaining the six IVR systems were established during pilot testing of all 24 systems (with the same five pilot test participants who were mentioned above). The first criterion eliminated systems that did not lend themselves to performing fictitious fully automated tasks; for example, the IVRs of retail stores were eliminated because the end point of calls to these systems is a conversation with a live clerk. Systems were also eliminated if their use could embarrass or worry sensitive participant-callers or oblige them to divulge personal information. This excluded Bell Canada, which requires callers to enter their telephone numbers, as well as banking and other financial IVR systems that ask callers to enter their account numbers or the numbers of their credit cards. Other IVRs were eliminated

because they were clearly too difficult for ordinary people, such as the Revenue Canada Automated Customs Information Service, which used impenetrable jargon. Finally, to assess the impact of different types of IVR features, the six IVRs and tasks that were retained were the ones that seemed to offer the widest possible range of features and of degrees of difficulty to participants.

The pre-tests suggested that the Income Security task was the easiest to perform successfully while the Canadian Airlines task was the most difficult; the other four tasks seemed to present intermediate degrees of difficulty. As far as usability was concerned, pre-test participants rated the Income Security and United Airlines systems/tasks most positively and the Canadian Airlines system/task most negatively, with the other three tasks receiving intermediate ratings.

Usability Questionnaire

As soon as participants finished an IVR task, they were asked 13 questions to find out their impressions of the system they had just used (see questionnaire in Appendix G). The same questions were asked for all systems/tasks. The first 10 questions were based on a 22-item Likert-scale IVR usability questionnaire (1=strongly disagree to 5=strongly agree) developed by Dutton, Foster and Jack (1999). They included five positive items ("I found the service easy to use") and five negative items ("I felt very frustrated using this service"). A participant's total usability rating for an IVR system was the sum of the 10 numbers corresponding to their answers (such as 1 for "strongly disagree" and 5 for "strongly agree"), with the numbers being reversed for the negative items. The possible range was 10 to 50 points, with a higher rating reflecting a more favourable impression.

Performance Measures

The four measures of performance were the number of successful tasks, the performance scores, the length of time required to succeed, and the number of keystrokes/choices required to succeed.

Number of successful tasks. A successful task was one in which participants made the right choices and obtained the correct answer to the question posed. For example, the IVR task involving the telephone system of the Regional Municipality of Ottawa-Carleton asked participants to find the answer to the question "Is the Regional Municipality hiring any new employees at this time?" To succeed, participants had to make four consecutive correct menu choices, expressed in four keystrokes (2, 5, 8 and 3), and to note down the answer to the question. As each participant performed six tasks, the number of successful tasks could range from 0 to 6.

Performance scores. Unlike the dichotomous definition of a successful task (participants succeeded or did not succeed), performance scores took into account partial successes, for example by giving 50% of the top score to participants who gave up after entering two of the four correct keystrokes required to succeed in the Regional Municipality task described above. Performance scores could range from 0 (total failure) to 1 (success); 50% of the top score was 0.5.

Length of time required to succeed. This was measured in terms of the number of seconds a participant needed to successfully complete a task. The assumption behind this measure was that participants who took longer to succeed expended more effort to achieve the same result than participants who took less long to succeed.

Number of keystrokes/choices required to succeed. This was the number of menu choices successful participants made on their way to the correct answer. For five of the six IVRs, participants' choices were indicated by keystrokes. For the voice-activated United Airlines system, choices were almost always expressed through verbal commands, with keystrokes sometimes used as fallbacks when the voice system failed. In the case of United Airlines, the number of choices to succeed is the sum of the voiced commands and the keystrokes. The assumption underlying this measure was that if two participants succeeded at the same task but one of them used more keystrokes/choices to achieve this, the one who needed more choices had more difficulty in performing the task because she made some wrong choices that took her out of the shortest, most direct path.

Hypotheses

Based on research about the effects of aging, I hypothesized that my tests would find significant age differences in using IVR systems. Specifically, I predicted that (a) older women would perform less well than young ones on all tasks; (b) older women's usability ratings would be lower than those of young women for all systems; (c) compared to young women who succeed, older women would take longer to succeed, would need more keystrokes/choices to do it, and would be less satisfied with the IVR systems they used successfully; (d) older women's scores on prior experience of technology and on the attention/memory test would be lower than those of younger women, and these lower scores would be correlated with inferior performances and lower usability ratings.

On the basis of the pre-tests and of usability research, I hypothesized that the different IVR systems would produce different performance scores and usability ratings.

Specifically, I expected that (a) the tasks that require users to make the largest number of choices/decisions to succeed would produce the lowest performance scores, while the tasks that require the smallest number of choices would produce the highest performance scores; and (b) the system requiring the smallest number of choices (Income Security) and the voice-activated system (United Airlines) would produce the highest usability ratings, while the system requiring the largest number of choices (Canadian Airlines) would produce the lowest usability ratings. As the effects of aging vary depending on the tasks involved, I also expected to find a significant interaction between age and the type of IVR system. Because so little is known about the impact of specific IVR features on their users or about the differential effects of IVR features on old and young users, my research on these subjects was exploratory.

Design and Data Analysis

This study has a 2 x 6 mixed factorial design. The between-subjects independent variable is age (old vs. young) and the within-subjects independent variable is the type of IVR system (systems 1 to 6). Two correlational variables, scores on prior technological experience (0 to 96) and attention/memory scores (0 to 84), are used in some of the analyses. The dependent variables include the number of successful tasks (0 to 6), performance scores (0 to 1), usability ratings (10 to 50), time to succeed (in seconds), and number of keystrokes/choices to succeed.

A first set of analyses assesses age differences in the dependent variables. For data that are not normally distributed, the statistical methods used are non-parametric Mann-Whitney U rank tests for two independent samples and Ordinal Pattern Analysis (OPA).

Data that are normally distributed are analysed by means of t-tests where appropriate and with 2 x 6 repeated-measures mixed-design factorial univariate analyses of variance (ANOVA) with follow-up pairwise comparisons. The t-tests and pairwise comparisons are performed using Bonferroni's correction ($\alpha .05/6=.008$).

A second set of analyses examines differences between the IVR systems in performance scores and usability ratings. These analyses are done separately for each of the two age groups. For data that are not normally distributed, non-parametric Friedman tests for K related samples are used, with Wilcoxon follow-up tests for two related samples. Data that are normally distributed are analysed in pairwise tests following up on the 2 x 6 repeated-measures mixed-design factorial ANOVA mentioned above, using Bonferroni's correction ($\alpha .05/12=.004$).

Other analyses are performed to examine the relationships between performance scores and usability ratings; between time to succeed and keystrokes/choices to succeed; and between age, prior technological experience, attention/memory scores, performance scores and usability ratings. This is done with the non-parametric Kendall's tau measure of association. Finally, I report case studies that focus on each IVR system to provide details of the tasks performed, to analyse participants' comments, and to examine tabulations based on their answers to each of the ten main questions in the usability questionnaire.

RESULTS

The analyses presented in this section address the following questions: Are young women more likely than old ones to succeed at IVR tasks? Do women from both age groups evaluate their IVR experiences differently? Are successful old participants different from successful young participants on usability measures and on performance measures such as the length of time or the number of keystrokes they required to succeed? Can the inferior performance of old or young women be attributed to their lesser familiarity with new technologies, to age-related losses in short-term memory, or to both? Are some of the six IVR systems easier or more agreeable to use than others, and if so, what features make systems easier or harder, more pleasant or unpleasant, for either or both old and young people?

Age Differences in Successes

All participants successfully completed at least one of the six IVR tasks they performed, so their success scores ranged from 1 to 6. Figure 1 shows the distributions of the numbers of successful tasks for the 22 old and the 22 young participants. The distribution for old participants is bimodal and ranges from one to six successes, while the distribution for young participants is unimodal and ranges only from three to six. For old people, the median number of successes was 2.5 and the mean 2.8 (SD=1.8); for young people, the median was 5 and the mean 5.1 (SD=0.8). Eighty-two percent of the young participants were successful on at least five of the six tasks, while only 32% of the old participants were this successful.

Figure 1. Histogram of Successes by Age Group

Because of the non-normal distribution of the scores for old participants, it is desirable to use a non-parametric method of analysis. To determine if young participants succeeded at more tasks than old participants, a Mann-Whitney rank test for two independent samples was performed on the number of successes achieved by young and old participants. The results indicate a significant difference, $U = 77.5$, $p < .001$, which suggests that young people are more likely to succeed at IVR tasks than old people.

Is this age difference consistent across IVR systems? Figure 2 shows the number of old and young participants who successfully completed IVR tasks on Systems 1 to 6. As seen in Figure 2, young people are more likely to succeed than old people on all systems except System 3 (Income Security). For System 3, all but one participant from each age group succeeded at the task.

Figure 2. Graph of Successes by Age and System

Age Differences in Performance Scores

The previous analysis of successes can be refined by comparing old and young participants on their performance scores, which reflect partial as well as full successes: participants who made half of the correct choices for a task received performance scores of 0.5. Performance scores for both age groups on all six systems are shown in Figure 3. Almost none have normal or near-normal distributions. Four of the distributions for old participants are bimodal (Systems 1, 4, 5, 6). Five of the distributions for young people (all but System 2) and one distribution for old people (System 3) are positively skewed. According to the medians, means and standard deviations of the performance scores, shown in Table 1, young people obtained higher scores on all systems except System 3 (Income Security), on which both groups obtained the same score. The average scores for young people were very near the maximum score for five of the six systems, and the scores for old people had substantially greater deviations on four of the six systems.

Figure 3. Histograms of Performance Scores by Age and System

Table 1**Medians, Means and SDs of Performance Scores by Age and System**

	Old participants		Young participants	
	Median	<u>M (SD)</u>	Median	<u>M (SD)</u>
System 1 (Regional Municipality)	1.0	0.7 (0.4)	1.0	0.9 (0.2)
System 2 (Canadian Airlines)	0.4	0.4 (0.3)	1.0	0.7 (0.3)
System 3 (Income Security)	1.0	0.9 (0.1)	1.0	0.9 (0.1)
System 4 (Statistics Canada)	0.8	0.7 (0.4)	1.0	0.9 (0.1)
System 5 (AccuRate)	0.4	0.5 (0.5)	1.0	0.9 (0.2)
System 6 (United Airlines)	0.6	0.7 (0.3)	1.0	0.9 (0.0)

Table 2**Results of Mann-Whitney Tests and OPA of Age Differences in Performance Scores**

	M-W (<u>U</u>)	OPA (IOF)
System 1 (Regional Municipality)	143.5**	0.9**
System 2 (Canadian Airlines)	127.5**	0.5**
System 3 (Income Security)	242.0	0.0
System 4 (Statistics Canada)	118.0**	0.8**
System 5 (AccuRate)	122.5**	0.9**
System 6 (United Airlines)	92.0***	0.9***

** $p < .01$ *** $p < .001$

Because of the non-normal nature of the distributions, non-parametric methods of analyses are preferable. To determine if young people have significantly higher performance scores in IVR tasks, six Mann-Whitney rank tests for two independent samples were performed on the performance scores of old and young participants on each of the IVR systems. The results, shown in Table 2 on the previous page, indicate that except for System 3 (Income Security), on which old and young people performed equally well, young people achieved significantly better performances than old people.

Ordinal Pattern Analysis (OPA) provides another non-parametric way of examining the same data. Starting with the prediction that young people will perform better than old ones, it examines each possible pair of scores for young and old participants and determines in each case whether the scores confirm the hypothesis (Hit) or contradict it (Miss). Hits and Misses are summed separately and are used to produce an Index of Observed Fit (IOF). $IOF = (Hits - Misses) / (Hits + Misses)$. The IOF ranges from +1.0 (all hits) to -1.0 (all misses), with chance results producing an IOF of 0. A negative IOF always defeats the prediction, while a positive IOF may or may not be statistically significant (Thorngate, 1992). As shown in Table 2, OPA results confirm those of the Mann-Whitney tests.

Do these significant age differences in performance scores reveal more than the significant age differences in the number of successes that were found earlier? It does not appear so. Figure 4, which shows the sums of the performance scores for the 22 old and 22 young participants for each system, is almost identical to Figure 2, which shows the number of their successes. Both figures illustrate that young people performed better than

old people on all systems except System 3 (Income Security), on which they achieved the same degree of success. The only new information revealed by comparing these figures is that while System 6 (United Airlines) caused the greatest age difference in successes (Figure 2), it was System 5 (AccuRate) that caused the greatest age difference in performance scores (Figure 4).

Figure 4. Graph of Performance Scores by Age and System

To determine whether the bimodal distribution of old participants' performance scores was caused by a polarisation of their scores between young-old and old-old people, a Mann-Whitney rank test was performed to examine the difference between the sum of the performance scores of the 12 participants aged 65 to 74 and the sum of the scores of the 10 participants aged 75 to 86. The results indicated that their scores were not significantly different ($U=33$, $p= n.s.$).

Age Differences in Usability Ratings

Participants' evaluations of each IVR system produced usability ratings ranging from 15 to 50, with a higher rating indicating a more positive impression. These ratings had near-normal distributions for both young and old participants on all systems except System 3 (Income Security), which produced a skewed distribution because most old and young participants gave it very high ratings. Table 3 provides the medians, means and standard deviations of the usability ratings given by old and young participants to each of the IVR systems. The means are invariably larger for younger than for older participants, while the standard deviations are greater for old participants on all but System 2 (Canadian Airlines), reflecting the greater variability in the ratings of old people.

Table 3

Medians, Means and SDs of Usability Ratings by Age and System

	Old participants		Young participants	
	Median	Mean (<u>SD</u>)	Median	Mean (<u>SD</u>)
System 1 (Regional Municipality)	26.0	26.5 (7.8)	34.0	33.5 (4.7)
System 2 (Canadian Airlines)	28.0	28.5 (6.0)	31.0	31.9 (7.3)
System 3 (Income Security)	38.0	36.5 (5.0)	38.0	37.8 (4.4)
System 4 (Statistics Canada)	31.5	32.2 (7.1)	35.5	35.2 (6.6)
System 5 (AccuRate)	26.0	28.3 (7.4)	35.5	33.7 (7.3)
System 6 (United Airlines)	28.5	29.8 (9.0)	39.0	39.9 (7.4)

To maintain consistency in the analyses while acknowledging the non-normality of the usability data for System 3, all the analyses were performed using both parametric and non-parametric methods. To establish if old and young participants gave different usability ratings to the IVR systems, a 2 (age: old, young) $\times$ 6 (system: 1 to 6) mixed-design factorial analysis of variance (ANOVA) was performed on the usability ratings of old and young participants for the six telephone systems. The between-subjects variable was age and the within-subjects variable was type of system. The results revealed at least one significant difference among the usability ratings given to the various systems, $F(5,210)=9.7$, $p < .001$; $\eta^2=.19$, a significant difference between the ratings of old and young participants, $F(1,42)=16.8$, $p < .001$, $\eta^2=.29$, and a significant interaction between age and type of system, $F(5,210)=3.3$, $p < .01$, $\eta^2=.07$.

To determine if the interaction effect between system and age qualified the main effect of age, a graph of this interaction is shown in Figure 5. The graph shows that young people gave higher mean usability ratings than old people to all systems, so that the significant effect of age is meaningful in spite of the interaction. Figure 5 also reveals a pattern that is very similar to the one shown in Figure 2 for successful tasks and Figure 4 for performance scores. On the other hand, Figure 5 differs from Figures 2 and 4 in some respects: old people's mean usability ratings for System 1 (Regional Municipality) appear low and their ratings for System 2 (Canadian Airlines) appear high compared with their performance and success scores; and young people's mean usability ratings of System 6 (United Airlines) seem high relative to their performance and success scores. For System

6, the difference between young people's performance scores and their usability ratings is probably due to the fact that the performance measures are limited by a ceiling effect.

Figure 5. Graph of Mean Usability Ratings by Age and System

To verify age differences in the usability ratings old and young participants gave to the IVR systems, follow-up pairwise comparisons (using Bonferroni's correction, $\alpha = .05/6 = .008$) were performed between the ratings for the six systems. The results, shown in Table 4, indicate that young people gave significantly higher ratings than old people to System 1 (Regional Municipality of Ottawa-Carleton) and System 6 (United Airlines), but that their ratings of the other four systems were not significantly different. As also shown in Table 4, parallel analyses using the non-parametric Mann-Whitney U test found that young people gave higher usability ratings to Systems 1, 6 and 5 (AccuRate).

Table 4

Results of Pairwise Comparisons and Mann-Whitney U Tests of Age Differences in Usability Ratings

	p for pairwise comparisons	<u>U</u>
System 1 (Regional Municipality)	.001***	114.0**
System 2 (Canadian Airlines)	.101	117.5
System 3 (Income Security)	.546	229.0
System 4 (Statistics Canada)	.162	194.0
System 5 (AccuRate)	.019*	146.5*
System 6 (United Airlines)	.000****	89.0****

* $p < .05$ ** $p < .01$ *** $p < .008$ **** $p < .001$

Do these age differences in usability ratings reflect differences in performance? If so, the link is not apparent in the statistical results, since consistent age differences in usability are found only for Systems 1 and 6, while age differences in performance scores were found for all except System 3. To gain a better understanding of the relationship between performance scores and usability ratings for old and young participants, the performance scores and usability ratings for each age group were analysed with a measure of association. Because of the non-normal nature of the performance data, the non-parametric Kendall's tau was used. The results revealed a moderately strong significant

positive correlation between the sums of the usability ratings and the sums of the performance scores for old people, $\tau(22) = .53$, $p < .01$, but no significant correlation for young people, $\tau(22) = .20$, $p = \text{n.s.}$ A closer examination of the plots showing the associations between usability ratings and performance scores for each age and system reveals that for each of the pairs reporting non-significant correlations, the performance scores showed ceiling effects.

Age Differences in the Length of Time Required to Succeed

Unlike the previous analyses, which looked at all the IVR tasks performed by all the participants, the analyses in this section and in the following two are restricted to a subset of the data. This subset includes only data relating to the tasks on which the participants were successful, and the questions to be answered are whether old and young participants who were successful at IVR tasks are different from each other in terms of (1) the length of time (in seconds) they needed to succeed at a task, (2) the number of keystrokes or verbal choices they used in their successful tasks, and (3) the usability ratings they gave to the systems with which they performed successful tasks.

The distributions for length of time are almost all non-normal; in particular, the distributions for young people are all negatively skewed or leptokurtic, or both, with times that were mostly concentrated at the low end around the median. The medians, means and standard deviations for length of time, in Table 5, show that successful old participants were faster at some tasks while successful young participants were faster at others. Old participants' times have greater standard deviations on only three of the IVR systems (Systems 1, 5, 6), while young participants' times vary more on Systems 2, 3, 4.

Table 5**Medians, Means and SDs of Time (in Seconds) to Succeed, by Age and System**

	(n)	Median	Mean(SD)
System 1 (Regional Municipality)			
Old participants	(12)	161.0	194.8 (69.5)
Young participants	(21)	154.0	160.8 (38.9)
System 2 (Canadian Airlines)			
Old participants	(3)	183.0	193.7 (30.4)
Young participants	(12)	158.5	166.5 (43.3)
System 3 (Income Security)			
Old participants	(21)	92.0	109.8 (46.7)
Young participants	(21)	98.0	107.1 (50.0)
System 4 (Statistics Canada)			
Old participants	(8)	110.0	123.4 (25.7)
Young participants	(17)	114.0	143.9 (58.9)
System 5 (AccuRate)			
Old participants	(10)	165.0	187.8 (61.2)
Young participants	(20)	155.5	167.9 (56.4)
System 6 (United Airlines)			
Old participants	(8)	246.0	271.6 (101.2)
Young participants	(21)	211.6	215.9 (40.6)

Because of the non-normal distributions, non-parametric methods of analysis were used. To determine if young people who succeed at IVR tasks accomplish them faster than old people who also succeed at these tasks, six Mann-Whitney rank tests were performed on the length of time young and old people took to complete the tasks. The results indicate that there were no significant differences in the length of time old and young participants required to achieve successful performances on any of the six IVR tasks (System 1: $\underline{U}=85.0$, $p=n.s.$; System 2: $\underline{U}=0.1$, $p=n.s.$; System 3: $\underline{U}=190.5$, $p=n.s.$; System 4: $\underline{U}=62.5$, $p=n.s.$; System 5: $\underline{U}=79.0$, $p=n.s.$; System 6: $\underline{U}=52.0$, $p=n.s.$).

Age Differences in the Keystrokes/Choices Required to Succeed

The medians, means and standard deviations for the number of keystrokes/choices participants used in performing successful tasks are in Table 6, as well as the minimum number of strokes needed to complete each task. Old participants have higher means than young ones on all systems except System 4 (Statistics Canada). They also have greater standard deviations on all but Systems 2 (Canadian Airlines) and 4. On System 4, the standard deviation for the old is zero because all those who were successful used the same, minimum number of keystrokes. The distributions for the keystrokes/choices closely resemble the distributions for "time to succeed;" they are very negatively skewed.

Because of the lack of distribution on System 4 and the several non-normal distributions, non-parametric tests were used. To determine if young people who succeed at IVR tasks accomplish them with fewer keystrokes/choices than old people who also succeed at these tasks, six Mann-Whitney rank tests were performed on the number of keystrokes/choices for these two groups.

Table 6

Medians, Means and SDs of Keystrokes/Choices to Succeed, by Age and System

	Minimum number of choices needed to succeed	(n)	Median	Mean(SD)
System 1 (Regional Municipality)	4			
Old participants		(12)	4.0	4.9 (1.4)
Young participants		(21)	4.0	4.1 (0.4)
System 2 (Canadian Airlines)	25			
Old participants		(3)	25.0	27.3 (4.9)
Young participants		(12)	25.0	26.8 (4.9)
System 3 (Income Security)	2			
Old participants		(21)	2.0	2.8 (1.6)
Young participants		(21)	2.0	2.2 (0.9)
System 4 (Statistics Canada)	5			
Old participants		(8)	5.0	5.0 (0)
Young participants		(17)	5.0	5.4 (0.9)
System 5 (AccuRate)	8			
Old participants		(10)	10.0	11.2 (3.7)
Young participants		(20)	8.0	8.9 (2.5)
System 6 (United Airlines)	10			
Old participants		(8)	15.5	19.9 (10.4)
Young participants		(21)	14.0	14.7 (2.7)

The results indicate that successful old and young participants did not use significantly different numbers of keystrokes/choices to complete the tasks for System 2 (Canadian Airlines; $\underline{U}=15.5$, $p=n.s.$, System 4 (Statistics Canada; $\underline{U}=56$, $p=n.s.$) and System 6 (United Airlines; $\underline{U}=52$, $p=n.s.$), but that old people needed more keystrokes than young ones to succeed on System 1 (Regional Municipality; $\underline{U}=89$, $p<.05$), System 3 (Income Security; $\underline{U}=159.5$, $p<.05$) and System 5 (AccuRate; $\underline{U}=45$, $p<.05$).

These results are very different from the results for "time to succeed," which found no significant differences between old and young participants. Similar results were expected since more keystrokes/choices require more time, as shown in Table 7 by analyses using Kendall's tau that find significant positive correlations on most systems between participants' time to succeed and the number of their keystrokes/choices.

Table 7

Correlations Between Time to Succeed and Keystrokes/Choices to Succeed

	Old participants		Young participants	
	(n)	τ	(n)	τ
System 1 (Regional Municipality)	(12)	.75**	(21)	.31
System 2 (Canadian Airlines)	(3)	.33	(12)	.63*
System 3 (Income Security)	(21)	.60**	(21)	.41*
System 4 (Statistics Canada)	(8)	--- ¹	(17)	.58**
System 5 (AccuRate)	(10)	.51*	(20)	.66**
System 6 (United Airlines)	(8)	.64*	(21)	.48**

* $p<.05$ ** $p<.01$ *** $p<.001$

¹ no variability in keystrokes

Age Differences in Usability Ratings for Successful Tasks

The distributions of the usability ratings of successful participants are very similar to those of the usability ratings for all participants, with near-normal distributions on four of the five systems. As before, the exception is System 3 (Income Security), whose distributions are positively skewed.

The median, mean and standard deviations of usability ratings for successful participants are shown in Table 8. Unlike the medians and means of the usability ratings for all participants shown earlier in Table 3, which were invariably higher for young people than for old ones (except for System 3, where their medians were equal), the mean ratings of successful old people are higher than those of successful young people for System 2 (Canadian Airlines) and System 4 (Statistics Canada). Overall, the usability ratings of successful older participants do not show greater variability than those of successful young participants.

Analyses were performed to examine the difference between the usability ratings of successful old and successful young participants for each of the six IVR systems. As these ratings are normally distributed for all systems except System 3, parametric and non-parametric methods were used. None of the six t-tests that were performed (using Bonferroni's correction, $\alpha .05/6=.008$) produced significant results. Parallel non-parametric Mann-Whitney U tests also yielded non-significant results.

Table 8**Medians, Means and SDs of Usability Ratings for Successful Tasks, by Age and System**

	(n)	Median	Mean(SD)
System 1 (Regional Municipality)			
Old participants	(12)	32.0	30.8(7.9)
Young participants	(21)	34.0	33.3(4.7)
System 2 (Canadian Airlines)			
Old participants	(3)	37.0	37.7(3.1)
Young participants	(12)	38.0	35.4(7.0)
System 3 (Income Security)			
Old participants	(21)	38.0	36.6(5.1)
Young participants	(21)	38.0	37.8(3.9)
System 4 (Statistics Canada)			
Old participants	(8)	40.0	38.3(5.3)
Young participants	(17)	36.0	36.1(6.2)
System 5 (AccuRate)			
Old participants	(10)	34.5	33.8(6.4)
Young participants	(20)	36.0	34.8(6.5)
System 6 (United Airlines)			
Old participants	(8)	35.5	33.1(8.6)
Young participants	(21)	40.0	40.1(7.5)

Association Between Age, Attention/Memory, Technological Experience and IVR Results

The participants' scores on the test of attention/memory (referred to henceforth as the memory test) had near-normal distributions, as did their scores on the questionnaire about their prior experience with IVRs and other technological products. For the memory test, the mean score for old people was 24.2 ($SD=6.9$), with a median of 21 and a range from 17 to 42; for young people, the mean score was 30.5 ($SD=7.3$), with a median of 31.5 and a range from 20 to 47. For prior technological experience, the mean score for old people was 40 ($SD=14.8$), with a median of 41.5 and a range from 3 to 62; for young people, the mean was 68.1 ($SD=9.0$), with a median of 68 and a range from 49 to 86.

To understand the relationships between participants' ages, their memory scores, their prior technological experience, the sum of their performance scores and the sum of their usability ratings, these data were analysed with a measure of association. Because of the non-normal nature of the performance scores, the non-parametric Kendall's tau was used. The results, in Table 9, indicate that age is significantly and negatively correlated with prior technological experience and with performance on a memory test. As the graph on the left in Figure 6 illustrates, the first of these correlations reflects the fact that almost all young participants have more technological experience than almost all old participants. The situation is quite different for attention/memory; the graph on the right shows that although young people are more likely than old ones to obtain higher scores on attention/memory tests, there is a considerable degree of overlap in the scores of old and young participants, including one old woman who obtained a higher attention/memory score than all but two of the young participants.

Table 9

Intercorrelations (Kendall's tau) Between Age, Technological Experience, Memory Scores, Sums of Performance Scores and Sums of Usability Ratings (N=44)

	Age	Tech. Experience	Memory	Performance
Tech. experience	-.48**			
Memory	-.33**	.27*		
Performance	-.45**	.46**	.40**	
Usability	-.38**	.40**	.14	.52**

Note: * $p < .05$ ** $p < .01$

Figure 6. Scatterplots of Age with Technological Experience and Memory Scores

The results in Table 9 also show that performance is positively correlated with both technological experience and attention/memory, that usability ratings are positively correlated with technological experience but not with attention/memory, and that technological experience and attention/memory are correlated with each other. The results mean that in this study, people who had better memories and who had been more exposed to new technologies tended to do better at IVR tasks, people who were more technologically savvy gave more positive evaluations to IVR systems, and good attention/memory abilities and technological experience tended to be found in the same persons.

The difficulty with concluding anything from these associations, such as for example that better cognitive abilities appear to cause people to obtain higher scores, is that in these analyses the effects of cognitive abilities (and the effects of prior technological experience) are confounded with the effects of age, so that any effects attributed to having worse attention/memory capacities could in fact be due to other characteristics of old people, such as a less acute hearing. To explore whether the correlations shown in Table 9 are still significant when the age confound is removed, the same analyses of association were performed separately with Kendall's tau on the data for each age group.

These results, in Table 10, suggest that when the age confound is removed, the only significant (positive) though not very strong correlations that remain are between attention/memory and performance for young participants ($\tau=.37$), and between attention/memory and technological experience for old participants ($\tau=.34$). This suggests

that better cognitive abilities are linked with better performances only among young people, and that prior technological experience is not linked with either performance or usability ratings for either age groups. The moderately strong correlation shown between performance scores and usability ratings ($r=.53$) reaffirms earlier results.

Table 10

Intercorrelations (τ) Between Age, Technological Experience, Memory Scores, Sums of Performance Scores and Sums of Usability Ratings, by Age (N=22 each)

	Age	Tech. Experience	Memory	Performance
Tech.experience				
Old participants	-.17			
Young participants	.27			
Memory				
Old participants	-.26	.34*		
Young participants	-.04	-.25		
Performance				
Old participants	-.25	.14	.19	
Young participants	.12	.15	.37*	
Usability				
Old participants	-.15	.19	-.01	.53*
Young participants	.01	.12	-.09	.20

Note * $p < .05$

To further clarify the relationship between attention/memory and performance scores, analyses of association were performed with Kendall's tau for each system separately for each age groups. These twelve analyses confirm the weakness of the link between attention/memory and performance by producing only one significant result, possibly obtained by chance, showing a link between higher cognitive abilities and better performance scores for old people on System 2 (Canadian Airlines; $\tau(22) = .42$, $p < .01$).

Differences Between the IVR Systems in Performance Scores

The purpose of this last series of analyses is to compare the different IVR systems to each other, first on the performance scores they produced, and secondly on the usability scores they were assigned by the participants. Performance scores are preferred to success scores in this context because they include partial as well as full successes and are therefore a more sensitive measure of task difficulty.

The medians, means and standard deviations of the performance scores for each system were shown above in Table 1. As mentioned earlier, the distributions of the performance scores of old and young people on each system are not normally distributed. As a result, non-parametric methods are used to analyse the effects of the different IVR systems on the scores of old and young people. No analysis is done on the total performance scores for both age groups because the gap between the scores of old and young participants varies so much from one system to another (see the interaction of system and age in Figure 4) that total results for both age groups could be misleading.

To establish whether different IVR systems produce different performance scores, a Friedman test for K related samples was conducted on the scores of old participants and

another Friedman test was conducted on the performance scores of young participants. The test for old participants, $\chi^2(5N=22) = 34.4$, $p < .001$, and the test for young participants, $\chi^2(5N=22) = 22.2$, $p < .001$, were both significant, indicating at least one difference in each age group among the scores produced by the six systems.

Table 11 shows the ranks of the sums of the performance scores for each system for old and young participants. System 3 (Income Security) produced the highest total scores for old people, with all the other systems producing much lower scores. The ranking for young people is the same as for the old with one reversal: System 6 (United Airlines) produced the highest scores for them, followed by System 3. The range of the sums of performance scores for the old (9.3 to 21.5) is much wider than the range for the young (16 to 21.5). Table 11 also shows a reverse linear relationship between performance scores and the number of keystrokes needed to succeed - the more keystrokes are required, the lower the score - but verbal choices such as those required in using System 6 (United Airlines) seem to follow a different rule.

To determine if the performance scores old and young people obtained on each of the systems were significantly different from each other, Wilcoxon tests were performed separately for old and young people on the scores for pairs of systems, starting with the pair that showed the largest differences in the sums of their performance scores in Table 11, continuing in descending order of differences, and stopping after two consecutive non-significant results. The results of these tests appear in Table 12.

Table 11**Ranking of IVR Systems by Sums of Performance Scores, Old and Young Participants**

	Minimum number of choices needed to succeed	Old		Young	
		Rank	Sum of Scores	Rank	Sum of Scores
System 3 (Income Security)	2	1	21.5	2	21.5
System 6 (United Airlines)	10	2	14.9	1	21.8
System 1 (Regional Municipality)	4	3	14.8	3	21.3
System 4 (Statistics Canada)	5	4	14.2	4	21.0
System 5 (AccuRate)	8	5	11.5	5	20.9
System 2 (Canadian Airlines)	25	6	9.3	6	16.0

The Wilcoxon test results for old participants shown in Table 12 divide the systems into three distinct groups. System 3 (Income Security), alone at the top, produced significantly higher performance scores than all the other systems. System 2 (Canadian Airlines), at the bottom, produced significantly lower scores than all the other systems except System 5 (AccuRate), which straddles the bottom and middle groups. The middle group includes the rest of the systems (6, 1 and 4), which produced scores that are not significantly different from each other.

Table 12

Results of Wilcoxon Tests of Differences Between the IVR Systems in Performance Scores for Old and Young Participants

Old participants	z	Young participants	z
Systems 3 and 2	-3.8***	Systems 2 and 6	-2.7**
Systems 3 and 5	-3.1**	Systems 2 and 3	-2.7**
Systems 3 and 4	-3.1**	Systems 2 and 1	-2.1*
Systems 3 and 1	-2.9**	Systems 2 and 4	-2.6**
Systems 3 and 6	-3.2**	Systems 2 and 5	-2.5*
Systems 2 and 1	-2.5*	Systems 6 and 4	-2.1*
Systems 2 and 6	-2.8**	Systems 6 and 5	-1.0
Systems 2 and 4	-2.3*	Systems 6 and 1	-0.4
Systems 5 and 6	-1.3		
Systems 5 and 1	-2.1*		
Systems 5 and 4	-1.0		
Systems 5 and 2	-1.0		

* $p < .05$ ** $p < .01$ *** $p < .001$

The Wilcoxon results for young participants shown in Table 12 reveal a division of the systems into two distinct groups. The first is composed of System 2 (Canadian Airlines), alone at the bottom, which produced significantly lower performance scores than all the other systems. The second group includes the rest of the systems, which produced

scores that were not significantly different from each other. The only exception was a significant difference between the scores for System 6 (United Airlines) at the top and System 4 (Statistics Canada) in the middle.

Differences Between the IVR Systems in Usability Ratings

When age differences in usability ratings were examined earlier, it was concluded that the distributions of the ratings for old and young participants were normally distributed except for System 3 (Income Security), whose distributions were positively skewed because both old and young participants gave it high ratings. The medians, means and standard deviations of the usability ratings for old and young participants for the six IVR systems were shown in Table 3.

Because of the non-normal distribution for System 3, these data are analysed using both parametric and non-parametric methods. The parametric tests are follow-ups of the 2 x 6 (age x system) mixed-design ANOVA performed earlier on the usability scores. As described earlier, the results of the main analysis indicated at least one significant difference among the usability ratings for the various systems and a significant interaction between age and type of system. The interaction, which was illustrated in Figure 5, revealed that the gaps between the usability ratings of old and young participants are so different from one IVR system to another that no meaningful conclusions could be drawn from analysing the total usability ratings for both age groups combined. Separate analyses are therefore performed for each age group.

Table 13 shows the ranks of the usability ratings for each system for old and young participants. Once again, System 3 (Income Security) is in first place for old people and

System 6 (United Airlines) is in first place for young people. The main difference between these usability rankings and the rankings for the performance scores in Table 12 is the position of System 1 (Regional Municipality). Although this system (1) produced the third highest performance scores for both old and young people, (2) was in a tie for producing the highest number of successes for young people, and (3) produced the second highest number of successes for old people, it was rated the least usable system by old people and the second least usable by young people.

Table 13

Ranking of IVR Systems by Sums of Usability Ratings, Old and Young Participants

	Old		Young	
	Rank	Sum of ratings	Rank	Sum of ratings
System 3 (Income Security)	1	802	2	821
System 4 (Statistics Canada)	2	709	3	774
System 6 (United Airlines)	3	655	1	878
System 2 (Canadian Airlines)	4	627	6	701
System 5 (AccuRate)	5	622	4	741
System 1 (Regional Municipality)	6	583	5	736

To determine if there were differences in the usability ratings the participants gave to the IVR systems, the earlier ANOVA analyses of these ratings were supplemented by 24 follow-up pairwise comparisons (12 for old participants and 12 for young ones, using Bonferroni's correction, $\alpha .05/24=.002$) performed on the 12 pairs of usability ratings for which hypotheses were formulated. According to these hypotheses, which were based on pre-test results, System 3 (Income Security) and System 6 (United Airlines) were expected to obtain the highest usability ratings while System 2 (Canadian Airlines) was expected to obtain the lowest ratings.

Old people's results (see the *p* column on the left in Table 14) confirm that they gave top usability ratings to System 3, as predicted, but do not support the predictions relating to Systems 6 and 2. Young people's results (in the *p* column on the right in Table 14) support the hypothesis predicting that Systems 6 and 3 would obtain the highest ratings and System 2 the lowest. Overall, these results reveal different patterns for old and young participants: old people's results show System 3 alone at the top with significantly higher ratings than all the other systems except for System 4 (Statistics Canada), which straddles the top and bottom positions; young people's results show Systems 6 and 3 at the top, System 2 at the bottom with significantly lower ratings than System 6, and the rest of the systems straddling these two poles.

Table 14

Results of Pairwise Comparisons and Wilcoxon Tests of Differences Between the IVR Systems in Usability Ratings, for Old and Young Participants

Old participants	p ¹	z ²	Young participants	p ¹	z ²
Systems 3 and 1	.000****	-3.7****	Systems 6 and 2	.000****	-3.3**
Systems 3 and 5	.000****	-3.4**	Systems 6 and 1	.004**	-3.1**
Systems 3 and 2	.000****	-3.6****	Systems 6 and 5	.008**	-2.8**
Systems 3 and 6	.001***	-2.4*	Systems 2 and 3	.002**	-2.4*
Systems 4 and 1		-2.6**	Systems 6 and 4	.023*	-2.6**
Systems 4 and 3	.015*	-2.3*	Systems 3 and 1	.020*	-2.4*
Systems 4 and 5		-2.1*	Systems 3 and 5	.043*	-2.2*
Systems 4 and 2	.038*	-2.3*	Systems 6 and 3	.187	-1.0
Systems 6 and 1	.133	-0.9	Systems 2 and 4	.064	-1.5
Systems 6 and 4	.228	-1.4	Systems 3 and 4	.205	
Systems 6 and 2	.505		Systems 2 and 1	.332	
Systems 6 and 5	.504		Systems 2 and 5	.340	
Systems 2 and 1	.224				
Systems 2 and 5	.905				

¹ p for pairwise comparisons

² Wilcoxon results

* p < .05 ** p < .01 *** p < .002 **** p < .001

As well as these parametric tests, two non-parametric Friedman tests were also conducted on the usability ratings to determine whether they differed for the different IVR systems. The test for old participants, $\chi^2(5N=22) = 24.8$, $p < .001$, and the test for young participants, $\chi^2(5N=22) = 20.2$, $p < .01$ were both significant. To follow up on these results, Wilcoxon tests were performed separately for old and young people on the usability ratings for pairs of systems, starting with the pair that showed the largest differences on Table 13, continuing in descending order of differences, and ending after two consecutive non-significant results.

The results of these tests, shown in Table 14 (in the z columns) are slightly different but consistent with the results of the ANOVA follow-up comparisons. Old people's ratings divide the systems into three distinct categories: System 3 (Income Security) comes first, with significantly higher ratings than all the other systems; System 4 (Statistics Canada) comes second, with ratings lower than those for System 3 but higher than those for all the other systems except System 6 (United Airlines); then come the rest of the systems, including System 6, with ratings that are not significantly different from each other's. Young people's results show the same two groups as the parametric test results: Systems 6 and 3 together at the top, with higher ratings than the rest of the systems which are together at a slightly lower level.

Case Studies - Features of IVR Systems and their Effects

The following case studies provide details about the IVR systems and the tasks performed by the participants, summarize the experiences of old and young participants with each system, and examine errors, ratings and comments that provide additional

information on the reasons why participants performed well or badly and rated each system as they did. The systems are examined in descending order starting with the one on which old participants obtained the most successes and ending with the system on which they did least well.

Income Security (System 3). This system, whose structure is shown in Appendix E-3, started with a long public announcement that changed every few weeks during the testing period, followed by instructions to press 1 if you have a touchtone phone, which brings the caller to a first-level menu of three options, the third being a live operator (during business hours). The first two options called up other announcements and second-level menus that led to recorded information. The task consisted of finding out what action the caller should take if she was a pensioner whose pension cheque, which normally came on the same date every month, was four days overdue (the answer was to wait one more day and then call again and speak to an operator). Of the six tasks, this one required the fewest number of choices to succeed: only two keystrokes, one to indicate that the caller had a touchtone system and one choice (press 2) from the first-level menu. This system produced the highest performance scores and usability ratings for old participants and the second highest for the young: 21 old and 21 young participants succeeded at this task. As shown in Table 15, most old and young participants found this system easy (19 and 19 out of 22) and felt in control while using it (19 and 17); 11 seniors and 6 young people enjoyed the experience. On the negative side, four seniors found it very frustrating, while six old people and seven young ones found it confusing.

Table 15 - Details of Answers to Usability Questionnaires (N = 22 each old and young)

	System 1				System 2				System 3			
	Old		Young		Old		Young		Old		Young	
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
<u>Positive statements</u>												
Easy to use	9	3	10	21	1	0	9	2	11	12	2	8
Felt in control	10	1	11	18	3	1	10	0	12	11	2	9
Voice very clear	16	1	5	17	0	5	21	0	1	20	1	1
Good impression	7	3	12	10	7	5	6	6	10	11	6	5
Enjoyed using	2	1	19	5	8	9	4	4	14	7	6	9
<u>Negative statements</u>												
Very frustrating	15	1	6	6	2	14	16	1	5	10	1	11
Confusing choices	17	0	5	12	0	10	17	0	5	13	1	8
Must concentrate hard	18	0	4	14	1	7	15	1	6	9	4	9
Sometimes felt lost	13	0	9	7	4	11	13	0	9	12	0	10
Voice too quick	10	0	12	5	2	15	6	1	15	4	1	17
<u>Positive statements</u>												
Easy to use	17	1	4	19	2	1	10	1	11	16	1	5
Felt in control	14	1	7	17	1	4	9	0	13	17	1	4
Voice very clear	20	0	2	12	2	8	16	1	5	14	4	4
Good impression	12	3	7	14	6	2	8	4	10	14	5	3
Enjoyed using	9	4	9	9	6	7	6	2	14	11	5	6
<u>Negative statements</u>												
Very frustrating	9	1	12	5	3	14	11	1	10	3	4	15
Confusing choices	10	0	12	1	1	20	14	0	8	14	1	7
Must concentrate hard	12	2	8	8	3	11	17	2	3	11	3	8
Sometimes felt lost	9	0	13	3	0	19	14	0	8	8	3	11
Voice too quick	11	0	11	3	0	19	7	2	13	5	1	16

The most frequent comment many (at least six) participants made about System 3 was that it was "straightforward." The senior participants who were frustrated by this system, as well as four young ones, said "there was too much talk before getting to the point." Most of the participants who found this system confusing said it was because they felt that both first-level choices could have applied to the task question. Two participants (one old and one young) picked the wrong option because they found it illogical to pick the (correct) option which said "If you are currently receiving benefits," when they were calling because they had not received theirs.

In summary, the Income Security task was easy for most old and young participants and was the favourite IVR task of senior participants. The key elements which made this task easy appear to be the system's narrow first-level menu (only two choices apart from choosing a live operator) and the very small number of choices required to succeed (only one in addition to choosing English). In spite of their success, however, almost a third of old and young participants felt confused by the lack of specificity of the terms used to describe the IVR options.

Regional Municipality of Ottawa-Carleton (System 1). The choices offered by this IVR system are described in Appendix E-1. After hearing a long introductory message and choosing English or French, callers were presented with the following: a first-level menu of six possible choices/items in which each item consists of a long list of topics; a second-level menu with as many as eight items each; a third-level menu with a small number of choices, usually leading to taped information on one of a multitude of subjects under the jurisdiction of the regional municipality.

The participants' task consisted of finding the municipality's "employment line" and noting down how many jobs were being offered that day (the answer was always "none"). This involved making four choices reaching down to the deepest level. This task was the second most successful for old participants and the most successful (tied with two others) for the young: 12 of the old people and 21 of the young succeeded at it. The most frequent error made with this system, by six old participants and two young ones, was to miss hearing the "job" option because it was buried in a long enumeration. The second most common error, made by five old people, was to choose the wrong option; this happened when those who had not heard the "jobs" option opted for the "accessibility" option because they interpreted it to mean "access to jobs" (it referred to programs for handicapped people). In terms of usability, old participants rated this system lowest and young participants second-to-lowest.

As shown in Table 15, a majority of both old and young participants found this system confusing and said it required them to concentrate very hard in order to catch the correct option; 10 old people and 5 young ones found the voice too quick. Young people were not greatly troubled by these problems: almost all of them found the system easy to use and felt in control. Close to half of the old participants felt similarly, but the other half found the opposite, that the system was difficult to use and that they had not felt in control. Old participants were over twice as likely to feel frustrated (15 old compared with 6 young) and more likely to say they had felt lost (13 compared with 7 young) and had not enjoyed using this system (19 compared with 9 young). Many participants from both age groups complained that this system was boring, with badly organized and

excruciatingly long menus that made it very difficult to keep their concentration and gave the impression that the callers' time was not important. Young people were mostly resigned to this, but many old people expressed indignation and said they had "no time to waste on this nonsense."

To summarize, this Regional Municipality IVR task was easy though not pleasant for almost all young participants, but it was difficult for half of the old participants and disagreeable for almost all of them. The main problems seemed to be too many choices in each menu, too many items under each of these menu choices, and a monotonous and rapid voice that failed to make the options distinct. A secondary problem was the use of bureaucratic jargon in naming the "accessibility" option.

AccuRate (System 5). The structure of this system is shown in Appendix E-5. It started with a long introduction about the current rates for buying or selling U.S. cash and then went on to a first-level menu with four options including a live operator, followed by two more levels with four or five choices each. The task consisted of finding out how many Chilean pesos can be bought with one Canadian dollar. This involved making eight keystroke choices to reach the third-level menu and to enter the first three letters of the country name ("CHI").

This system produced the third largest number of successes among old participants (10 succeeded) and the fourth among young participants (20 succeeded). In terms of usability, it was in fifth place out of six for the old and in fourth place for the young. The most frequent error with this system, made by eight old people and six young ones, was to mistakenly choose U.S. instead of Canadian options. The participants who made this

mistake, along with several others, complained that U.S. options were so ubiquitous in this system that they were very confusing. For example, after listening to a menu offering "For U.S. rates, press 1" and "For other currency, press 2" and pressing the "2" key to choose option 2, callers were presented with a menu that started with, "For U.S. rates under \$10,000, press 1," "For U.S. rates over \$10,000, press 2," etc. The second type of error, made by four old persons and four young one, was caused by problems in entering "CHI" (for Chile); three young people and one senior were so confused that they tried to enter "CAN" (for Canada) instead. One old participant commented that she had never noticed there were letters on a telephone's keys.

As shown in Table 15, half or more of both old and young participants found this system confusing (14 old and 14 young) and felt it required them to concentrate very hard (17 old and 11 young), and a quarter to a third thought the voice was too quick (7 old and 5 young). Most young people had no difficulty surmounting these problems: all six who chose a U.S. option by mistake managed to correct this error and to succeed at the task; 16 young people found the system easy to use, 17 felt in control and only 3 felt frustrated. For their part, only one of the eight old people who mistakenly chose a U.S. option eventually succeeded at the task; only 10 old people found the system easy to use, only nine felt in control, and 11 found the experience very frustrating; 14 did not enjoy the experience, compared with 6 young people.

In summary, this AccuRate IVR task posed problems for all participants. These problems were quite easily surmounted by almost all young people but by less than half of the old. The main difficulty was a confused and illogical structure of menus. A second

problem which affected a small number of old and young people was their lack of experience or their lack of ability with entering information by means of keystrokes.

United Airlines (System 6). Unlike the other five systems, this one was almost entirely voice-activated, so that callers expressed their choices verbally and seldom used keystrokes. As illustrated in Appendix E-6, the exchanges between the United robot operator and callers do not follow a structure with a small number of pre-determined paths, but proceed instead like a real-life question-and-answer session in which there are a multitude of possible directions and outcomes. The task consisted of finding out if United Airlines had a flight from Toronto to New York (Kennedy Airport) leaving around 6:00 p.m. the next day, and to note down the flight number and departure time. A minimum of 10 verbal commands, consisting of one or a few words each, were required to succeed.

This task produced the best performance score of all systems for young people and the second best score for old people; 21 of the young participants and eight of the old ones successfully completed the task. In terms of usability, United produced the highest ratings for young people and the third highest ratings for old people. One of the most common caller error with this system, which was made by eight old persons and one young one, was to answer questions with full sentences instead of the required keywords. Five of these old persons said they had thought throughout the call that the United robot was a real person. One provoked old participant ended up saying "Listen to me, young man. I want you to tell me if there's a flight . . .". When another old participant was asked "Did you think it was a real person?", she answered "I don't know what it was but it sure was stupid." Most of these calls were eventually re-routed to a live operator.

Eleven young women and nine old ones ran into difficulties when the computer could not recognize their verbal answers. For example, after answering that their city of departure was "Toronto," some participants were prompted (sometimes several times in a row) to confirm that their flight was departing from "Tokyo," or "China," or "Atlanta;" this occurred more frequently with young participants (7) than old ones (4). Another example was that after asking callers to choose one of the various New York airports, the United computer recognized "Kennedy" but did not recognize "Kennedy airport," to the frustration of four young and six old participants who repeated "Kennedy airport" over and over again but had it rejected every time.

A common error, made by one young person and seven old ones, was the failure to catch, near the end of the call, that the flight number the computer was about to provide was for today, instead of tomorrow as the task required. If this was caused by a memory lapse, it is not a serious problem because people who phone airlines are unlikely to forget the date on which they want to depart. At this point in the call, two old women who were slightly hard of hearing said over and over that the "day" was wrong, but the computer could only recognize the word "date"; one student was caught in "IVR hell" when she repeated over and over that she wanted a different "time," which led to prompts about the hour she preferred, following which she was told again and again the flight number for today, until she finally hung up. Most participants who mistakenly accepted the flight number for "today" were left with the impression that they had succeeded at this task.

Of the 16 seniors who encountered problems with the United Airlines task, only two went on to succeed. By contrast, 14 of the 15 young participants who had difficulties

recovered and went on to succeed. When young people who had encountered problems using the United system were asked why they nevertheless gave it high usability ratings, they answered that they did not mind the mistakes so long as they were given an opportunity to recover. The United system is very good in this respect; callers can interrupt and restart or correct their answers at all stages of the process.

The United system elicited superlatively delighted responses ("pretty neat," "really cool," "awesome," "wonderfully weird") from many young participants and some old ones. When asked if they agreed with the statement "I enjoyed using this service," eight young participants said they agreed and another 11 said they strongly agreed; five old participants agreed and four strongly agreed, but 11 old participants said they had not enjoyed the experience. A deep age division is also found for all the other usability questions for this system: almost all young people gave top ratings to all aspects of it, but half to two-thirds of old people found it difficult and very frustrating, with feelings of being lost and without any sense of control.

To summarize, the United Airlines voice-activated system was enthusiastically welcomed by those who succeeded or thought they had succeeded at the task, including almost all young people and close to half of the old (some of whom had accepted the "today" flight number). On the other hand, this system was strongly criticized by the old participants who felt they had failed. The enthusiasm generated by the United system is particularly remarkable given the many evident flaws of this system, including its lack of a clear statement that it is an automated system and not a real person, its extreme literalness

and failure to recognize less-than-perfect speech and common possible answers, and its confusing lack of a specific prompt about the date of travel.

Statistics Canada (System 4). The structure of this system is shown in Appendix E-4. It had a very long introductory message, including an Internet address and the number of the Human Resources Department, followed by a choice of English or French, a first-level menu of three items (including a live operator during business hours), followed by second-level menus with three or four choices, and then in some cases by short third-level menus giving access to recorded information. The task consisted of finding out and noting down the population of Quebec and Ontario at the time of the 1996 Census. This involved making five choices leading to the system's deepest level. In terms of performance scores, this system ranked fourth out of six for both old and young participants; eight of the old people and 17 of the young ones succeeded at the task. In terms of usability, this system was second for old participants and third for young ones.

The most frequent error in using this system, made by five young people and three old ones, was to write down an incomplete or wrong population number. This was attributed to the fact that the taped voice reciting the population figures was much too fast, "like at an auction," said one participant, and to the lack of a direct option to have these figures repeated (it could only be done in a roundabout way by returning to the previous level, which few participants realized). Another main error, made by seven old people, was failing to hear the "population" option in the first or second-level menus. This was blamed on two things: one was the fact that there was "way too much talk" at the beginning of the call, causing participants to become distracted and miss their cue; the

other was the system's lack of an option to repeat the options just listed. Seven young people and one old one also complained about having to listen to many more figures than necessary (all the populations were read in order from the Atlantic to the Pacific); they said that callers should be able to select the jurisdiction they want.

Three old participants and six young ones said they were surprised at the almost-amateurish quality of this IVR system because they expected something more sophisticated from Statistics Canada. Two old participants and three young ones felt that Statistics Canada was purposefully trying to discourage people from telephoning and was encouraging them instead to obtain information from its Internet site, which few seniors are able to do. In spite of these criticisms, Table 15 shows that most of the old and young participants found this system easy to use, although less than half of either age group had enjoyed the experience.

In sum on Statistics Canada's system, half of old participants and close to a quarter of young ones failed at this task for reasons that can be blamed on the system's poor design, including a very long introduction that obscures its menu choices, a voice that is too fast to take numbers down, and the lack of features such as options to repeat and options to select individual items instead of having to listen to long lists.

Canadian Airlines (System 2). As illustrated in Appendix E-2, this system's message started with a long introduction, including a message in French followed by a pause. Callers were then presented with a series of questions or prompts that were very similar to the ones callers were asked in the voice-activated United Airlines system (System 6). In this case, however, callers had to express their answers by means of

keystrokes. The task, which was almost identical to the one used in calling United Airlines, was to find out if Canadian had a flight from Ottawa to Halifax leaving Ottawa around 5:00 p.m. the next day, and to note down the flight number and departure time. A minimum of 25 keystrokes were required to achieve success.

The Canadian Airlines system produced the lowest scores for both old and young people; only three old participants and 12 young ones succeeded at the task. In terms of usability, this system was rated fourth out of six by old participants and sixth by young participants. As shown in Table 15, three-quarters of the old participants and half of the young ones found this system very frustrating and confusing. Among the 31 types of errors participants made in accomplishing this task, 22 occurred when callers were trying to enter data by means of keystrokes. The most common entry error, made by three old people and four young ones, was to use three instead of four keystrokes to identify the departure or arrival cities. This was to be expected since the prompt said to "enter the first four letters of the name of the city, or the three-letter airport code." Other entry errors included using the zero key instead of "O" for Ottawa (three old and one young), entering the wrong information (four young and one old), entering the information too slowly (three old) and not knowing how to enter minutes (three old; the answer was "00"). Another common error (by four old and four young people) was due to a complicated question asking if the caller knew or did not know the departure and arrival cities and approximate departure time etc.

Recovery was impossible for almost all the errors; the calls were instantly aborted by the system and re-routed to a live operator (when available). This lack of recovery

options was strongly criticized by seven old participants and five young ones, who felt that the system's instant aborting of their calls for a minor error made them feel incompetent.

Another common complaint, by seven unilingual old people and two young ones, was that at the beginning of the call, when they heard a (seemingly) long and incomprehensible message in French, followed by a pause, they did not know what was going on and almost hung up. Five seniors commented that they found it slow and awkward to find the alphabet on the telephone keys, so that more time should be allowed to enter the information.

In summary, the Canadian Airlines system was so badly designed that few old participants and only half of young participants were able to use it successfully to perform a seemingly simple task. The system had numerous undesirable features, including an off-putting French introduction and pause, several unclear or confusing instructions, insufficient delays to enter the required information, and the lack of recovery options when errors occurred.

Typology of Difficulties Encountered in Using IVR Systems

To obtain an overview of the problematic features of IVR systems, the available information was boiled down to eight main types of problems as shown in Table 16. This information came from a combination of sources, including comments, answers to the usability questionnaires, and listening to the participants' struggles while they made the IVR calls. For example, participants who made an error during a call, or who answered a usability question by saying they had found a system confusing, were asked to explain what had caused them to make that particular error or what had made them feel confused.

Table 16**Main Problems with IVR Systems Reported by Participants, in Order of Frequency**

Type of problem & systems	Number	%	Old	Young	Ratio O/Y
Unclear choices or instructions	71	(26.5)	37	34	1.1
Canadian Airlines	30		17	13	1.3
AccuRate	28		14	14	1.0
Income Security	13		6	7	0.9
Voice spoke too quickly	59	(22.0)	39	20	2.0
Regional Municipality	15		10	5	2.0
Statistics Canada	14		11	3	3.7
AccuRate	12		7	5	1.4
All others	18		11	7	1.6
Too-long intros/menus/items	41	(15.3)	22	19	1.2
Regional Municipality	24		12	12	1.0
Statistics Canada	17		10	7	1.4
Verbal data entry problems - United	30	(11.2)	18	12	1.5
Voice recognition failure	21		10	11	0.9
Caller speaking in sentences	9		8	1	8.0
Lack of means of recovery from error at Canadian Airlines	25	(9.3)	16	9	1.8
Keystroke data entry problems	23	(8.6)	16	7	2.3
Canadian Airlines	19		13	6	2.2
AccuRate	4		3	1	3.0
Caller confusion about the task	14	(5.2)	9	5	1.8
- failure to reject "today"					
in United Airlines	8		7	1	7.0
- other caller errors	6		2	4	0.5
Use of jargon - Regional Municipality	5	(1.9)	5	0	5/0
TOTAL	268	(100%)	162(60.4%)	106(39.6%)	1.5

Table 16 reveals that unclear choices/instructions were the most frequent source of problems with IVRs in this study, accounting for 26.5% of the difficulties. More old participants (37 reports) than young ones (34 reports) took notice of this. Second most common (22% of the problems) were too-rapid automated voices, with 39 reports from old people and 20 from the young. The next two problems were too-long menus/texts (15.3% of problems; 22 reports from the old and 19 from the young) and verbal data entry difficulties in the United Airlines system (11.2%). These latter difficulties included (a) the system's rejection of less-than-perfectly voiced words and terms, which affected old people (10 reports) and young ones (11 reports) almost equally; and (b) callers using full sentences instead of the exact words of the prompts, which was much more frequent among old people (8 reports) than young ones (1 report). Next in frequency (9.3% of problems) was the lack of provision for recovery from error in the Canadian Airlines system, which affected many more old people (16 reports) than young ones (9 reports). This large age discrepancy reflects the fact that old people made more errors and therefore had more need to recover from them. Problems with keystroke data entry (8.6% of the problems; 16 reports from the old and 7 from the young) and with the use of jargon (1.9%, by five old people) were much more troublesome to old than to young participants. Of the eight participants who mistakenly failed to reject a flight for "today" in the United task, seven were senior women.

The last column of Table 16 shows the ratio of old to young reports for each problem listed. This was obtained by dividing the number of reports from old people by the number of reports from young people for each problem. A ratio of "1" indicates that

the problem was reported by an equal number of old and young participants. The higher the ratio, the greater the association between having that problem and being old. The highest ratios and their likely causes are as follows:

Ratio 8.0. Speaking in sentences to the United Airlines robot. These errors suggest that old women were slower than young ones at learning the requirements of this new (voice recognition) system that none of the participants had encountered before.

Ratio 7.0. Failure to reject the United flight for "today." Old people failed to ask for "tomorrow" because of a memory problem: they had forgotten the details of the task.

Ratio 5/0. Problems with jargon: "accessibility." Because of their lifestyles, old people are less exposed to contemporary jargon than young ones.

Ratio 2.3. Keystroke data entry. Old participants were less familiar with this process than students, who use it to register for classes. Older people were also slower at assimilating data entry instructions and acting on them, which triggered timed cut-offs; some also had problems with vision (in looking for the letters) or arthritic fingers.

Ratio 2.0. Too-rapid voices. Several old participants who complained of this commented that they could not think that fast and could not assimilate the choices, and that the more choices there were, the slower the voices should be. Some said it was a hearing problem because when the voice was faster, the words were less distinct.

Ratio 1.2. Long messages/menus/items. Many old participants said that long and monotonous enumerations made it very difficult for them to maintain concentration and that they could not remember the choices described.

DISCUSSION

The goal of this study was to identify the problems old and young people experience when they use IVR systems and to try to determine their cause. In a 2 x 6 mixed factorial design, 22 old and 22 young women performed the same six real-life IVR tasks and answered usability questionnaires after each task. They also answered a questionnaire about their prior experience with technological products and performed a test of attention and short-term memory.

As predicted, old age has a significant negative impact on IVR performance: old participants had fewer successes and obtained lower scores than young participants on all systems but one, for which they obtained identical results. However, these results must be qualified because the performance data for old participants have a bimodal distribution indicating that half of them obtained very poor scores, a third obtained perfect or near perfect scores, and a small number obtained intermediate scores. This polarisation suggests that a substantial proportion of people suffer a deterioration in their abilities and performance as they grow old, but that an almost equally substantial proportion retain their capacities and continue to perform as well or almost as well as young people. The polarisation did not reflect differences between young-old and old-old participants since the performance scores of the twelve participants aged 65 to 74 were not significantly different from those of the ten participants aged 75 to 86.

Old age was also expected to be associated with lower usability ratings on all IVR systems and with a greater degree of effort (measured by length of time and number of keystrokes) in performing successful tasks. This was partially verified as follows: (a) old

participants gave significantly lower usability ratings than young participants to two of the six systems (three in the non-parametric tests) but did not differ on the rest; and (b) old participants needed significantly more keystrokes than young participants to succeed on half of the tasks but did not differ on the other half. There were no significant differences in the length of time old and young participants spent on successful tasks, nor any differences in their usability ratings for successful tasks. These mixed findings reinforce the view that it is misleading to make general statements about the negative effects of old age. Other results showed, as predicted, that young participants had significantly higher scores than old participants on prior experience with technological products and on the attention/memory test, but there was no consistent association between these scores and success at the IVR tasks.

To identify the features of IVR systems that make them more difficult to use, the six systems were compared in terms of the performance scores and usability ratings they produced. With one exception, the results verified the prediction that performance scores would be related to the number of choices/decisions the users had to make to succeed, with the smallest number of choices producing the highest scores. The exception was the voice-activated United Airlines which, despite having the second-to-largest number of choices, produced the highest performance scores among young people and the second-to-highest scores among old people.

The predictions concerning usability ratings were that among both old and young people, the voice-activated United Airlines and the system with the smallest number of choices (Income Security) would obtain the highest ratings, while the lowest ratings would

go to the system with the largest number of choices (Canadian Airlines). This was partially verified: among old participants, the Income Security system obtained significantly higher ratings than the other five systems, but the other systems did not consistently differ from each other; among young people, United Airlines and Income Security were tied for highest ratings, but the other four systems did not significantly differ from each other. In addition, a significant association was found between the performance scores and usability ratings of old participants, indicating that they gave higher ratings to the IVRs they used most successfully. The lack of a similar relationship for young participants appeared to be due to ceiling effects on their performance scores.

Case studies provided detailed information about the main problems old and young participants encountered in trying to accomplish their IVR tasks. These problems were primarily related to flaws in the IVR systems and included the following: unclear choices or instructions, too-rapid automated voices, too-long introductions/menus/items, difficulties with verbal data entry, lack of clear means of recovering from error, problems with keystroke data entry, caller confusion about the task, and problems understanding jargon. The effects of these difficulties were often compounded. For example, Canadian Airlines had a lethal duo combining confusing instructions that produced errors with no means of recovering from errors. At United Airlines, verbal data entry errors were frequent but could be surmounted through an excellent error recovery system.

Almost all these problems have been recognized by IVR experts and represented violations of the IVR guidelines described in the literature review. These guidelines specify that the language used in IVRs should be simple, clear, unambiguous, concise, and

free of technical jargon unless the users are expected to be familiar with it (Schumacher et al., 1995); that menus should have no more than three levels, with no more than four choices per level (Schumacher et al.); and that callers who are stuck or lost should be given the means to recover by hearing a repetition of the menu or prompt or by being clearly told what they are expected to do next (Stentiford & Popay, 1999). The only problem that was not specifically mentioned in past research but that was identified by both old and young people in this study as their second most frequent difficulty was too-rapid automated voices. Participants specified that rapidity is a relative concept, and that the more choices there are, or the more complex, the slower the voices should be.

The IVR systems used in this study were not originally chosen because they were difficult. It was therefore surprising to find that all six IVRs contained obvious flaws that their designers could have avoided or corrected. This finding confirms the view of IVR experts who believe that a large proportion of the IVR systems in use today were developed too quickly and without sufficient testing (Bajorek, 2000). In this study, it would appear that IVR flaws were the main factor that caused participants to make errors, with personal characteristics playing a secondary role. This impression is reinforced by the lack of consistent association between the participants' IVR performances and their personal characteristics as revealed by their scores on the attention/memory test and their prior technological experience.

Another indication that system flaws played the most important role in creating difficulties for the participants in this study is that almost all the problematic IVR features that were listed above were reported by both young and old people. The most frequent

complaint, which was about unclear choices or instructions, was reported almost as often by young participants as by old ones. Overall, a greater number of old participants than young ones reported IVR flaws, and young participants who complained were not as disturbed by these flaws as their old counterparts. In many cases, it would seem, old and young people had the same problems with the IVR systems, but young people were able to overcome them and to succeed at the tasks while a significant proportion of old people were unable to do so.

This greater negative reaction by old participants has interesting methodological implications. IVR experts have written that "you need few subjects to test for big problems and a larger number of subjects to test for subtle problems" (Schaffer, Sorflaten, & Venkateswaran, 2000). More efficient, it seems, is to test all problems with old subjects. Because of their greater sensitivity, seniors resemble the canaries miners took down the mine shafts because they were more sensitive than humans and gave warning of oxygen shortages in time to get out.

In addition to evaluating IVR use, this study brings additional evidence to the longstanding debates between researchers such as Salthouse (1993, 1994), who mostly associates age with performance decline, and others such as Schaie (1994), who points out the enormous variability in age-related changes and the effectiveness of some intervention measures in preventing decline. The results of this study support Schaie's position by showing one large subgroup of old people who performed very poorly and another almost equally large subgroup who performed as well or almost as well as the young.

On the other hand, this study also supports the findings of Salthouse (1993, 1994) and many others about the decline in mental and physical capacities people generally experience as they age. As noted in the Results section, old participants made some IVR errors or complaints much more frequently than young participants, thus indicating age differences in certain types of skills. For example, seven times more old than young participants accepted the United Airlines offer of a flight number for "today," when the task description had clearly specified "tomorrow," suggesting that a greater proportion of old people have memory difficulties. Other differences in errors and complaints suggested that a greater proportion of old than of young people are slow at learning new tasks, slow at processing information and slow at acting on instructions; have greater difficulty maintaining concentration; have more problems with vision, with hearing, and with fingers made stiff by arthritis; and have inferior levels of experience with contemporary jargon and IVR systems. Design flaws were the primary cause of the difficulties participants encountered in this study, but it was the interaction of these flaws with the diminished capacities of many old people that produced age differences in IVR performance.

The finding that old participants gave higher usability ratings to the IVR systems with which they were most successful supports previous research suggesting that seniors do not have a predisposition for or against new technology, but are inclined instead to judge each new technological product on its own merits. When the Human Rights Commission of Australia held hearings in 1999 on the accessibility of information technologies for older people and people with disabilities, seniors' groups argued that well-designed IVRs have an important role to play in increasing older people's access to

essential services (Committee on Ageing of New South Wales, 2000). Taking banking as an example, they said, many young-old seniors under age 75 use ATMs regularly and enjoy their convenience. Other seniors, older or less healthy, have difficulty entering digits or remembering PIN numbers and feel that banking at a branch counter offers them greater security and convenience. Still other seniors, who are mentally alert but physically frail and unable to get around, are grateful for IVR banking that allows them to retain their financial independence and control by paying their bills and checking their bank balances from their homes (Committee on Ageing of New South Wales).

Many organizations throughout the world that are concerned with the quality of life of older and disabled people argue that these people do not want special products adapted to their particular problems and lesser capacities (Australia, Human Rights and Equal Opportunity Commission, 2000; Gill & Shipley, 2000; Hyppönen, 1999). Instead, they support "universal design," which means building products for everyone that are robust and accommodating and take account of differences in sight, hearing, mobility, speech and cognition: "Universal design helps not only people with disabilities, but also any of us when we're tired, busy, or juggling many tasks" (Francik, 1996). The slogan of these groups is that "Good design for disabled people is good design for everyone" (Gill & Shipley).

Calls for universal design are fully supported by the results of this study, which found that most of the problems IVR users experience are troublesome for both young and old people. If anything, these results understate the problem because many, if not most, users of IVR systems are not people in their twenties, but middle-aged people

whose capacities are somewhere between those of the old and the young. According to the Finnish authors of the Handbook on Inclusive Design of Telematics Applications (Hyppönen, 1999), it is illogical to base the design of IVRs and similar systems on the capacities of young people when 25% of the population of industrialised countries is composed of older people or people with a disability and when statistics reveal that, because of improved pension benefits, people over 45 years old have become the top consumers in almost every product group in Finland. Hawthorn (2000) points out that the effects of age start to become noticeable from the mid-forties onward, so that aging people are not just another minority group, but an important segment of the population that is growing in all industrialized countries.

Universal design is good for business for other reasons as well. Badly designed and unpopular IVR systems that cause callers to press the "0" button to speak to an operator can cost IVR owners a great deal of money. According to firms that produce IVR systems, "Some [owners] have indicated that every second they save on the average length of a call means a savings of \$120,000 a year. Where CSR (customer service representative) employees number in the thousands, savings per second have been reported at a million dollars a year" (Schaffer et al., 2000). Almost all the seniors in this study, and about a quarter of the young participants, said they always or almost always press the "0" button to speak to a live operator.

This study was the first to analyse in detail the problems young and old IVR users experience with these systems. Many more studies, using different IVR systems and participants of different ages and of both sexes, are needed to confirm the conclusions of

this research and to obtain a fuller understanding of the problems involved and the measures needed to correct them. Because the six IVR systems I used were not a random sample of all the IVR systems in existence, the results of this study cannot be generalized to all IVR systems. Another drawback of this study is that the senior participants gave the impression of being healthier, more capable and more educated than other community-dwelling seniors. More than half of the senior women who originally expressed interest in participating withdrew upon learning that the subject was IVR systems, saying that they feared and loathed these systems and would have nothing to do with them. Another flaw in this study is that the dependent variables "length of time" and "keystrokes" do not appear to have been successful proxies for participants' degree of effort. Part of the problem was that longer times and a larger number of keystrokes also reflected the cautious nature of some old and young participants, who took advantage of the repeat options to listen to the answers more than once.

IVR systems are here to stay. Most younger people have become used to them and even prefer them for many types of transactions (Shelton, 2000). A recent study found that when people who called an IVR system were transferred to a live operator, half of them hung up (Noonan, 2000). The IVR technology of the future, speech recognition of natural language (McTear, 2000), is potentially very superior to the old keystroke data entry system that seniors hate. Much research still needs to be done to find out how automated telephone-answering systems can be made to serve human beings instead of the other way around.

REFERENCES

- Australia, Human Rights and Equal Opportunity Commission. (2000). Accessibility of electronic commerce and new service and information technologies for older Australians and people with a disability. Report to the Attorney-General. Internet: www.humanrights.gov.au/disability_rights/ecomrep.htm
- Baddeley, A. (1994). Working memory: The interface between memory and cognition. In D.L. Schacter & E. Tulving, (Eds.), Memory systems 1994. (pp. 351-68). Cambridge, MA: MIT Press.
- Bajorek, C. (2000). Back to IVR Basics. Ask Dr. CT. Internet: www.ct-labs.com
- Balota, D.A., & Duchek, J.M. (1988). Age-related differences in lexical access, spreading activation and simple pronunciation. Psychology and Aging, 3, 84-93 .
- Barr, R.A., & Giambra, L.M. (1990). Age-related decrement in auditory selective attention. Psychology and Aging, 5, 597-599.
- Beswick, R., & Reinsch, N. (1987). Attitudinal responses to voice mail. Journal of Business Communication, 24, 23-35.
- Blanchard, H.E., & Lewis, S.H. (1999). Voice messaging user interface. In D. Gardner-Bonneau, Ed., Human factors and voice interactive systems. (pp. 257-284). Boston: Kluwer.
- Boettinger, H.M. (1977). Our Sixth-and-a-half sense. In I. de Sola Pool (Ed.), The social impact of the telephone. (pp. 200-207). Cambridge, MA: MIT Press.
- British Telecommunications. (1998). Now you're talking... Internet: www.labs.bt.com/projects/voice/talk/index/htm
- Bunce, D.J., Morris, I., & Barrowclough, A. (1996). The moderating influence of physical fitness on age gradients in vigilance and serial choice responding tasks. Psychology and Aging, 11, 671-682.
- Charness, N., & Bosman E.A. (1994). Age-related changes in perceptual and psychomotor performance: Implications for engineering design. Experimental Aging Research, 20, 45-59.
- Cohen, M. (1999) Common dialogue tasks examples. Mission: Telephone Speech Standards Committee. Internet: American Voice I/O Society (AVIOS).

Committee on Ageing of New South Wales. (2000). Submission to the Human Rights and Equal Opportunity Commission on accessibility of electronic commerce and new service and information technologies for older Australians and people with a disability. Internet: www.hreoc.gov.au/cgi-bin/textonly.cgi/disability_right.../COA_Ecommerce_subm.ht

Communications International. (1993) VM gets the message, 20, 14-15.

Cook, R. (2000). How *not* to run a voicemail system. Microtimes Magazine, 207. Internet: www.microtimes.com

Craik, F.I.M., & Jennings, J.M. (1992). Human memory. In F.I.M. Craik & T.A. Salthouse, Eds., The handbook of aging and cognition. (pp. 51-110). Hillsdale, NJ: Erlbaum.

Cross, B.A., Hayhow, D.V., Postle, K.J., & Bloomfield, M.H. (1994). Interactive voice, facsimile and TouchTone dialogues. BT Technology Journal, 12, 26-33.

Cross, R. (1995, June 1). Broadcast radio tunes in to listeners. Direct Marketing. Internet: www.crossworldnetwork.com

Czaja, S.J. (1996). Aging and the acquisition of computer skills. In W.A. Rogers, A.D. Fisk, & N. Walker, Eds., Aging and skilled performance. (pp. 201-220). Mahwah, NJ: Erlbaum.

Czaja, S.J., Guerrier, J.H., Nair, S.N., & Landauer, T.K. (1993). Computer communication as an aid to independence for older adults. Behaviour & Information Technology, 12, 197-207.

Czaja, S.J., & Sharit, J. (1998). Age differences in attitudes toward computers. Journal of Gerontology, 53B, P329-P340.

Daneman, M., & Carpenter, P.A. (1980). Individual differences in working memory and reading. Journal of Verbal Learning and Verbal Behavior, 19, 450-466.

Danowski, J.A., & Sacks, W. (1980). Computer communication and the elderly. Experimental Aging Research, 6, 125-135.

de Sola Pool, I., Ed. (1977). The social impact of the telephone. Cambridge, MA: MIT Press.

Dutton, R.T., Foster, J.C., & Jack, M.A. (1999). Please mind the doors - Do interface metaphors improve the usability of voice response services? BT Technology Journal, 17, 172-177.

Dutton, W.H., & Meadow, R.G. (1987). A tolerance for surveillance: American public opinion concerning privacy and civil liberties. In K.B. Levitan (Ed.), Government infostructures. (pp. 147-170). New York: Greenwood Press.

Dyck, J.L., & Smither, J.A. (1994). Age differences in computer anxiety: The role of computer experience, gender and education. Journal of Educational Computing Research, 10, 239-247.

Edwards, R., & Engelhardt, K.G. (1989). Microprocessor-based innovations and older individuals: AARP survey results and their implications for service robotics. International Journal of Technology and Aging, 2, 56-76.

Feldman, R.M., & Reger, S.N. (1967). Relations among hearing, reaction time and age. Journal of Speech and Hearing Research, 10, 479-495.

Fitzgibbons, P.J., & Gordon-Salant, S. (1999). Profile of auditory temporal processing in older listeners. Journal of Speech, Language and Hearing Research, 42, 300-311.

Francik, E. (1996). Telephone interfaces: Universal design filters. Internet: [//park.org/Guests/Trace/pavilion/pbfilter.htm](http://park.org/Guests/Trace/pavilion/pbfilter.htm)

Gardner-Bonneau, D. (1999). Guidelines for speech-enabled IVR application design. In D. Gardner-Bonneau, Ed., Human factors and voice interactive systems. (pp. 146-162). Boston: Kluwer.

Gardner-Bonneau, D.J. (1992). Human factors problems in interactive voice response (IVR) applications: Do we need a guideline/standard? Proceedings of the Human Factors Society 36th Annual Meeting - 1992. 222-226.

Gartner Group. (1998, January-February). Now entering voice mail hell. Marketing Tools. Internet: www.marketingtools.com

Gick, M.L., Craik, F.I.M., & Morris, R.G. (1988). Task complexity and age differences in working memory. Memory and Cognition, 16, 353-61.

Gill, J., & Shipley, T. (2000). Telephones - What features do disabled people need? Internet: www.tiresias.org/phoneability/telephones/

Hawthorn, D. (2000). Possible implications of aging for interface designers. Interacting with Computers, 12, 507-528.

Henderson, R., Deane, F., Barrelle, K., & Mahar, D. (1995). Computer anxiety: correlates, norms and problem definition in health care and banking employees using the Computer Attitude Scale. Interacting with Computers, 7, 181-193.

Howard, J.H., & Howard D.V. (1997). Learning and memory. In A.D. Fisk & W.A. Rogers, Eds., Handbook of human factors and the older adult. (pp. 7-26). New York: Academic Press.

Hyppönen, H., Ed. (1999). Handbook on Inclusive Design of Telematics Applications. Internet: www.stakes.fi/include/handbook.htm

Jay, G.M., & Willis, S.L. (1992). Influence of direct computer experience on older adults' attitudes toward computers. Journal of Gerontology: Psychological Sciences, 47, 250-257.

Katz, J., Aspden, P., & Reich, W.A. (1997). Public attitudes toward voice-based electronic messaging technologies in the United States: A national survey of opinions about voice response units and telephone answering machines. Behaviour & Information Technology, 16, 125-144.

Kausler, D.H. (1994). Learning and memory in normal aging. New York: Academic Press.

Kline, D.W., & Scialfa, C.T., (1997). Sensory and perceptual functioning: Basic research and human factors implications. In A.D. Fisk & W.A. Rogers, Eds., Handbook of human factors and the older adult. pp. 27-54. New York: Academic Press.

Kloosterman, S.H. (1994). Design and implementation of a user-oriented speech recognition interface: the synergy of technology and human factors. Interacting with Computers, 6, 41-60.

Kraus, S.J. (1995). Attitudes and the prediction of behavior: A meta-analysis of the empirical literature. Personality and Social Psychology Bulletin, 21, 58-75.

Lehto, J. (1996). Are executive function tests dependent on working memory capacity? Quarterly Journal of Experimental Psychology, 49A(1), 29-50.

Lemme, B.H. (1995). Development in adulthood. Boston: Allyn and Bacon.

Lennig, M., Bielby, G., & Massicotte, J. (1995). Directory assistance automation in Bell Canada: Trial results. Speech Communication, 17, 224-234.

Lindquist, C. (2000, June 15). Phones still don't listen. CIO Magazine. Internet: www.cio.com

Li-Ron, Y. (2000, September 18). The new Web: no strings attached (Special Advertising Section). Newsweek.

Lockey, M. (2000, August). Listen up! Hear e-mail and web information by phone. PC World Magazine. Internet: www.pcworld.com

Marx, G.T. (1994). New telecommunications technologies require new manners. Telecommunications Policy, 18, 538-551.

McAteer, J.F. (1997). New directions in interactive voice response: Emerging opportunities for speech recognition technology. Business Intelligence Program Abstract. Internet: Business Intelligence Program.

McTear, M.F. (2000). Spoken dialogue technology: Enabling the conversational user interface. Internet: www.infj.ulst.ac.uk/~cbd23/survey/spoken_dialogue_technology.html

Meeks, L.L. (1992). Selecting an interactive voice response system. In Proceedings of the 11th annual conference of the American Voice I/O Society, San Jose, California, 223-227.

Morrell, R.W., & Echt, K.V. (1996). Instructional design for older computer users: The influence of cognitive factors. In W.A. Rogers, A.D. Fisk & N. Walker, Eds., Aging and skilled performance: Advances in theory and applications. (pp. 241-265). Mahwah, NJ: Erlbaum.

Morris, R.G., Gick, M.L., & Craik, F.I.M. (1988). Processing resources and age differences in working memory. Memory and Cognition, 16, 362-66.

Myerson, J., & Hale, S. (1993). General slowing and age invariance in cognitive processing: The other side of the coin. In J. Cerella, J., J. Rybash, W. Hoyer, & M.L. Commons, Eds., Adult information processing: Limits on loss. (pp. 115-141). New York: Academic Press.

Nabelek, A., Letowski, T.R., & Tucker, F.M. (1989). Reverberant overlap and self-masking in consonant identification. Journal of the Acoustical Society of America, 86, 1259-1265.

National Council of Welfare. (1990). Women and poverty revisited. Ottawa.

Noonan, T. (1998). Building user-friendly voice systems. Internet: www.softspeak.com.au

Ollari, R.J. (1997). Over one million served: The quiet voice of service. AS 400. Internet: Missing Link NAS, Inc.

Rogers, W.A., Cabrera, E.F., Walker, N., Gilbert, D.K., & Fisk, A.D. (1996). A survey of automatic teller machine usage across the adult life span. Human Factors, 38, 156-166.

Rogers, W.A., Meyer, B., Walker, N., & Fisk, A.D. (1998). Functional limitations to daily living tasks in the aged: A focus group analysis. Human Factors, 40, 111-125.

Salthouse, T.A. (1994). The nature of the influence of speed on adult age differences in cognition. Developmental Psychology, 30, 240-259.

Salthouse, T.A. (1993). Speed mediation of adult age differences in cognition. Developmental Psychology, 29, 722-738.

Salthouse, T.A., Mitchell, D.R.D., Skovronek, E., & Babcock, R. (1989). Effects of adult age and working memory on reasoning and spatial abilities. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 507-16.

Schaffer, E., Sorflaten, J., Venkateswaran, M., & Miracle, G. (2000). Interactive voice response (IVR) case study: Testing your telephone-based e-commerce support. Human Factors, 12.2. Internet: www.humanfactors.com/library/ivr.asp

Schaie, K.W. (1996). Intellectual development in adulthood: The Seattle Longitudinal Study. New York: Cambridge University Press.

Schaie, K.W. (1994). The course of adult development. American Psychologist, 49, 304-313.

Schmandt, C. (1994). Voice communication with computers. New York: Van Nostrand Reinhold.

Schumacher, R.M., Jr., Hardzinski, M.L., & Schwartz, A.L. (1995). Increasing the usability of interactive voice response systems: Research and guidelines for phone-based interfaces. Human Factors, 37, 251-264.

Schwartz, J. (1988). The computer market. American Demographics, 10, 38-41.

Settle, R.B., Dillon, T.W., & Alreck, P.L. (1999). Acceptance of the phone-based interface for automated call direction. Behaviour & Information Technology, 18, 97-107.

Shelton, K.M. (2000). Twenty-first century interactive voice response. Internet: www.tmcnet.com/tmcnet/articles/ts0700.htm

Slawinski, E.B., Hartel, D.M., & Kline, D.W. (1993). Self-reported hearing problems in daily life throughout adulthood. Psychology and Aging, 8, 552-561.

Smith, A.D. (1996). Memory. In J.E. Birren & K. Warner Schaie, Eds., Handbook of the psychology of aging, 4th edition (pp. 236-250). New York: Academic Press.

Smither, J.A. (1993). Short term memory demands in processing synthetic speech by old and young adults. Behaviour & Information Technology, 12, 330-335.

Society of Consumer Affairs Professionals in Business (SOCAP). 1998. Press release: Consumer acceptance of automated telephone systems is mixed; Can lead to hang ups, new report finds. Internet: www.socap.org/press/ivrstudy.html

Stentiford, F.W.M., & Popay, P.A. (1999). The design and evaluation of dialogues for interactive voice response services. BT Technology Journal, 17(1), 142-148.

Thorngate, W. (1992). Evidential statistics and the analysis of developmental patterns. In J.B. Asendorpf & J. Valsiner, Eds., Stability and change in development: A study of methodological reasoning. (pp. 63-83). London: Sage.

Turner, A. (2000, February). The less obvious business case for call centres. Business Connect. Internet: www.lucentbcsuk.co.uk.

Vercruyssen, M. (1997). Movement control and speed of behavior. In A.D. Fisk & W.A. Rogers, Eds., Handbook of human factors and the older adult. (pp. 55-86). New York: Academic Press.

Verhaeghen, P., Vandenbroucke, A., & Dierckx, V. (1998). Growing slower and less accurate: Adult age differences in time-accuracy functions for recall and recognition from episodic memory. Experimental Aging Research, 24, 3-19.

Virzi, R.A., & Huitema, J.S. (1997). Telephone-based menus: Evidence that broader is better than deeper. Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting - 1997. 315-319.

Walker, N., Philbin, D.A., & Spruell, C. (1996). The use of signal detection theory in research on age-related differences in movement control. In W.A. Rogers, A.D. Fisk, & N. Walker, Eds., Aging and skilled performance: Advances in theory and applications. (pp. 45-64). Mahwah, NJ: Erlbaum.

Whittaker, S.J., & Attwater, D.J. (1996). Interactive speech systems for telecommunications applications. BT Technology journal, 14(2) 11-22.

Wingfield, A., Stine, E.L., Lahar, C.J., & Aberdeen, J.S. (1988). Does the capacity of working memory change with age? Experimental Aging Research, 14, 103-107.

Zandri, E., & Charness, N. (1989). Training older and younger adults to use software. Educational Gerontology, 15, 615-631.

Zeithaml, V.A., & Gilly, M.C. (1987). Characteristics affecting the acceptance of retailing technologies: A comparison of elderly and nonelderly consumers. Journal of Retailing, 63, 49-68.

APPENDIX A

Informed Consent Form

Project: Study of the usability of automated telephone response systems in real-life tasks performed by older and younger women

Researcher: Louise Dulude
Department of Psychology
Carleton University
Tel.: 746-7668

Faculty sponsor: Dr. Warren Thorngate
Department of Psychology
Carleton University
Tel. 520-2600, ext. 1403

The purpose of this informed consent form is to ensure that you understand the objectives of the study and the nature of your involvement. The informed consent must provide you with enough information that you have the opportunity to determine whether you wish to participate in the study. Please ask the researcher to clarify any concerns you may have.

I (the participant) consent to participate in a study about automated telephone response systems. The purpose of this study is to try to find out if older people have more difficulty with these systems than younger people and to identify the features of the systems that make them more difficult to use. Testing will take place at Carleton University, will last from 50 to 80 minutes, and will involve the following:

- 1) providing my birth date;
- 2) doing a quick telephone test to identify problems such as difficulties with hearing or eyesight that could make it harder to accomplish the main telephone tasks (described in 5 below);
- 3) completing a questionnaire on my experience with technological products (such as microwave ovens and VCRs);
- 4) doing a test of attention and memory;
- 5) performing six real-life tasks using automated telephone response systems; and
- 6) filling out a short questionnaire after each real-life task to indicate my impressions of the telephone system I just used.

I realize that the telephone tasks will be recorded with an audio tape recorder in order to analyse my responses. None of these tests should pose any risks or cause me any discomfort. I understand that I do not have to answer any questions I do not wish to answer and that I may withdraw from this study any time I want. I further understand that the information I provide, the audio tapes and the results of my tests will be coded anonymously, will be kept confidential, and will be used for research purposes only.

- 2 -

If I want more information about this study, I can contact the researcher or faculty sponsor named above. If I have any ethical concerns about the study, I can contact one of the following:

- Chair, Department of Psychology: Dr. K. Matheson, tel. 520-2600, ext. 2648
- Chair, Ethics Committee: Dr. M. Gick, tel. 520-2600, ext. 2664.

Subject:

Researcher:

Printed name: _____

Printed name: _____

Signature: _____

Signature: _____

APPENDIX B

Information/Debriefing Form

Thank you very much for participating in this study, whose purpose was to try to identify the features of automated telephone systems that make them easier or more difficult to use, and to determine whether these features are different for older and younger people. Very little research has been done on this subject, and many people who are obliged to use these systems feel frustrated and dissatisfied. We hope that our study will help remedy the problem by indicating ways in which these systems can be improved and made more acceptable to their users.

All the information you have given us in the study will be kept confidential. Feel free to contact us at any time if you have any questions about the study. As already mentioned to you, we will be sending you a brief summary of the results as soon as they become available. If there is anything you wish to find out before or after you receive this summary, do not hesitate to contact any of the people listed below.

For questions or comments about this research:

- Researcher: Louise Dulude, tel. 746-7668
- Faculty sponsor: Dr. Warren Thorngate, tel. 520-2600, ext. 1403

If you have any ethical concerns about this research:

- Dr. M. Gick, Chair, Department of Psychology Ethics Committee, tel. 520-2600, ext. 2664, or
- Dr. Kim Matheson, Chair, Department of Psychology, tel. 520-2600, ext. 2648.

You can also write to Louise Dulude or any of the people named above at the Department of Psychology, Carleton University, Ottawa, Ontario K1S 5B6

Thanks again for your contribution. It was greatly appreciated.

APPENDIX C

**Questionnaire on Prior Experience
with Technological Products**

Note: This questionnaire was administered verbally in an informal manner which made participants feel relaxed and at ease.

Instructions: Various technological products are listed below. For each one, indicate whether you have had experience with it in the past month, how many times, and, if you have ever used it in the past, how you would rate your degree of success in using it on the scale described below.

1) Automated telephone systems encountered in calling stores, businesses:

Have used it in the past month (circle correct answer below)		Number of times used in past month	If ever used in the past, degree of success in using it (0= <u>unsuccessful</u> ; 1= <u>neutral</u> ; 2= <u>successful</u>)
YES	NO	_____	_____

2) Microwave oven:

Have used it in the past month (circle correct answer below)		Number of times used in past month	If ever used in the past, degree of success in using it (0= <u>unsuccessful</u> 1= <u>neutral</u> ; 2= <u>successful</u>)
YES	NO	_____	_____

And so on for each of the following: automated bank teller machine (ATM), debit card scanner, home telephone answering machine, video cassette recorder (VCR), video camera, photocopy machine, fax machine, compact disk player, computer, home security system.

(Scoring: Add together 1 point for YES or 0 for No, the actual number of times used in the past week up to a maximum of 5, and the degree of success noted. Minimum for each item: 0. Maximum for each item: 8. Maximum points for all 12 items: 96.)

APPENDIX D

Word Span Test of Attention and Short-Term Memory

Instructions (given verbally): You are going to hear recorded words, presented at a rate of about one word per second. The words are presented in sets, and after each set you will be given the time you need to write down the words you remember, not necessarily in the original order. We will start with two practice sets of two words each, then two sets of three words each, and so on, increasing by one word each time.

Practice 1 - happy, winter

Practice 2 - cherry, danger

1 - novel, recess, actor

2 - silver, prose, beauty

3- spirit, today, money, report

4- action, water, raisin, heaven

5- rabbit, coffee, visit, modest, clinic

6- pasta, story, joker, puzzle, record

7- minor, patron, tiptoe, title, hurry, lizard

8- topic, ruler, lucky, indian, nature, soccer

9- spider, market, adverb, token, settle, rebel, profit

10- vulgar, better, stable, tissue, sandal, gallop, recent

11- party, hollow, glory, pocket, differ, tulip, vocal, sister

12- reason, fellow, letter, penny, cable, summer, vital, pivot

13- sailor, ready, urgent, panel, social, cotton, shiver, pupil, angel

14- neutral, topic, daily, satin, candle, safer, medal, dinner, proper

(Scoring: The score is the total number of correctly recalled words - maximum 84.)

APPENDIX E - STRUCTURE OF IVR SYSTEMS

System 1 - Regional Municipality of Ottawa-Carleton**Regional Municipality of Ottawa-Carleton - 233-4636**

System 2 - Canadian Airlines

Canadian Airlines - 1-888-298-8888

System 3 - Government of Canada Income Security**Income Security Programs - 1-800-277-9914**

Welcome to . . . If you are calling about OAS or CPP benefits, have your S.I.N. # ready, etc. (long introduction).
To learn more about the selections available, please listen to all the short following messages.

If you are currently receiving benefits, Press 1

Recording re time when cheques arrive
0- speak to agent
1- change address
0 - direct deposit change
2- general information
3- application by mail
0- speak to agent
* to repeat
first menu
0- speak to agent

For general information, Press 2

1 - General Information +
(1) more on OAS
(2) more on GIS
(3) more on SPA
(4) more on WSPA
* to repeat
return to first menu
(0) agent
2- General Information on CPP +
(1) info retirement
(2) info disability
(3) info contributions
* to repeat
return to first menu
(0) agent
(6) to apply (Voice Box)
3- Application kit (Voice Box)
0- speak to agent

To speak to an agent, Press 0
To repeat, Press *

System 4 - Statistics Canada

Statistics Canada

1-800-263-1136 or 951-8116

System 5 - AccuRate Foreign Exchange

AccuRate - 596-0612

System 6 - United Airlines

United Airlines - 1-888-824-6200

APPENDIX F

Instructions for the IVR Tasks
(read by the participants from index cards)

System 1 - Your cousin is unemployed and you are trying to help him. To find out if any jobs are available at the Regional Municipality of Ottawa-Carleton, call 233-4636 and note down the answer.

System 2 - You want to visit a friend who lives in Halifax and think tomorrow would be a good time for it. Call Canadian Airlines at 1-888-298-8888 to find out if there is a flight from Ottawa to Halifax around 5:00 P.M. tomorrow afternoon. Note down the flight number and departure time.

System 3 - You are an old-age pensioner who normally receives her cheque by mail on the same date every month. But this month it did not arrive on time and is now four days late. Call 1-800-277-9914 to find out what you should do now.

System 4 - You want to know how many people lived in Quebec and Ontario at the time of the 1996 census. Call Statistics Canada at 1-800-263-1136 to find out the answers. Note them down.

System 5 - Accu-Rate is a private company that deals in foreign monies. Phone them at 596-0612 to find out how many Chilean pesos you can buy with one Canadian dollar. Note down the answer.

System 6 - A friend who lives in New York wants to see you right away and offers to pay your way there. You decide to leave by plane from Toronto tomorrow. You call United Airlines at 1-800-824-6200 to find out if they have a flight from Toronto to New York (Kennedy airport) around 6:00 P.M tomorrow evening. Note down the flight number and departure time.

APPENDIX G

Usability Questionnaire
(administered verbally)

On the scale from one to five shown below, indicate if you agree or disagree with the ten statements on this page.

1	2	3	4	5
strongly disagree	disagree	neutral	agree	strongly agree

Rating 1-5

1. I found this service easy to use..... _____
2. I felt very frustrated using this service..... _____
3. I felt in control while using this service..... _____
4. I thought some of the choices were confusing..... _____
5. The operator's voice was very clear..... _____
6. I am left with a good impression of that organization... _____
7. I had to concentrate really hard in doing this task..... _____
8. I sometimes felt lost while using this service..... _____
9. I enjoyed using this service..... _____
10. I thought the operator spoke too quickly..... _____

Complementary questions:

Had you ever used this system or a similar one before? How many times?

What was going on in your head when you were doing the task?

Do you have any other comments you would like to add about this task?