

RICE UNIVERSITY

**The Impact of Voice Characteristics on User Response in an
Interactive Voice Response System**

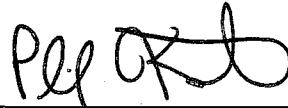
by

Rochelle E. Evans

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Master of Arts

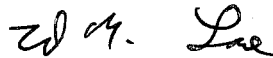
APPROVED, THESIS COMMITTEE:



Philip Kortum, Professor in the Practice,
Chair, Psychology



Michael Byrne, Associate Professor,
Psychology



David M. Lane, Associate Professor,
Psychology, Statistics, and Management

HOUSTON, TEXAS

MAY 2009

UMI Number: 1466775

INFORMATION TO USERS

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 bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send 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.

UMI[®]

UMI Microform 1466775
Copyright 2009 by ProQuest LLC
All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

The Impact of Voice Characteristics on User Response in an Interactive Voice Response System

by

Rochelle Evans

System voice within interactive voice response systems (IVRs) was investigated. Specifically, users were randomly assigned a system voice personality (upbeat, professional, and sympathetic) and voice gender (male and female) when completing a health survey over IVR. Disclosure rates were not affected by the type of voice heard, nor did they differ by user gender. Additionally, disclosure was higher on the IVR version of the health survey than on a web-based version, further recognizing the privacy offered by IVRs.

Acknowledgements

I would like to thank sincerely Dr. Philip Kortum for taking me in as one of his advisees. He has not only been a great teacher, instilling in me the fundamentals of human factors design, but a truly fantastic mentor. Additionally, I want to thank my committee members, Dr. David Lane and Dr. Michael Byrne, for their thoughtful feedback and for their assistance not only throughout the revision process, but during my tenure at Rice.

Finally, I would like to dedicate this thesis to my husband, Corry Edwards, for his unwavering support.

TABLE OF CONTENTS

LIST OF TABLES.....	vi
LIST OF FIGURES.....	vii
LIST OF APPENDICES.....	viii
1.0 INTRODUCTION.....	1
1.1 IVRs in the Medical Community.....	1
1.2 Accuracy.....	4
1.3 Privacy.....	5
1.4 Disclosure.....	7
1.5 Persona and Social Interface Theory.....	10
1.6 The M.D. Anderson Symptom Inventory.....	12
1.7 The Current Study.....	13
2.0 EXPERIMENT 1.....	15
2.1 Stimulus Creation.....	16
Participants.....	16
Materials.....	16
Procedure.....	17
Preliminary Selection.....	18
2.2 Stimulus Selection.....	18
Participants.....	18
Materials.....	18
Procedure.....	19
2.3 Results & Discussion of Experiment 1.....	20
Graphical Depiction of Voices.....	21
ANOVAs Examining Within-Voice Differences.....	21
Contrasts Examining Within-Voice Differences.....	23
Final Selection of a Male and Female Voice.....	26
Trust & Liking of Voices.....	29
The Current MDASI-IVR Voice.....	32
Experiment 1 Discussion.....	32
3.0 EXPERIMENT 2.....	34

3.1 Measurement of Response Differences.....	35
Participants.....	35
Materials.....	35
Procedure.....	38
3.2 Results & Discussion of Experiment 2.....	41
Between-Voice Comparisons.....	42
Scoring of the MDASI-IVR & CDC-HRQOL.....	43
Accuracy of Response on the CDC-HRQOL.....	46
Reliability of the MDASI-IVR.....	48
Usability of the MDASI-IVR.....	50
Experiment 2 Discussion.....	51
4.0 GENERAL DISCUSSION.....	52
REFERENCES.....	55

LIST OF TABLES

Table 1: ANOVAs for voices being considered after initial elimination.....	24
Table 2: Correlations between M21's trust and like ratings.....	30
Table 3: Confidence Intervals and Significant Tests for User Gender x Voice Gender x Voice Type.....	43
Table 4: Comparison of the current sample's CDC-HRQOL scores to state- and nation-wide averages.....	47
Table 5: Comparison of the current sample's CDC-HRQOL scores to Zullig (2005).	48

LIST OF FIGURES

Figure 1. Bar graphs comparing user ratings of dominant items (i.e., upbeat rating for upbeat voice) for male and female voices.....	22
Figure 2 Bar graphs demonstrating user ratings of the selected male and female voices.....	28
Figure 3 Ratings of trust and liking by rater gender for F19.....	31
Figure 4 Ratings of trust and liking by rater gender for M21.....	33
Figure 5 Rating for the current MDASI-IVR's voice.....	34
Figure 6 Box plots of ratings on the MDASI-IVR for all voices presented.....	41
Figure 7 Rating comparison of CDC HRQOL, MDASI-IVR and web-based MDASI.....	45
Figure 8 Comparison of ratings on MDASI depending on version (IVR or web-based) and presentation (IVR given first or second).....	49

LIST OF APPENDICES

Appendix A: M.D. Anderson Symptom Inventory (MDASI) Core Items.....	63
Appendix B: Sample Script.....	64
Appendix C: Personality Scale.....	65
Appendix D: MDASI-IVR Script.....	66
Appendix E: System Usability Scale.....	69
Appendix F: CDC-HRQOL-9.....	70

The Impact of Voice Characteristics on User Response in an Interactive Voice Response System

1.0 Introduction

Interactive voice response systems (IVRs) are the interfaces that step in for a live operator or telephone attendant to route a user through a company's telephony system. One might be familiar with such phrases as, "for English, press 1," or "please enter your 9-digit social security number now." The IVR interface first entered the market in the mid-1970's (Witten & Madams, 1977). The first IVRs used Dual Tone Multi-Frequency (DTMF) entry whereby the telephone's touch-tone keypad was used for information entry. As IVRs have been prevalent in telephony for over thirty years, anyone calling a company for routine billing or customer support has certainly crossed paths with an IVR. By 1993, receptionists were no longer the front line, as 97% of large corporations in the United States turned to IVRs for routing incoming calls (Moeller & Bort, 1993). However, widespread usage of an interface does not guarantee user satisfaction. While 60% of users could access the desired information or person via the IVR, 70% found it hard to describe what they were looking for and believed there were too many irrelevant options that had to be heard (Katz, Aspden, & Reich, 1997). However, with improvements in IVR design, user satisfaction has increased (Bushey, Martin, & Joseph, 2001).

1.1 IVRs in the Medical Community

IVRs span from voicemail to consumer 1-800 numbers. They have alternative uses beyond managing call flow or facilitating movement through menus. For example, the DTMF keypad of a touchtone phone is well-suited for data entry, particularly when

the entry can be limited to digits from 0-9. Members of the medical community have taken advantage of this functionality, converting what would be paper-and-pencil questionnaires or in-person interviews to those that users can complete away from the doctor's office. IVRs appear in many areas of medicine. Simpson, Kivlahan, Bush, and McFall (2005) utilized an IVR protocol for recovering alcoholics. These individuals called the system daily, weekly, or monthly to fill out a behavioral survey. Wang et al. (2002) had employees who utilized a particular insurance group fill out a survey via IVR to collect information about health risks. This method was found to be cost-effective due to the removal of an interviewer from the process. Indeed, compared to computer-assisted telephone interviews (CATIs), where an interviewer follows an automated script from a computer and enters responses into the system which then prompts the next script to be read, IVRs proved to be more cost-effective (Corkrey & Parkinson, 2002). The aforementioned studies utilized call-in access whereby the users were given a number and called the system. IVR systems can also be designed as call-out where the system contacts the users. For example, Cleeland et al. (2000) validated a survey which is also used via IVR to collect daily information on a cancer patient's symptom levels. The IVR sends its results to the physician on call, who can be alerted to individual responses (or changes in responses from each time a user fills it out) above a certain threshold. Tanke and Leirer (1994) utilized an IVR to send out reminders for patients with an upcoming appointment at a tuberculosis clinic. These automated reminders increased patient attendance by 10%. There is also a voice user interface (VUI) being tested that contacted patients a day after their discharge from outpatient surgery (Forster et al., 2008). This

study not only demonstrated that collecting adverse events via the VUI is feasible, but that patients appreciated that the call was automated.

The IVR saves time for doctors or nurses who may otherwise need to call patients at home to gather patient information. The physician can also design the IVR in multiple languages or dialects to accommodate his or her patient population (Abu-Hasaballah, James, & Aseltine Jr., 2007). This would remove complications associated with misunderstanding the patient. In addition to the benefits on the physicians' side, there are several for the patient as well. For example, if the user does not have to rely on a call from the physician, he can instead call the IVR or schedule times for the system to call him at his convenience. These calls can be made and received around the clock which gives the user more freedom than initially provided. Patients also minimize trips to the doctor's office, using the IVR to give the physician information that may have required an in-clinic appointment. IVRs are also beneficial for patients with a lower education status who may not be highly literate (Abu-Hasaballah et al.)

On the downside, IVR response rate is not 100%, and rates typically dwindle when repeated collection is necessary. For example, Agel, Rockwood, Mundt, Greist, and Swiontkowski (2001) had patients fill out a paper-and-pencil survey in an orthopedic clinic and upon leaving, patients were given a card with the number for the IVR version of the survey, which they were to complete within a week. Compliance for the IVR version was a dismal 49%. Simpson et al. (2005) offered a very small incentive (\$0.50) per call made to alcoholics in the sample, who were reporting alcohol use and cravings. IVR compliance over one month was not as low as the previous study where no incentive was offered. Individuals calling daily completed 88.9% of calls and those calling weekly

completed 70.4% of calls. A lower percentage was recorded for users who were considered homeless or living in shelters. Also, while physicians may appreciate the time saved by IVRs, patients may not prefer IVRs. This may particularly be the case when the IVR is poorly designed (i.e. Reidel, Tamblyn, Patel, & Huang, 2008).

1.2 Accuracy

To reap the benefits of utilizing an IVR, the data collected by the IVR must be accurate. Specifically, symptoms and behaviors reported via the IVR ought to maintain high convergent validity with information that would have been collected via the physician or paper-and-pencil/web-based inventories utilized prior to the intervention of the IVR technology. Alemagno, Frank, Mosavel, and Butts (1998) examined IVR-collected data from adolescents between the ages of 11 and 18 being screened for risky behavior such as suicidal tendencies, alcohol and drug use, and depression. These adolescents completed the IVR which utilized “1” for yes and “2” for no. Twenty-two physicians then examined the responses to determine if they believed the adolescents had overreported, underreported, or accurately reported risky behaviors. They strongly agreed with adolescents’ ratings, giving the system an accuracy rating of 84.2%. It was believed that depression may have been slightly overestimated, with agreement for this subscale dropping to 72.8%. In another example, the Short Musculoskeletal Function Assessment (SMFA), a 46-question inventory assessing musculoskeletal complaints, was built into an IVR (Agel et al., 2001). Patients completed the SMFA-IVR or the written SMFA at the orthopedic clinic, and then completed the alternate method 3 to 7 days later from home. The two forms took roughly the same amount of time to complete (9.5 minutes for the IVR and 10 minutes for the written SMFA), and there was no reliable difference in

responses between the two forms, regardless of order of administration. Toll, Cooney, McKee, and O'Malley (2005) investigated cigarette usage reported via IVR and compared this data to a timeline follow-back method (TLFB) where a thorough history of smoking history was recorded for the previous month. The TLFB method utilizes a calendar, in which participants record usage over a predefined interval. All 381 participants were heavy smokers who smoked a minimum of 20 cigarettes a day for the past year. The IVR was completed over 7 subsequent days. Across all 7 days, the correlation between the TLFB and IVR remained very high, ranging from .84 to .95 (all p -values less than .05). There was high accuracy between TLFB and IVR ratings, so it did not appear that individuals were over- or under-reporting on the IVR compared to the TLFB method which had already been assessed for reliability and validity as an instrument.

1.3 Privacy

The benefit of privacy is global across all IVR applications – a user inputting a password or social security number via DTMF appreciates the ability to securely enter in such protected information. It is true that a voice user interface (VUI) is also capable of maintaining privacy when information is coded, as the user would utter “one” to represent the answer he wishes to denote (e.g. “If you would like to check your HIV test results, say 1. If you would like to check your hepatitis B results, say 2.”). However, an IVR is a more effective interface at capturing responses due to its lower error rates – in order for a VUI to work to its true potential it would have to accurately comprehend numerous accents and dialects, a challenge for many untrained VUIs (Olive, 1999).

Patient privacy is an added benefit in IVRs as they can be designed to maintain the confidentiality of participants. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) requires subject de-identification. Via the HIPAA Privacy Rule, information that identifies an individual such as date of birth or Social Security number cannot be included unless the data is sufficiently de-identified (United States Department of Health & Human Services, 2003). IVR designers can guarantee de-identified data through using participant IDs and coding additional demographic information numerically. In addition, the privacy that IVRs offer both via the comfort of de-identification and the ability to complete calls in the absence of health care professionals or other people that may make the user uncomfortable is quite beneficial. Frank et al. (1997) found that HIV + patients preferred to receive over-the-phone counseling from an IVR than from a live counselor. Kim, Bracha, and Tipnis (2007) gave the 10-item Edinburgh Postnatal Depression Scale (EPDS) via IVR to 57 disadvantaged pregnant women in order to screen for antenatal depression. Of the 21 women who scored highly on the EPDS, only 5 of the women indicated that they would be willing to speak with a psychiatrist or mental health professional, whereas 10 were willing to speak to their obstetrician or midwife about how they were feeling. The IVR offered privacy that allowed these women to reveal their depression, however following up with a psychiatrist or their current obstetrician would have removed that level of privacy, so fewer were willing to utilize that option.

When an individual faces sensitive questions, he or she is likely to under-report undesirable behaviors and over-report desirable behaviors (Tourangeau & Yan, 2007). For example, people are likely to under-report substance abuse (Fendrich & Vaughn,

1994) and racism (Krysan, 1998). This pattern is likely in situations where privacy is not secured, and in this case an IVR has the ability to increase privacy. Moskowitz (2004) had adolescents between the ages of 12 and 17 complete a survey of smoking behaviors that was interviewer-administered or automated via IVR. Both smoking susceptibility among non-smokers and current smoking rates among smokers were higher for adolescents reporting over the IVR compared to the interviewer-administered survey. In addition, the adolescents' parents were more likely to be present in the interviewer-administered condition, so it can be hypothesized that with a lack of privacy, underreporting of behaviors was occurring. Corkrey and Parkinson (2002) compared responses for Australian participants between computer-assisted telephone interviews (CATI), IVR, and a combination of the two who received the 5-question AUDIT which measures alcohol use and questions on drug use. Response rates for alcohol and marijuana use were much higher for participants who filled out the questionnaires via IVR than CATI or the IVR/CATI combination. This finding indicates that the privacy that IVRs offer do indeed increase response rates for undesirable behaviors that would otherwise be underreported.

1.4 Disclosure

In order to minimize or eliminate the degree to which users under- or over-report information, one ought to be able to take advantage of customization of an IVR's voice beyond making it sound more natural. One difference across the numerous IVRs utilized in the above studies is that the voices used in the IVRs are quite different. Some are males; some, females, and within gender, the voices vary as to what personality they may convey (concern, professionalism, etc.). For example, it was found that gender

stereotypes can influence user perceptions; specifically, a male-voiced computer was rated as better suited for a dominant role than a female voice, and the male voice was also found to be more competent and friendly than its female counterpart (Nass, Moon, Morkes, Kim, & Fog, 1997). Nass et al. also discovered that individuals preferred to interact with a voice that was closer to themselves in dominance or submission, regardless of gender. This factor altered refusal rates in traditional phone interviews when considering vocal aspects such as pitch and tempo (Oksenberg, Coleman, & Cannell, 1986).

IVRs carry an added factor depending on the quality of voice implemented into the system. Designers can utilize text-to-speech (TTS), which converts text to speech. TTS, which may be traditionally considered a very robotic computer voice, varies greatly in quality and some TTS is practically indistinguishable from a natural voice. One can manipulate gender, accent, and many other vocal characteristics. The other voice type that can be implemented in the IVR is a natural voice where a speaker's recorded voice is directly utilized in the IVR. While a natural voice sounds better, it is much less cost-effective than TTS. Thus, the IVR designer is granted many options when selecting the voice for an IVR, and this may in turn affect the user's responses or satisfaction with the system. When examining the distinction between a synthetic (robotic) or natural (human-like) voice, users may not consistently prefer one voice. In a study enquiring about participants' deviant behaviors (sex life, lying, cheating, drug use), the natural-voiced IVR evoked more responses from participants than the synthetic voice (Nass, Robles, Heenan, Bienstock, & Trienen, 2003). An increase in unanswered questions was tied to a decrease in disclosure, so there was more disclosure when using the natural-voiced IVR.

However, Couper, Singer, and Tourangeau (2004) discovered that while users were able to distinguish accurately between the voices, they did not disclose more information to a natural voice IVR, human-like TTS IVR, or a machine-like TTS IVR. While there was no difference between the three types of IVRs, users' responses averaged across the IVR types displayed greater disclosure when compared to CATI.

It is not clear whether a particular gender elicits the most disclosure. Tannen (1996) holds that there is a preference to disclose information to a female voice. However, within the female voice, there are several vocal attributes that can alter how it is perceived. For example, a female voice that was lower in pitch was rated as less emotional, yet more mature (Aronovitch, 1976). Preferences for the male gender can exist, as well. Lines and Hone (2002) found that individuals over the age of 65 preferred a male voice, referring to it as more pleasant, intelligent, natural, clear, and soothing. However, these individuals were not listening to this voice for the purpose of disclosure. When individuals were asked to disclose personal information such as "have you ever taken part in sexually-deviant behavior," those who listened to a natural-voiced IVR in the experiment by Nass et al. (2003) were more likely to disclose to a natural-voiced female than a male. Tourangeau, Couper, and Steiger (2003) found mixed results in terms of gender and disclosure, which depended on whether or not a live interviewer asked demographics questions before switching the participants over to the IVR. When these innocuous questions were asked first, there was more disclosure on the IVR when a male voice was heard. However, when the questions were not asked by the live interviewer, but added to the IVR, there was more disclosure when a female voice was heard over the IVR. This study exemplifies the complexities in understanding the effects of voice

qualities on disclosure. Additionally, it can be argued that it is not the gender behind the voice, but the communication skills relayed by the male or female speaker, that may drive gender differences in disclosure or satisfaction. For example, Christen, Alder, and Bitzer (2008) discovered that patients did not prefer a male or female physician, but a patient-centered communication style, that just happened to be more frequently conveyed by female physicians. When controlling communication style across gender with an IVR, a preference for a particular gender may not exist.

1.5 Persona and Social Interface Theory

IVR research investigating differences in disclosure or preferences for different voices have not expanded beyond gender, and the aforementioned studies fail to delineate the tone of the voices. For example, it is possible that in Lines and Hones (2002) the female voice was exceptionally high-pitched and the male voice was a more comfortable pitch to listen to, hence the preference for the male voice. No studies have focused on IVR voices along this type of dimension to examine the differences reported in subjective preferences or responses between personae. A persona is defined as the “personality of a speech interface inferred by users based on the behavior of the VUI” (Hura, 2008). This dimension is crucial, as speaker type and form that ought to be utilized would vary depending on the IVR’s application (Stentiford & Popay, 1999; Reeves & Nass, 1997).

The voice heard over the receiver affects a user’s response to an IVR because of the user’s interpretation of the IVR’s persona. One identifies the interface’s personality through its human-like characteristics. Beyond IVRs, this quality is found in computers (GUIs) and robotics that produce sounds or images that have such human-like qualities associated with them (van Mulken, André, & Müller, 1998; Wagner, Van der Loos, &

Leifer, 2000; Reeves & Nass, 1997). Although the concept of a persona is debated among the human-computer interaction community (Klie, 2007; Rolandi, 2007), evidence of its existence is quite pervasive within IVR research. Synthetic voices produced by IVRs can possess emotional qualities (e.g. happy, sad; Nass, Foehr, & Somoza, 2001). Such synthetic voices can also be identified as male or female voices, and users will attribute stereotypic gender qualities to these voices (Lee, Nass, & Brave, 2000). For example, the “female” voice was rated as being more attractive than the “male” voice despite the fact that no human was attached to these synthetic voices. Although synthetic voices are given a persona, their quality is not as human-sounding as that produced by natural speech recordings. When participants were asked to compare various IVR voices based on how human-like they sounded, Couper et al. (2004) found that natural (recorded) voices were ranked higher than human-like TTS, which outranked computer-like TTS, indicating that IVR voices were accurately rated according to how human-like they sounded. Although Couper et al. found no effects of voice gender on disclosure, Nass et al. (2003) discovered that users were less responsive with a male and female synthetic voice and more responsive with a male and female natural voice. The female voice surfaced as more productive than the male voice in inducing disclosure.

The concept of persona can be encompassed by another term called social interface theory (Reeves & Nass, 1996; Dryer, 1999; Tourangeau et al., 2003). Not only do people assign a personality to machines, but they will act and respond in a social manner. For example, a voice user interface that identified itself with a person’s name (e.g. “Hello, I’m John”) elicited not only more interaction from more users, but the interaction contained more direct and complete responses than if the VUI did not identify

itself (Knott & Kortum, 2006). Related to this, Tourangeau et al. (2003) found that an IVR utilizing the first person (“I will read you a few statements...”) induced more disclosure of embarrassing information from participants than if it had utilized the third person (“Please listen to a few statements...”).

1.6 The M. D. Anderson Symptom Inventory

Disclosure is at the forefront in the medical arena, and one such application where persona effects may impact user responses is the Community Cancer Care Symptom Monitor, as patients with cancer may find discussion of their symptoms to be a sensitive topic. Cancer patients are affected by symptoms such as vomiting, fatigue, and physical and emotional pain that are caused from both the cancer itself as well as cancer treatment (Cleeland et al., 2000). The M. D. Anderson Symptom Inventory (MDASI; see Appendix A), which gauges the symptom level and interference of 21 items, is a paper-and-pencil inventory that uses an 11-point scale from 0 (no symptom at all) to 10 (the symptom is as bad as you can imagine it could be). It was created by reducing 26 items related to specific symptoms and 6 items related to the interference of one’s activity down to 15 and 6 items, respectively. The items, which were taken from similar scales as well as from input from medical professionals, were reduced using hierarchical cluster analysis to remove redundant items. The IVR version, identical to the MDASI in regards to the items used, was implemented into the Community Cancer Care Symptom Monitor, which remotely monitors cancer patients’ pain levels between office visits. It is the system utilized in this thesis. The system calls a patient at regular intervals determined by the patient’s physician. The patient has the flexibility to schedule the phone calls within a particular hour of the day that is convenient to him or her. Upon answering the phone, the

patient is then guided through the MDASI-IVR. Each question receives a rating from 0 (no symptom at all) to 10 (the symptom is as bad as you can imagine it to be). The interference questions also follow this 0 to 10 rating scale, where 10 is utilized for severe interference. The voice utilized in the current MDASI-IVR is a natural-voiced female who is not a professional voice actor. She was selected to replace a TTS voice. Considering the question that voice may alter user response, it is unclear if the current MDASI-IVR voice is optimal for the system.

1.7 The Current Study

The MDASI-IVR was used to examine users' responses to different personae, as research examining responses to exemplars within gender (e.g. an upbeat versus professional female voice) as well as across genders (comparing these two exemplars for both male and female voices) has yet to be explored thoroughly. The aforementioned research has investigated whether or not a particular gender of physician or system voice may be preferred when being asked sensitive questions, but there is no additional research investigating what characteristics of a female or male voice may reduce response bias. It also seems that the preferred voice will vary not only based on the IVR system, but on the user population.

Both genders were tested as male and female voices are both highly requested by companies hiring voice talent (Klie, 2007). Additionally, "upbeat" and "professional" voices were utilized as they represented two ends of the persona spectrum which are usable in a commercial IVR. Professionalism is highly desired by patients, ranking expertise as the most desired physician quality (Schattner, Rudin, & Jellin, 2004). Breast cancer patients also shared this belief that expertise, defined as "being efficient,

acclaimed, or frank” was necessary in order to trust a physician (Wright, Holcombe, & Salmon, 2004). This concept of expertise could not be demonstrated in a scripted IVR, however the type of voice used could, indeed, convey more or less expertise or professionalism. The following descriptors were given to the voice talent to encompass professionalism: “professional, dispassionate, matter-of-fact, somber, and trustworthy.” Upbeat, on the other hand, was defined as “happy, outgoing, caring, interested, perky, optimistic, and passionate about life.” Aside from interest, these characteristics are not defined as ideal physician behaviors, and an upbeat voice in the absence of additional context could remove the ability to interpret physician expertise, and therefore trust. In addition to the upbeat and professional voices, a sympathetic voice was utilized due to patient preferences for physician characteristics associated with sympathy. Quirk et al. (2008) found that patients in a focus group identified physicians with a soft voice who speak with a slow pace as preferred and more caring. Bendapudi, Berry, Frey, Parish, and Rayburn (2006) examined good and bad physician behaviors collected from patients via an oral critical incident technique. Two of the seven ideal incidents included empathy and humanity. In addition, Beck, Daughtridge, and Sloane (2002) examined verbal behaviors through a metaanalysis and found that empathy, reassurance and encouragement were positively associated with patient compliance with recommended therapy. From the above research, caring, empathy, and encouragement seem to encompass a sympathetic vocal style.

Additionally, patients may react differently to the physician’s personality depending on both the physician’s and patient’s gender. Mast, Hall, and Roter (2007) examined male and female patient reactions to both male and female physicians who

were high or low in dominance and/or caring. For male patients, they were not influenced by communication style, regardless of physician gender. However female patients preferred a caring female physician over a non-caring female physician. Female patients were dissatisfied with a male physician who was non-caring and dominant, as well as the reverse – a male physician who was caring and non-dominant. As previously mentioned, communication style may be more important than gender (Christen et al., 2008). This includes being patient-centered and sharing in the decision-making process.

Unfortunately, aspects such as these that patients seek in their physician may not be easily conveyed through voice, alone. For example, dominance (expressing a power difference between patient and physician) was not expressed in the exemplars utilized in the MDASI-IVR, but the caring role (defined by the authors as concern and empathy) was similar to sympathy. The present study determined whether individuals gave different ratings when listening to voice prompts presented by different voice dimensions and genders.

2.0 Experiment 1

The current research utilized both genders (male and female) and three dimensions of voice (upbeat, professional, and sympathetic) when recording the voice prompts for the experimental version of the MDASI-IVR. In Experiment 1, vocal stimuli were created and subsequently rated to determine if they exemplified the appropriate vocal dimensions. Upon selection of a male and female voice at the completion of Experiment 1, the voices were implemented into the full MDASI-IVR in a second experiment (Experiment 2) in order to determine if there were or were not rating differences as a result of the system voice received.

2.1 Stimulus Creation

Participants

Twenty-three individuals recorded a sample of the MDASI-IVR script in the three voice dimensions (upbeat, professional, and sympathetic). Of the 23 individuals, five (three males and two females) were professional voice talents sought via the internet. They offered a free sample script that the experimenter was able to assess. Four of the professionals were willing to re-record the sample script to accommodate changes requested. The remaining 18 individuals consisted of four adults residing in the Houston, Texas area who recorded the test script on a voluntary basis, and 14 undergraduates at Rice University who recorded the test script and the full script for experimental credit or payment. The four adults were told that they would only be asked to record the full script if they were selected as a final male or female voice because they were performing on a volunteer basis. On the other hand, the undergraduates were asked to record the full script at the same time as the sample script because some of them may have been harder to reach by the time re-recording was in order.

Materials

The undergraduates and four Houstonians recorded their voices using an AKG Perception 200 microphone. The professional voice talent utilized their own equipment. All individuals were given directions as to what was expected for the three voices. For the upbeat voice, one was to use a voice that was “happy, outgoing, caring, interested, perky, optimistic, and passionate about life,” yet not “silly, teasing, sexy, or excited.” These adjectives have been extracted from the participant’s instructions; the full instructions are described in Appendix B. The professional voice was described as

“professional, dispassionate, matter-of-fact, somber, and trustworthy,” yet not “sarcastic, angry, depressed, unhappy, or cold.” Finally, the sympathetic voice was described as “sympathetic, compassionate, sincere, kind, warm, and trustful,” yet not “forced, sarcastic, excited, eager, or whiny.”

Procedure

The undergraduate students met the experimenter in an acoustically-dampened recording room, and they first read the test script to give the experimenter an idea of their ability to perform on this task. One female was dismissed after recording of the test script due to a noticeable speech impediment. The remaining 13 individuals (eight males, five females) continued to record the full MDASI script. Each line of the MDASI was read as many times as necessary to reach the experimenter’s satisfaction; no recording session took longer than one hour. Of these students, six had a background in theater or improvisation.

The experimenter met the four Houstonians at separate times outside of the laboratory. The microphone was set up with a laptop in a quiet room. The test script of the MDASI was recorded several times until the experimenter was satisfied with the quality of the recordings. Because they did not utilize our recording facility, the professional voice talent emailed attachments containing their recordings of the sample script. The experimenter listened to the recordings and emailed back suggested changes, and the individual would then email back an improved recording. This procedure was continued until the experimenter was satisfied with all recordings.

Individuals’ whose voices that appeared to sound quite distinct between the upbeat, professional, and sympathetic recordings while accurately capturing the correct

dimension were sought. Utilization of the same male and female individuals for all three exemplars (upbeat, professional, and sympathetic) was the main objective. Otherwise, any rating differences found in the second portion of Experiment 1 might be due to the dissimilar voices (i.e. the sympathetic male voice selected may have a naturally higher pitch and be more nasal, whereas the professional male voice selected may have a lower pitch and be less nasal) as opposed to a difference in the voice dimensions.

Preliminary Selection

The researchers listened to a portion of the full scripts from all 23 individuals and selected 9 voices (5 males and 6 females) who demonstrated superior performance on the three voice dimensions. From the 9 selected, two of the five professional voices were chosen (one male and one female), as were three of the four adult Houstonians (two males and one female) and four undergraduates (two males and two females, all of whom happened to have a theater/improvisation background). A third researcher who was blind to the conditions listened to several of the voices and could not distinguish between the dimensions for most of the unselected individuals.

2.2 Stimulus Selection

Participants

Sixteen undergraduate psychology students (10 females, 6 males) from Rice University who did not participate in stimulus generation took part in stimulus selection. Participants had a mean age of 18.50 years ($SD = 0.73$ years) and 10 of the 16 were native speakers of English.

Materials

Instructions were read at the beginning of the experiment. Specifically, the participants were told,

For this experiment, you will be listening to several voice recordings and rating each of them on a personality scale. Please rate the recordings as honestly and as accurately as you can. As you listen to the recordings, you will notice that some have a better recording quality than others. Please do not consider the *recording* quality when rating the voices and instead focus on the actual *vocal* qualities, such as pitch or intensity.

The personality scale was utilized to rate each voice along the three key vocal dimensions (perkiness, professionalism, and sympathy) on a 7-point scale from 1 (not at all) to 7 (very much). In addition, 11 personality facets were selected for inclusion in the personality scale. These items were selected from the adjectives utilized by the voice actors when recording the scripts or from the articles examining patient satisfaction with physicians or ideal physician behaviors which helped surface professionalism and sympathy as key voices to utilize (e.g. Quirk et al., 2008; Wright et al., 2004). These facets were also rated on the same 1-7 scale. It was important to determine if the adjectives utilized by the voice talent or by patients to describe “professionalism,” for example, do accurately reflect how a third party would rate a professional-sounding voice. A final question was also asked in regards to how much the individual liked the voice they had just heard (again, on the 1-7 scale from “not at all” to “very much”). This rating scale is in Appendix C.

Procedure

Participants were tested one at a time. They entered the acoustically-dampened laboratory and sat at a computer station. On the computer monitor was a playlist with the 28 voices to be played. These 28 voices were comprised of the 9 selected individuals' recordings from Stimulus Creation (each individual having 3 recordings -- an upbeat, professional, and sympathetic track), plus the female voice currently utilized on the MDASI-IVR system, the personality of which is unidentified. The latter was added in to determine the qualities perceived in the current system's voice. Participants were given a consent form, after which they were read instructions and then given a personality scale to use for the first voice. After one voice was heard, the participant filled out a personality scale. The experimenter, who was seated beside the participant at the computer, casually observed the participant to make sure that no items were missed and to take notes on which items were rated particularly high or low. Once the scale was complete, the experimenter would play the next voice, which would be followed up with another personality scale, and so on until all 28 voices were played. The voices were randomly presented. Following this portion of the experiment, the participant was asked follow-up questions to determine why a voice was rated on the high or low end of the scale. In addition, they were asked which voice dimension or gender they would prefer for the sample script of the MDASI-IVR they had heard. Finally, they were asked basic demographic questions, debriefed, and dismissed.

2.2.1. Results & Discussion of Stimulus Selection

In order for a male and female voice to be considered for use in Experiment 2, the voice must demonstrate that it was highly rated on the appropriate dimensions on the personality scale. In addition, if one personality dimension (i.e. professionalism) was

being rated, one would have expected a high rating in this domain and a low rating on the other two unrelated domains of perkiness and sympathy. For example, if a listener rated a professional voice as highly professional and perky, then the voice would not have stood out as an exemplar of a professional voice, but would have been entwined with the upbeat voice, which is a separate dimension being investigated.

Graphical Depiction of Voices

Bar graphs for the 27 voices were created to determine which, if any, of the voices visually stood out. All individuals are denoted by their gender and ID number, i.e. female number 18 will be called “F18” for the duration of this paper. The original MDASI-IVR voice will be examined independently after this section. From the bar graphs (see Figures 1a & 1b), two voices, F19 and M21, appeared to have consistently high ratings across all three vocal dimensions, whereas other voices such as M10 had a high mean rating on two of the items, but a lower comparative rating on the third item. Voice talents whose means fell below 4 were judged to have sub-par performance in expressing the desired personality trait. These poorly-rated voices were M15, M20, and F18. They were excluded from further consideration. Overall, the upbeat voices had higher means, indicating that they not only fit the appropriate “upbeat” classification, but that they were easy to identify as such. The sympathetic voices, on the other hand, had lower means overall as it was harder for the raters to give this more abstract personality a high rating.

ANOVAs examining within-voice differences

While the bar graphs demonstrated how the participants rated the superior dimension for each voice (i.e. “upbeat” for the upbeat voice), they did not consider

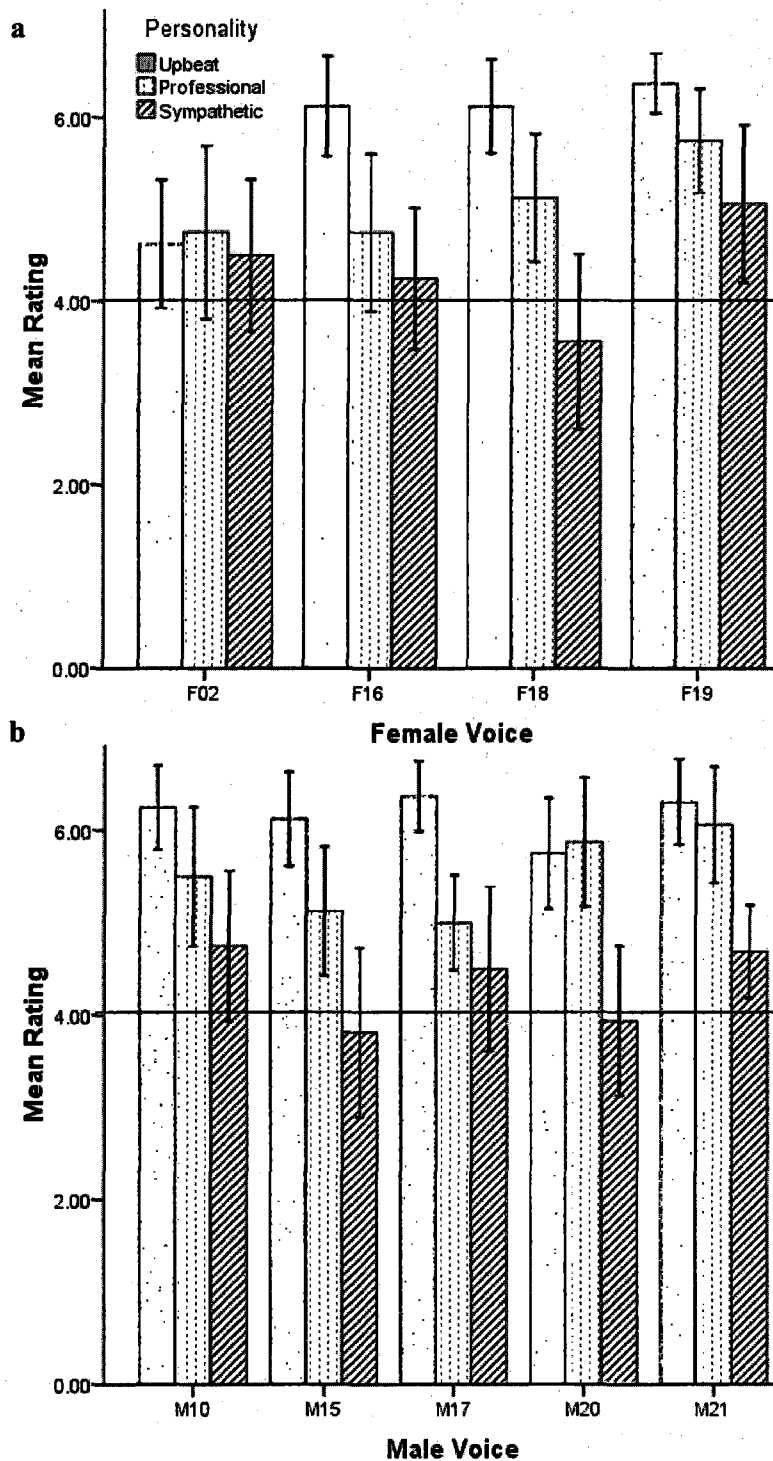


Figure 1 Bar graphs comparing user ratings of dominant items (i.e., upbeat rating for upbeat voice) for male and female voices. *Note.* Error bars = 95% CI. Horizontal line at Mean Rating = 4 indicates average response of 4 on 7-point scale. Mean ratings under this line were considered exceptionally low.

within-voice rating effects. Inferior dimensions needed to be rated lower than superior dimensions in order to determine that a particular voice indeed exemplified the desired personality. For example, the upbeat voice needed to have a high rating on the superior dimension of “upbeat,” but a low rating on the inferior dimensions of “professionalism” and “sympathy.” For the professional voice, “professionalism” was the superior dimension while “upbeat” and “sympathy” were the inferior dimensions. Table 1 displays the repeated measures ANOVAs for each voice still being considered. F02’s upbeat voice and M17’s professional voice had an unreliable difference between means, so they were both excluded, as these two voice dimensions were not discrete, as demonstrated by the ANOVAs.

Contrasts examining within-voice differences

The remaining four voices were followed up with contrasts to examine if, indeed, the 3 superior dimensions were rated higher than their inferior counterparts for a particular personality. To control for Type I error rate, all contrasts used a global Bonferroni correction, utilizing a critical α of .017 for this set of analyses. For female options, F16 and F19 remained. Contrasts on F16 revealed that upbeat ratings for her upbeat voice ($M = 6.13$, $SD = 1.02$) differed reliably from ratings of sympathy ($M = 3.75$, $SD = 1.39$), $t(15) = 4.93$, $p < .001$, $d = 1.23$, and professionalism ($M = 3.94$, $SD = 1.48$), $t(15) = 4.36$, $p = .001$, $d = 1.09$. For F16’s professional voice, the professional rating ($M = 4.75$, $SD = 1.61$) did not reliably differ from the sympathetic rating ($M = 3.69$, $SD = 1.45$), $t(15) = 2.64$, $p = .019$ or the upbeat rating ($M = 4.69$, $SD = 1.40$), $t(15) = 0.11$, $p = .912$. Finally, F16’s sympathetic voice rating indicated that raters thought the voice was more sympathetic ($M = 4.25$, $SD = 1.44$) than upbeat ($M = 2.69$, $SD = 1.35$), $t(15) = 3.57$,

Table 1: ANOVAs for voices being considered after initial elimination

Voice	<i>M</i>	<i>SD</i>	<i>df (tx, error)</i>	<i>F</i>	<i>p</i>
Upbeat					
F02	4.63	1.31	2, 30	0.83 ^a	.444
F16	6.13	1.02	2, 30	17.36	<.001
F19	6.38	0.62	2, 30	40.13	<.001
M10	6.25	0.86	1.14, 17.07 GG	35.79	<.001
M17	6.38	0.72	2, 30	80.39	<.001
M21	6.31	0.87	2, 30	55.67	<.001
Professional					
F02	4.75	1.77	1.48, 22.17, HF	21.16	<.001
F16	4.75	1.61	2, 30	3.46	.044
F19	5.75	1.06	2, 30	28.44	<.001
M10	5.67	1.29	2, 28	17.95	<.001
M17	5.00	0.97	2, 30	2.47	.101
M21	6.06	1.18	2, 30	45.49	<.001
Sympathetic					
F02	4.50	1.55	1.70, 25.46, HF	10.67	.001
F16	4.25	1.44	2, 30	9.20 ^a	.001
F19	5.06	1.61	2, 30	18.23	<.001
M10	4.75	1.53	2, 30	16.87 ^a	<.001
M17	4.50	1.67	2, 30	13.32	<.001
M21	4.69	0.95	1.67, 25.08, HF	24.68 ^a	<.001

Note. ^a: For these voices, the rating for professionalism (inferior dimension) was higher than the superior dimension being rated.

$p = .003$, $d = 0.89$, however not any different from professionalism ($M = 4.31$, $SD = 1.54$), $t(15) = 0.16$, $p = .879$.

Indeed, the rating of upbeat for the upbeat voice for F19 ($M = 6.38$, $SD = 0.62$) was reliably different from the ratings of sympathy ($M = 3.44$, $SD = 1.67$), $t(15) = 6.14$, $p < .001$, $d = 1.52$, and professionalism ($M = 2.56$, $SD = 1.46$), $t(15) = 8.86$, $p < .001$, $d = 2.22$. The same pattern held for F19's professional voice, with the ratings for professionalism ($M = 5.75$, $SD = 1.06$) being higher than the ratings of sympathy ($M = 4.44$, $SD = 1.67$), $t(15) = 2.84$, $p = .013$, $d = 0.71$, and upbeat ($M = 2.50$, $SD = 1.03$), $t(15) = 8.06$, $p < .001$, $d = 2.02$. However, F19's sympathetic voice demonstrated that the rating of sympathy ($M = 5.06$, $SD = 1.61$) was reliably different than the rating of upbeat ($M = 2.25$, $SD = 1.18$), $t(15) = 5.43$, $p < .001$, $d = 1.36$, but was not so from the rating of professionalism ($M = 4.81$, $SD = 1.47$), $t(15) = 0.44$, $p = .669$.

For M10, the male who was not a professional voice talent, custom contrasts revealed that the upbeat rating for his upbeat voice ($M = 6.25$, $SD = 0.86$) was reliably different from both the rating of sympathy ($M = 3.19$, $SD = 1.56$), $t(15) = 5.89$, $p < .001$, $d = 1.47$ and professionalism ($M = 3.06$, $SD = 1.48$), $t(15) = 6.35$, $p < .001$, $d = 1.59$. Similar to F19, his professional voice demonstrated that the rating of professionalism ($M = 5.50$, $SD = 1.41$) was rated reliably different than the sympathetic rating ($M = 3.13$, $SD = 1.30$) and upbeat rating ($M = 3.13$, $SD = 1.20$) where $t(15) = 5.21$, $p < .001$, $d = 1.34$ and $t(15) = 4.34$, $p = .001$, $d = 1.09$, respectively. However once again, the sympathy rating for the sympathetic voice ($M = 4.75$, $SD = 1.53$) was reliably different from the rating for upbeat ($M = 2.63$, $SD = 1.31$), $t(15) = 6.14$, $p < .001$, $d = 1.52$, however it was not rated

reliably different from the professional rating ($M = 4.88$, $SD = 1.36$), $t(15) = 6.14$, $p < .001$, $d = 1.52$.

Finally, for M21, custom contrasts revealed that the upbeat rating for his upbeat voice ($M = 6.31$, $SD = 0.87$) was reliably different from both the rating for sympathy ($M = 2.25$, $SD = 1.18$) and professionalism ($M = 2.69$, $SD = 1.30$) for the upbeat voice, where $t(15) = 9.84$, $p < .001$, $d = 2.46$ and $t(15) = 7.39$, $p < .001$, $d = 1.85$, respectively. M21's professional voice revealed a reliable difference between ratings on professionalism ($M = 6.06$, $SD = 1.18$) and sympathy ($M = 3.63$, $SD = 1.36$), $t(15) = 7.72$, $p < .001$, $d = 1.92$, or professionalism compared to upbeat ($M = 2.44$, $SD = 1.36$), $t(15) = 7.66$, $p < .001$, $d = 1.91$, as well. However, once again, while the sympathetic voice demonstrated that the rating for sympathy was reliably different from the rating for upbeat ($M = 2.88$, $SD = 1.50$), $t(15) = 4.04$, $p = .001$, $d = 1.01$, it was reliably different from the rating for professionalism, $t(15) = 4.04$, $p < .001$, $d = 1.01$, however it was not in the desired direction – professionalism ($M = 5.94$, $SD = 0.85$) was rated higher than sympathy ($M = 4.69$, $SD = 0.95$).

Final selection of a male and female voice

There was a consistent pattern across all four voices whereby, for the sympathetic voice, sympathy was tied to professionalism. Raters found that the professional voice could be absent of sympathy, however the sympathetic voices were rated high on both sympathy and professionalism. This implied that the raters sought professionalism in the sympathetic voices. This was confirmed through the interviews with individuals after the experiment who believed that a sympathetic voice also needed to sound professional or else it would sound too “over the top,” “phony,” or “forced.” Their specific comments

regarding what they liked about some of the sympathetic voices presented included the voices sounding sympathetic, yet professional. This preference is confirmed in Klie (2007) where it is stated that users calling an IVR expect to hear a professional-sounding voice. Therefore, the fact that sympathy was not distinguished from professionalism was not considered a shortcoming and was not used to distinguish a better or worse voice talent from another.

For female options, both F16 and F19 remained. Examination of the planned contrasts indicated that F16 did not have a good professional voice, as it was not ranked as more professional than upbeat or sympathetic. This was the case for F16's sympathetic voice, as well. On the other hand, all three of F19's voices were separable and had high means on the 7-point scale across the board (see Figure 2a). F19's ratings were consistently higher than F16. This was confirmed with paired samples *t*-tests, indicating that F19 outperformed F16 on her professional voice ($t(15) = 2.28, p = .037, d = 0.57$), however the means were not reliably different with the upbeat voice ($t(15) = 1.07, p = .300$) or the sympathetic voice ($t(15) = 1.52, p = .149$). Because F19's three voices were distinct representations of the three exemplars and her ratings were even higher than her closest competition, F16, F19 was selected as the final female voice.

For the males, M10 and M21 remained. Their means across the three voice types seemed quite comparable, and this was confirmed as all three *t*-tests did not reveal a reliable difference between the two male voices ($t(15) = 0.29, p = .774$ for upbeat; $t(15) = 1.38, p = .188$ for professional; and $t(15) = 0.19, p = .855$ for sympathetic). Ultimately, M21 was selected over M10 due to the quality of M21's voice. As M21 was one of the professional voices, his voice was lower and more consistent. M10, on the other hand,

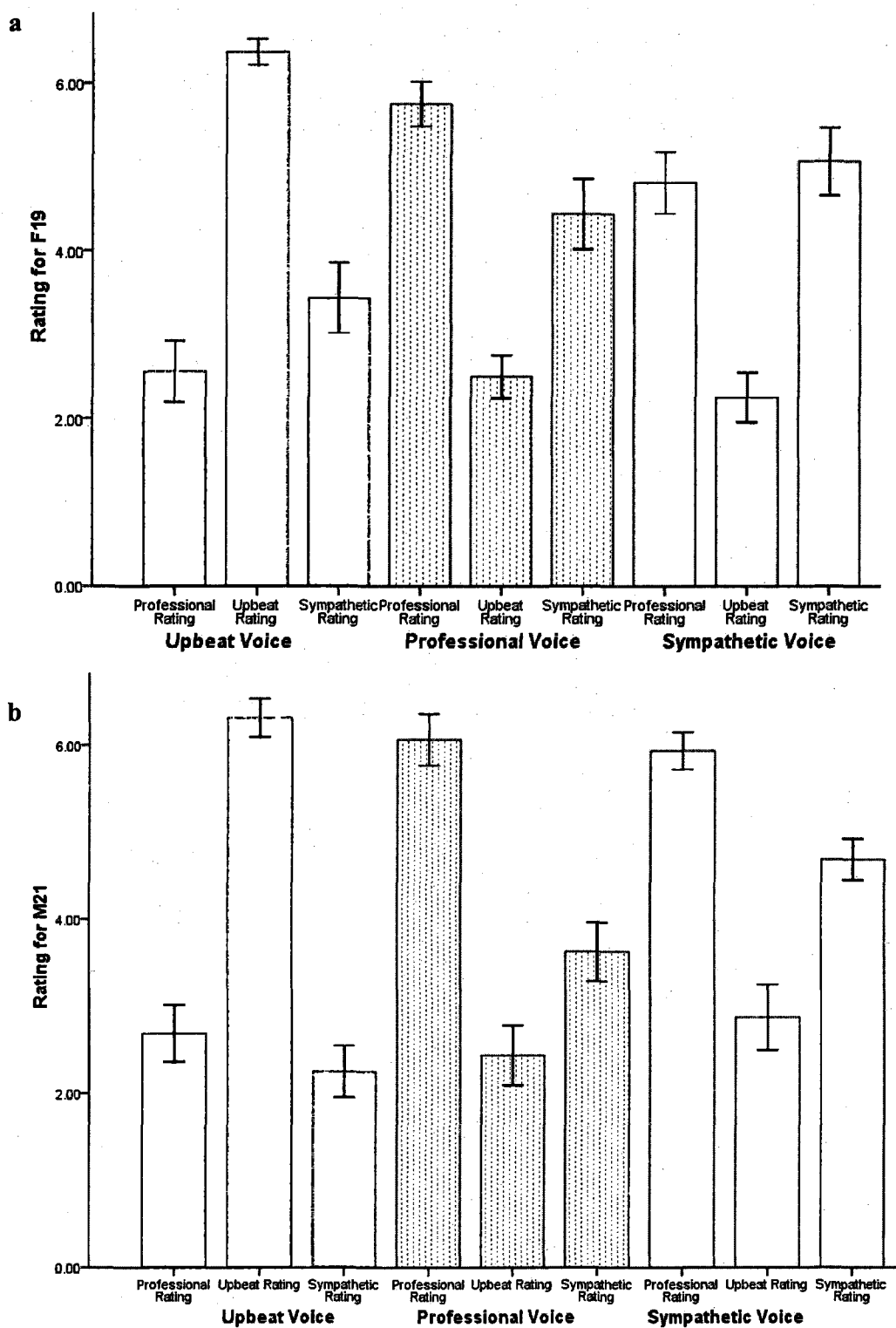


Figure 2 Bar graphs demonstrating user ratings of the selected male and female voices.
Note. Error bars ± 1 SEM.

had a higher-pitched voice that indeed sounded like a young adult. While this would not necessarily be problematic with the undergraduate sample to be utilized in Experiment 2, it may become an external validity issue if the voices are to be used in the future on a population that does not involve individuals around the same young age as M10. Further, Klie (2007) indicated that 81% of companies seeking voice talent selected a middle-aged individual so the voice produced by a middle-aged individual would be more familiar to most users as it is highly predominant across media. Therefore, as their voices were highly equivalent in terms of performance, quality helped lead to a decision toward M21. Figure 2b showed the separation found in M21's upbeat, professional, and sympathetic voices.

Trust & Liking of Voices

The voice that was trusted the most was M21's sympathetic voice ($M = 5.31$, $SD = 1.40$), followed by F02's upbeat voice ($M = 5.06$, $SD = 1.06$). These findings are related to how much the users liked particular voices. M21's sympathetic voice ($M = 5.25$, $SD = 1.13$) and F02's upbeat voice ($M = 5.38$, $SD = 1.15$) were the most liked voices. Interestingly, there is a strong positive correlation between trust and liking for all of the voices – this trend is demonstrated with M21 in Table 2. Examining F19's voice, there was not a main effect of personality on trust, $F(2, 28) = 2.64$, $p = .089$, however there was an interaction with the raters' gender, $F(2, 28) = 5.07$, $p = .013$. Specifically, simple main effects revealed that females rated F19's professional voice higher than males, $t(14) = 2.84$, $p = .013$, however there was no rating difference based on user gender for F19's upbeat voice ($t(14) = 1.86$, $p = .084$) or sympathetic voice ($t(14) = 0.90$,

Table 2: Correlations between M21's trust and like ratings ($n = 16$)

	<i>M</i>	<i>SD</i>	1	2	3	4	5	6
1. M21 U ^a -- Trust	3.13	1.02	(1)					
2. M21 U -- Like	3.19	1.33	.72**	(1)				
3. M21 P -- Trust	4.69	1.49	-.32	-.10	(1)			
4. M21 P -- Like	4.63	1.41	-.15	.08	.80**	(1)		
5. M21 S -- Trust	5.31	1.40	.39	.15	.05	-.07	(1)	
6. M21 S -- Like	5.25	1.13	.49	.37	.17	.32	.62**	(1)

Note. ** $p < .01$, two-tailed; ^a: U = Upbeat; P = Professional; S = Sympathetic.

$p = .384$, see Figure 3a). When examining how much users liked F19's voice, there was an interaction between user gender and F19's personality, $F(2, 28) = 3.68$, $p = .038$.

An analysis of simple main effects in figure 3b suggested that females may have rated F19's professional voice higher than males did, $t(14) = 2.13$, $p = .052$, however once again there was no rating difference based on user gender for F19's upbeat voice ($t(14) = 1.72$, $p = .108$) or sympathetic voice ($t(14) = 0.16$, $p = .876$, see Figure 3b).

For M21, on the other hand, there was no interaction with raters' gender, $F(2, 28) = 0.45$, $p = .643$, however there was a main effect of personality on raters' trust, $F(2, 28) = 10.28$, $p < .001$. A linear contrast indicated that the upbeat voice was less trusted than the professional voice, which was less trusted than the sympathetic voice, $F(1, 14) = 34.55$, $p < .001$. Examining the degree to which users liked M21's voices indicated the same trend, whereby there was no interaction between personality and raters' gender, $F(2, 28) = 0.14$, $p = .867$, however there was a main effect of personality on raters' trust, $F(2, 28) = 12.14$, $p < .001$. Once again, a strong linear contrast demonstrated that the

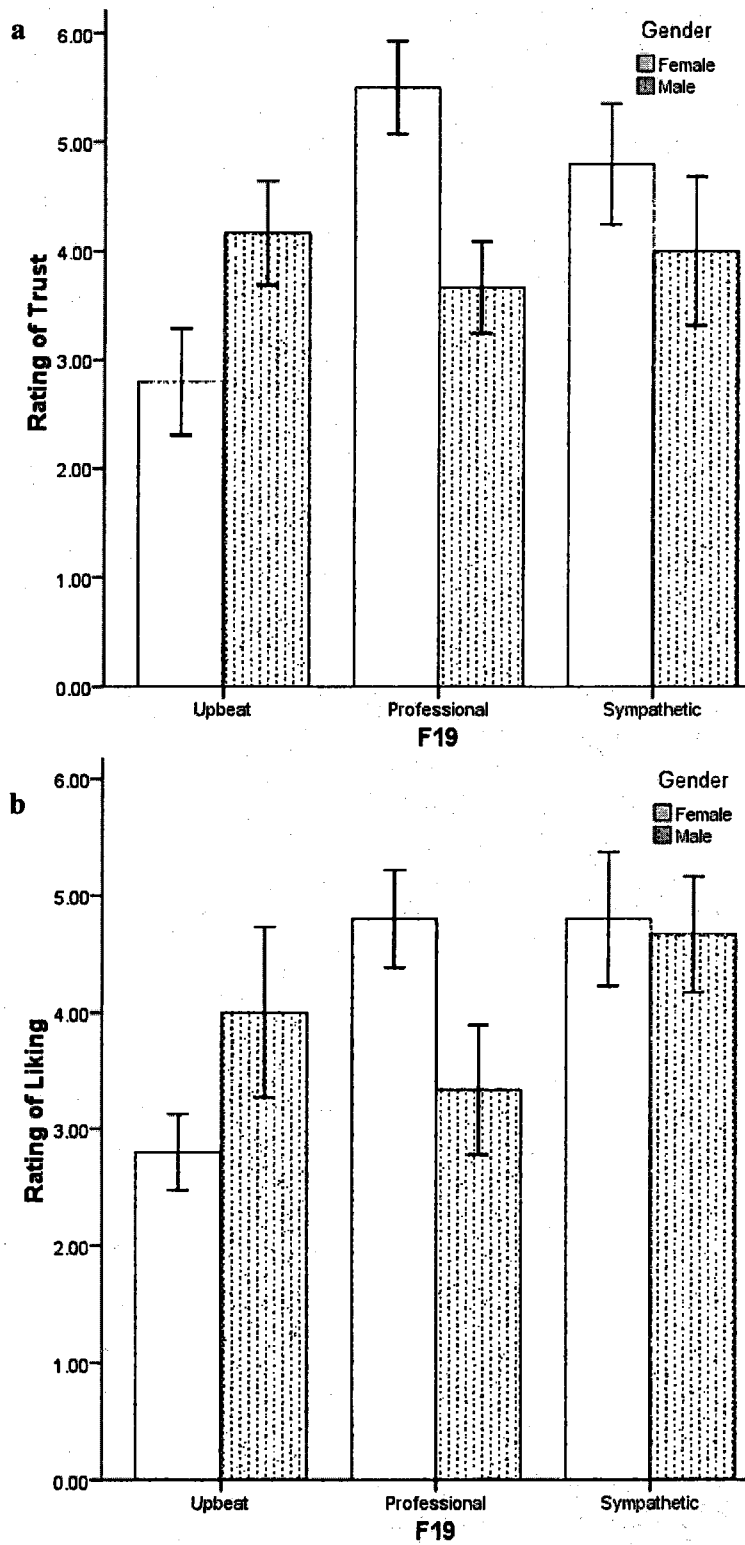


Figure 3 Ratings of trust and liking by rater gender for F19. *Note.* Error bars = 95% CI

upbeat voice was liked less than the professional voice, which was less liked than the sympathetic voice, $F(1, 14) = 30.54, p < .001$ (see Figures 4a & b).

The Current MDASI-IVR Voice

From an examination of the current MDASI-IVR voice, a repeated-measures ANOVA indicated that there was a difference between at least one of the means (upbeat, professional, or sympathetic), $F(2, 30) = 7.23, p = .003$, see Figure 5. Following adjustment with Scheffé, custom contrasts indicated that participants rated the current voice higher on professionalism ($M = 5.13, SD = 1.31$) than on sympathy ($M = 3.75, SD = 1.00$), however there was not a reliable difference in rating between upbeat ($M = 3.81, SD = 1.11$) and professionalism, where $t(15) = 3.47, p = .003, d = 0.87$ for the first contrast comparing professionalism to sympathy and $t(15) = 2.68, p = .017, d = 0.67$ for the second contrast comparing professionalism to perkiness. Additionally, the rating for upbeat did not reliably differ from the rating for sympathy, where $t(15) = .19, p = .849$. This indicated that the current MDASI-IVR voice does not exhibit strong sympathetic characteristics but it does exhibit professional and upbeat characteristics. To date, the current voice had not been categorized.

Experiment 1 Discussion

In Experiment 1, it was demonstrated that both a male and female were able to exhibit separable exemplars. Of the nine voices tested, three voices (M10, M21, and F19) had a high level of voice separation. Their sympathetic voices were rated as highly professional as well as highly sympathetic, however the raters expressed a preference for this pattern. The discovery of both a male and female who demonstrated the ability to record an upbeat, professional, and sympathetic script that encompassed the

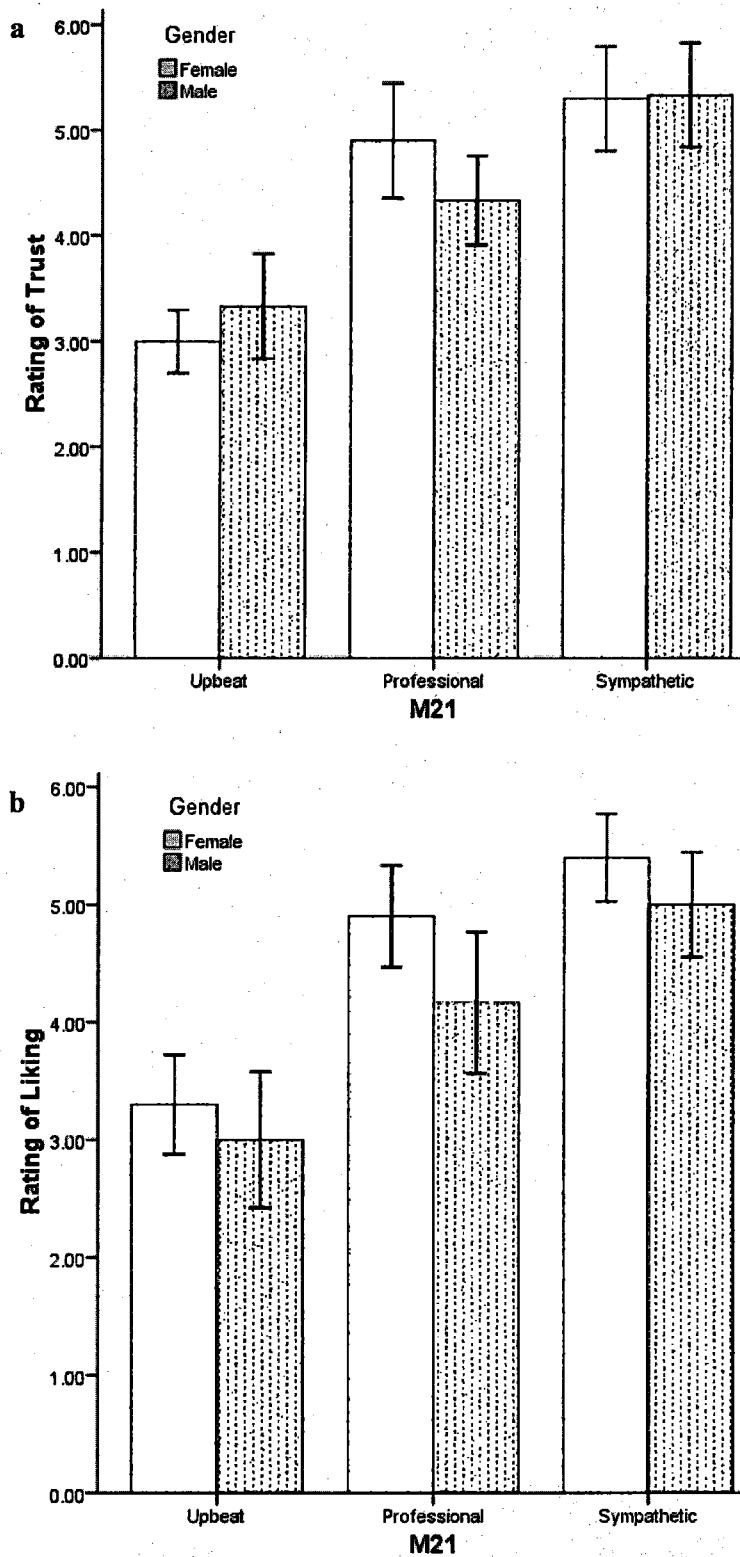


Figure 4 Ratings of trust and liking by rater gender for M21. *Note.* Error bars = 95% CI

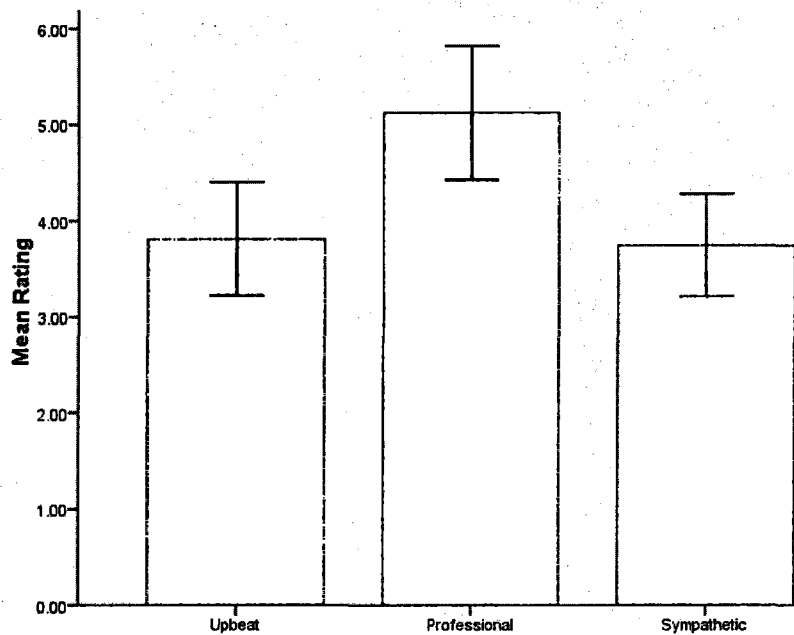


Figure 5 Rating for the current MDASI-IVR's voice. *Note.* Error bars = 95% CI.

appropriate characteristics of the personality to be captured is noteworthy. The finding in Experiment 1 supports the belief that one individual has the ability to express a variety of different personalities which do not overlap on measured dimensions. More importantly, identification of voices that could perform in all three persona dimensions allowed the creation of single-voice stimuli that could be utilized in Experiment 2.

3.0 Experiment 2

The purpose of Experiment 2 was to determine if voice affects the ratings given by users in a medical interactive voice response system. Using the voices from Experiment 1, it could be determined if system voice was an important factor when designing an IVR. In order to discover this, users were assigned to different system voices (F19 and M21's upbeat, professional, and sympathetic voices) and their symptom

ratings on the MDASI-IVR were measured. Comparisons were made across system voice type, system voice gender, and user gender. In addition, the relationship was examined between the MDASI-IVR scores and baseline health scores to examine if system voice induced over- or underreporting compared to a baseline health rating.

3.1 Measurement of Response Differences

Participants

Two-hundred ninety five undergraduates who did not participate in Experiment 1 were recruited from the experiment pool from November 9, 2008 to December 5, 2008. Users completed the experiment over the web, using a telephone when cued. Of these 295 users, 3 failed to finish the IVR to completion and 5 failed to finish the web portion. These eight users' partial responses were discarded, giving a final participant count of 287. Of the 287 users, 155 were female and 132 were male, with a mean age of 19.29 years ($SD = 1.70$). 243 (84.3%) were native English speakers, with the others reporting bilingualism (1.4%) or having learned a language other than English at birth (13.9%). Two hundred fifty-five users completed the MDASI-IVR in one attempt, whereas the rest took more than one try (9.8%, or 28 users, taking 2 tries; 1.0%, or 3 users, taking 3 tries; and 0.3%, or one user taking 4 tries). Debriefs from the participants indicated that the most common reason for multiple attempts was that their cell phones lost reception during the IVR portion of the experiment.

Materials

The full scripts for these voices (upbeat, professional, and sympathetic for male and female) were implemented into the MDASI-IVR. A replica of the MDASI-IVR was built using Pronexus VBVoice software. The replica matched the IVR identically in

giving users instructions upfront that could be repeated by pressing “2,” allowing all questions to be repeated by pressing the STAR (*) key, repeating all questions requiring DTMF input 3 times before resorting to an error prompt and subsequent disconnection, and informing users if an input was not recognized (e.g. if an “11” was entered when a number from “0” to “10” was required). Because undergraduates were tested in lieu of cancer patients, a few lines of the MDASI script were altered to reflect this difference. For example, instead of saying "People with cancer frequently have symptoms that are caused by their disease or by their treatment," the script was altered to read: "People who do not feel well frequently have symptoms that are caused from being sick or injured." In addition, users were asked to enter their “participant ID” instead of their “patient ID and birth date” (see Appendix D for the full MDASI-IVR script). Aside from these few changes, the 21 symptom and interference questions were not altered in any way. The 21 DTMF responses to the four different voices were recorded, and from these, it could then be determined if there were rating differences between the voices administered and within a voice, if ratings were related to overall subjective health scores.

A web-based version of the MDASI which mirrored the paper-and-pencil version was also given to users (see Appendix A). This version was identical to the MDASI-IVR although it lacked the IVR-related instructions. Answers from 0 to 10 were selected via pull-down menus. Users received both the MDASI-IVR and the MDASI in order to check the reliability of responses over the MDASI-IVR. The order of presentation was counterbalanced so that half of the participants received the MDASI-IVR first and half received the web-based MDASI first. A dramatic difference in responses between the MDASI-IVR and the web-based MDASI would indicate need for further investigation.

A second check on the MDASI-IVR system reliability was conducted utilizing the System Usability Scale (SUS; Brooke, 1996). The SUS is a 10-item scale which measures subjective usability; it was administered immediately following the MDASI-IVR. The SUS was utilized because an underlying usability issue could potentially drive poor results. After hanging up the telephone, users clicked a button on the web survey indicating completion of the call and inserted comments if they encountered any issues when making the call. The next page on the web survey was a web-based version of the SUS, where the 10 items were listed with pull-down menu options. The SUS gives researchers or practitioners a picture of users' overall satisfaction with the product or system, in this case, the IVR interface. It is rated on a 5-point scale from 1 (strongly disagree) to 5 (strongly agree; see Appendix E). The SUS is scored according to the rubric in Brooke's paper, giving each completed scale a score from 0 to 100, with 100 indicating a highly usable product or system. Bangor, Kortum, and Miller (2008) noted that an acceptable SUS score falls above 70, while exceptional products have SUS scores above 90. Comparing the aggregate SUS score for the MDASI-IVR to this criterion indicates the desire for a SUS score from the high 70s to the 90s, understanding the difficulty of achieving a very high score.

To gauge overall subjective mental and physical health, users completed the Centers for Disease Control Health Related Quality of Life (CDC-HRQOL) 4-question Healthy Days Measure and 5-question Healthy Days Symptom Module for purposes of validation (see Appendix F). The CDC-HRQOL-4 has been utilized in the State-Based Behavioral Risk Factor Surveillance System (BRFSS) since 1993, and a full set of 14 questions has been available since 1995. The BRFSS is a standardized phone interview

process which tracks U.S. health conditions and trends. Other common health surveys looking at symptom duration over the past month such as the SF-12, a 12-item version of the SF-36 health survey (Ware Jr., Kosinski, & Keller, 1996) appeared to have more range restriction compared to the CDC-HRQOL. The CDC-HRQOL, on the other hand, can elicit much variance as it uses a 31-point scale (from 0 to 30 representing days in a month), whereas many SF-12 items use a 5 point scale. Since a 31-point scale is open to error, Andresen, Catlin, Wyrwich, and Jackson-Thompson (2003) investigated test-retest reliability of the HRQOL Core (Healthy Days Measure) from a sample of 868 Missourians in 1999. Individuals were administered the HRQOL Core two weeks apart, and the test-retest reliability ranged from $\kappa = 0.60$ to $\kappa = 0.76$ for individuals between ages 18 and 64. This moderate kappa score indicated that subjective reports of symptom duration were fairly stable across time. The CDC-HRQOL has been utilized to investigate quality of life in cancer survivors (Richardson, Wingo, Zack, Zahran, & King, 2008), thus prior research did indicate the test's stability with cancer patients, who are the focus of the MDASI. Construct validity has been investigated in both unhealthy (Mielenz, Jackson, Currey, DeVellis, & Callahan, 2006; Andresen, Fouts, Romeis, & Brownson, 1999), and healthy (Zullig, 2005; Zullig, Valois, Huebner, & Drane, 2004) samples.

Procedure

Undergraduates who did not participate in Experiment 1 were randomly assigned to listen to one of the six voices selected in the previous experiment. The voice was either female (upbeat, professional, or sympathetic) or male (upbeat, professional, or sympathetic). Users completed the entire experiment from any available computer with

an internet connection for the web survey portion and any available phone (land line or cellular) for the IVR portion. By performing the experiment away from the laboratory, it was hoped that fidelity (related to the IVR system) and honesty in responses would increase. Users were emailed a link to the web survey a day before their scheduled time which included their participant ID. Scheduled sessions were set to avoid congestion on the system that delivered the IVR. Users read a consent form online and gave their consent via checking a box and typing their name on a designated line. The next page asked for the 3-digit participant ID from the email previously sent. Users needed to have this number to continue beyond this point. This number was used to link up the user's IVR and web data, and it also designated which voice the user heard during the IVR portion.

The experiment was presented in two orders. In the first, users started with the MDASI-IVR, followed by the System Usability Scale, CDC-Health Related Quality of Life scale, web-based version of the MDASI, and ended with demographic questions. The second presentation started with the web-based version of the MDASI, followed by the CDC-Health Related Quality of Life Scale, the MDASI-IVR, System Usability Scale, and finally the demographic questions. These two orders were selected for two specific reasons. First, the SUS necessarily had to follow the MDASI-IVR, as it was a rating of the IVR's usability. Second, the web-based MDASI and MDASI-IVR were spaced out as the first and last surveys presented to avoid item recall when answering the identical items from one instance of the MDASI to the next. To access the MDASI-IVR, users were given a phone number on the web survey giving brief instructions about the short phone call they were about to make. Their 3-digit ID from the previous page reappeared

on this page, and they were told to use that number for the phone portion. The first digit of the ID was used to filter users into the 6 possible IVR voices. Following the 5-minute IVR, users confirmed call completion on the web survey by checking a box. Upon checking this box, a question appeared asking users how much they trusted the voice they just heard from 1 (not at all) to 7 (very much). There was also a text box in which users could comment on any errors with the system or with their phone (e.g. a prompt that did not work or if their cell phone disconnected) that may have prevented them from reaching the end of the survey were they unable to finish the call. After completing this page, users filled out the 10-question SUS. Following the SUS, users were presented with the 9-question CDC-HRQOL, and then the written MDASI. At the end of these questions, there was a short page requesting demographics, and then users were taken to the debriefing page and expected to hit the “finish” button to submit their responses. The “back” button was removed from the survey to eliminate the possibility of changing answers or utilizing the written MDASI as a template for the MDASI-IVR. The second presentation of the web survey was identical to the above, except the MDASI-IVR and SUS were moved to the end preceding the demographic questions.

While users completed the IVR, a computer recorded the DTMF responses (key presses from 0 to 10) entered by the users. Upon the IVR completion, users filled out the CDC-HRQOL for validation purposes. A user’s written responses on the CDC-HRQOL health survey were expected to correlate with his responses on the MDASI-IVR – if he reported severe pain on one survey and not the other, then any results made about the MDASI-IVR would have to be interpreted with caution. Finally, the users completed the SUS to measure satisfaction with the MDASI-IVR. Upon completion of the SUS, users

were sent to a page presenting the debriefing and they were then allowed to close their browsers.

3.2 Results & Discussion of Experiment 2

There were roughly 16 to 17 percent of users in each of the six groups, ranging from 46 users who listened to the upbeat male voice, to 49 users who listened to the sympathetic male voice. In addition, there was fairly equal distribution of gender across the six groups. There was a minimum of 24 females in a group (professional female) to a maximum of 27 (upbeat female). For male users, there was a minimum of 20 in a group (upbeat female) to a maximum of 24 (professional female). Box plots examining ratings

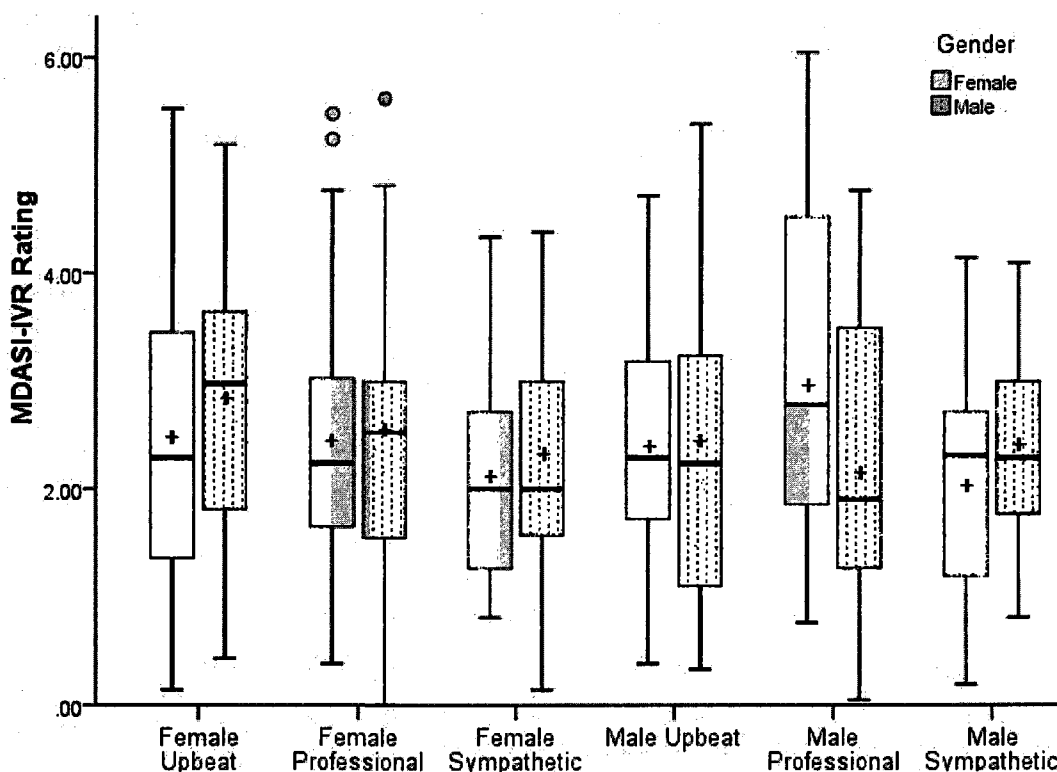


Figure 6 Box plots of ratings on the MDASI-IVR for all voices presented.

on the MDASI-IVR are separated across the 6 IVR voices (F19 and M21's upbeat, professional, and sympathetic voices), and are shown in Figure 6.

Between-voice comparisons

There was no overall rating difference for individuals based on the IVR voice heard while filling out the survey using the phone. A 3 x 2 x 2 ANOVA (voice type by voice gender by rater gender) showed no evidence of a 3-way interaction with rater gender, voice type, and voice gender, $F(2, 275) = 0.77, p = .465$. There was not a 2-way interaction between voice type and voice gender, $F(2, 275) = 0.64, p = .529$, or an interaction between voice type and rater gender, $F(2, 275) = 1.79, p = .169$, or an interaction between voice gender and rater gender, $F(1, 275) = 1.30, p = .255$. The main effect of voice type did not indicate that one voice personality was rated higher than another, $F(2, 275) = 1.65, p = .195$. In addition, there was no evidence for an effect of either system voice gender or user gender, where $F(1, 275) = 0.06, p = .809$, and $F(1, 275) = 0.06, p = .808$, respectively. The finding that no voice surfaced higher ratings than another seemed to indicate that users' ratings were not influenced by system voice gender or personality. Confidence intervals for the main effects and interactions are displayed in Table 3. It was unlikely that there was a practically-significant effect for user gender or voice gender, as every value in their respective confidence intervals is unimportant in a practical sense based on their narrow ranges. Additionally, in determining if there was an effect for the user gender x voice gender x voice type interaction (where voice type compared the upbeat voice to the average between the professional and sympathetic voices), the range of effects was wider, indicating that the true difference may be larger as well. However, the likelihood of missing a practically-important effect is unlikely.

Table 3: Confidence Intervals and Significant Tests for User Gender x Voice Gender x Voice Type ($n = 287$)

	<i>df (tx, error)</i>	<i>F</i>	<i>p</i>	<i>LL</i>	<i>UL</i>
User gender	1, 275	0.06	0.808	-0.187	0.240
Voice gender	1, 275	0.06	0.809	-0.240	0.188
User gender x Voice gender	1, 283	1.31	0.253	-0.478	0.127
Voice type ^a	1, 275	0.27	0.602	-0.334	0.194
Voice type ^b	1, 275	3.02	0.083	-0.488	0.030
User gender x Voice type ^a	1, 281	0.53	0.468	-0.508	0.234
User gender x Voice type ^b	1, 281	3.14	0.077	-0.694	0.037
Voice gender x Voice type ^a	1, 281	0.89	0.346	-0.193	0.548
Voice gender x Voice type ^b	1, 281	0.43	0.515	-0.245	0.487
User gender x Voice gender x Voice type ^a	1, 275	0.07	0.797	-0.596	0.458
User gender x Voice gender x Voice type ^b	1, 275	1.47	0.227	-0.837	0.199

Note. Voice type^a compares the upbeat voice with the average of the professional and sympathetic voices. Voice type^b compares the professional and sympathetic voices. All contrasts were normalized to 1.

Scoring of the MDASI-IVR & CDC-HRQOL

There is no predefined scoring rubric for the MDASI. Items have not been investigated on aggregate to date, however investigators have been moving in this

direction. While factor analysis demonstrated four distinct subscales for the MDASI-IVR (a mental, physical, fatigue, and distress factor), a grand mean was utilized for all MDASI-IVR items in lieu of the subscales as there was no prior research to confirm the accuracy of the subscales.

The CDC-HRQOL does not combine its items into subscales in order to maintain transparency (Moriarty, Zack, & Kobau, 2003). Instead, each individual item was designed for use as a single measure. However, when factor analysis was utilized to group the items, a mental and a physical group did clearly surface, similar to the SF-36 or SF-12 health surveys (Mielenz et al., 2006). While factor analysis revealed these two components in the current data, it would not be wise to utilize separate components when the overall MDASI-IVR score was to be used. Additionally, many researchers have used a subscore comprised of questions 2 and 3 from the HRQOL-4, but this would have excluded the questions included from the Healthy Days Symptom Module. A grand mean of the 8 CDC-HRQOL questions, excluding the first question which utilized a different scale, was taken.

The overall MDASI-IVR correlated very well with the overall CDC-HRQOL score, $r(285) = .50, p < .001$; utilizing the MDASI-IVR mental subscale with the CDC-HRQOL mental subscale did not dramatically increase the correlation, $r(285) = .58, p < .001$, so it did not seem crucial to use the subscores. The correlations, even when relying on the subscores, were not any higher because the MDASI-IVR asks about symptom severity over the past 24 hours and the CDC-HRQOL asks about the number of days over the past month in which a symptom persisted. A perfect correlation between the two was not expected, but a positive correlation indeed was.

When placing the MDASI-IVR, the web-based version of the MDASI, and the CDC-HRQOL on the same scale, it is important to see how similarly users rated their symptoms across the three items (see Figure 7). A repeated-measures ANOVA indicated that there was a rating difference across the three surveys, $F(1.36, 388.15) = 17.41, p < .001$. A contrast indicated that the CDC-HRQOL ($M = 7.49, SD = 3.91$) was rated reliably higher than the MDASI-IVR ($M = 6.87, SD = 3.84$), $t(286) = 2.80, p = .006, d = 0.16$, which was rated reliably higher than the web-based MDASI ($M = 6.36, SD = 3.62$), where $t(286) = 4.79, p < .001, d = 0.28$. The small effect sizes for both indicate that the difference between the means was not incredibly sizeable however there was a difference nonetheless.

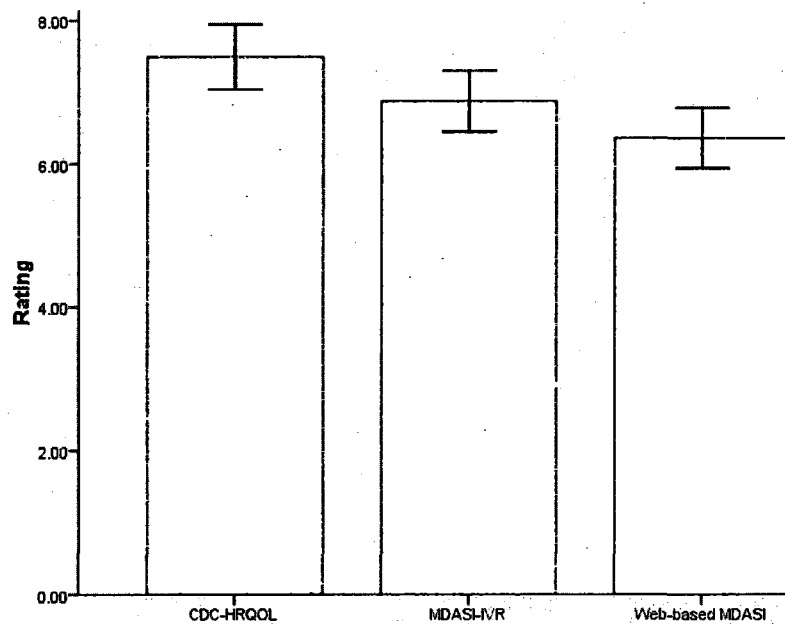


Figure 7 Rating comparison of CDC HRQOL, MDASI-IVR and web-based MDASI.
Note. Error bars = 95% CI.

Accuracy of response on the CDC-HRQOL

First it was determined if users accurately completed the CDC-HRQOL. Because it requested private health information, it was considered disclosure, in which case those who felt uncomfortable would have likely underreported on this web-based version of the CDC-HRQOL. The CDC has released state-wide and nation-wide means for some of the HRQOL items from its annual Behavioral Risk Factor Surveillance System survey (BRFSS; Prevalence Data, n.d.). Considering that the 2008-2009 undergraduate Rice population consisted of 51.0% Texans, Table 4 compared the Texas and nation-wide averages from the 18-24 year olds in 2007 to the obtained sample. The current sample appeared to have a higher reported average or percentage of unhealthy days on all items except for the overall health rating, where poor or fair was selected, and the percentage of physically unhealthy days, where 14 or greater days were selected. For these items, the current sample rated lower than the CDC state-wide and nation-wide population. This seemed to imply that there was accurate reporting, and that for the most part, the current sample had a lot of physical and mental distress compared to the CDC population, which could have been due to the heavy workload and end of semester burden faced by the students in the sample. In addition to examining the CDC-HRQOL scores to the archival data, Zullig (2005) utilized the CDC-HRQOL-4 on an undergraduate population. The data received from the 522 respondents very closely matched the current sample in positive skew, and the percentiles from Zullig near-mirrored those in the current sample (see Table 5).

The degree to which the ratings on the CDC-HRQOL corresponded to the MDASI-IVR ratings was also examined. A voice that was not well received over the IVR

Table 4: Comparison of the current sample's CDC-HRQOL scores to state- and nation-wide averages.

	Texas ^a		Nationwide ^b		Current Sample ^c	
	<i>M</i>	95% CI	<i>M</i>	95% CI	<i>M</i>	95% CI
Poor or Fair Health Rating (1-5)	12.2%	8.0 – 16.3	9.7%	8.6 – 10.8	6.3%	5.8 – 6.8
Physically Unhealthy Days	2.1	1.1 – 3.1	2.2	1.9 – 2.4	3.3	2.8 – 3.8
≥ 14 Days Physically Unhealthy	4.5%	1.2 – 7.8	5.1%	4.3 – 5.8	3.6%	2.8 – 4.4
Mentally Unhealthy Days	4.3	3.1 – 5.4	4.1	3.9 – 4.4	7.2	6.4 – 7.9
≥ 14 Days Mentally Unhealthy	13.3%	9.1 – 17.4	11.9%	10.9 – 12.9	14.7%	14.2 – 15.2
Activity Limitation Days	1.6	0.6 – 2.7	1.4	1.2 – 1.6	3.3	2.8 – 3.9
≥ 14 Days Activity Limitation	4.1%	0.5 – 7.6	3.4%	2.8 – 4.0	5.5%	5.4 – 5.6

Note. a: Texas $n = 664$; b: Nationwide $n = 14,872$; c: Sample $n = 287$. Days are rated from 0 to 30.

may have led a user to underreport symptoms on the MDASI-IVR which would have been accurately reported on the CDC-HRQOL. While the MDASI-IVR and written MDASI were very strongly positively correlated ($r(285) = .88, p < .001$), which makes sense given that they were the same instrument administered differently, the MDASI-IVR and the CDC-HRQOL were positively correlated as well, but the strength diminished slightly. Here, $r(285) = .50, p < .001$.

Table 5: Comparison of the current sample's CDC-HRQOL scores to Zullig (2005).

CDC-HRQOL Question	Response	<i>n</i>	% current study	% Zullig study
Self-perceived Health: "Would you say that in general your health is..."	Excellent	54	18.82	15.5
	Very Good	154	53.66	46.9
	Good	61	21.25	28.7
	Fair	16	5.57	5.9
	Poor	2	0.70	2.9
Physical Health: "Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?"	0 Days	66	23.00	20.5
	1 to 2 Days	91	31.71	35.1
	3 to 5 Days	92	32.06	23.8
	6 to 9 Days	13	4.53	9.8
	10 to 19 Days	20	6.97	7.0
	20 to 29 Days	2	0.70	1.9
	All 30 Days	3	1.05	2.1
Mental Health: "Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?"	0 Days	20	6.97	20.7
	1 to 2 Days	60	20.91	27.4
	3 to 5 Days	72	25.09	18.7
	6 to 9 Days	44	15.33	14.2
	10 to 19 Days	66	23.00	10.2
	20 to 29 Days	23	8.01	5.2
	All 30 Days	2	0.70	3.4
Activity Limitation: "During the past 30 days, for about how many days did poor physical or mental health keep you from doing your usual activities, such as self-care, work, or recreation?"	0 Days	93	32.40	40.4
	1 to 2 Days	75	26.13	30.1
	3 to 5 Days	65	22.65	13.4
	6 to 9 Days	22	7.67	6.9
	10 to 19 Days	25	8.71	3.8 ^a
	20 to 29 Days	5	1.74	2.9
	All 30 Days	2	0.70	2.5

Note. ^a: This percentage is corrected from Zullig (2005), where there is a misprint.

Reliability of the MDASI-IVR

Some users received the MDASI-IVR first, and some, second, in order to determine the reliability of the MDASI-IVR. In addition, the web-based version of the MDASI was given to users at the end of the experiment (after the SUS and CDC-

HRQOL) for those receiving the MDASI-IVR first and for those receiving the MDASI-IVR last, the web-based MDASI was given first. The stability of the MDASI across versions (IVR versus web-based) was also investigated. A mixed-design ANOVA indicated that there was a main effect of version, indicating that users rated their symptoms higher on the MDASI-IVR ($M = 2.44$, $SD = 1.29$) than on the MDASI ($M = 2.26$, $SD = 1.28$), $F(1, 285) = 23.39$, $p < .001$ (see Figure 8). Simple main effects revealed that ratings for the first presentation of the MDASI-IVR were higher than ratings for the first presentation of the web-based MDASI, where $t(145) = 6.19$, $p < .001$. This may imply that there was more disclosure on the IVR as symptom ratings were higher on this format. In addition, there was an interaction between version and presentation, indicating that not only did the user ratings differ across versions, but this was impacted by the

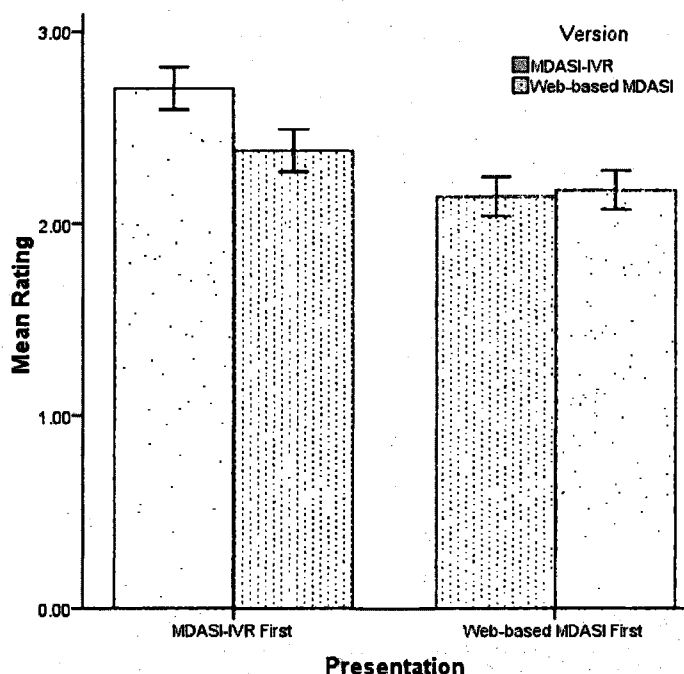


Figure 8 Comparison of ratings on MDASI depending on version (IVR or web-based) and presentation (IVR given first or second). *Note.* Error bars ± 1 SEM.

order of presentation (whether the MDASI-IVR was administered first or second), $F(1, 285) = 15.62, p < .001$. Investigating simple main effects, this revealed that MDASI-IVR scores were higher for users who completed it first ($M = 2.71, SD = 1.33$) versus last ($M = 2.16, SD = 1.20$), $t(285) = 3.64, p < .001, d = .43$, but MDASI scores did not differ regardless if users filled it out first ($M = 2.13, SD = 1.22$) or last ($M = 2.38, SD = 1.34$), $t(285) = 1.67, p = .097$.

There was no rating difference on the CDC-HRQOL depending on order of presentation, $t(285) = 0.15, p = .878$. Despite the different responses of the MDASI-IVR depending on presentation, the correlation between MDASI-IVR and CDC-HRQOL was not affected. When the MDASI-IVR was presented first, the correlation was $.46 (p < .001)$. When the MDASI-IVR was presented second, the correlation was $.58 (p < .001)$.

Usability of the MDASI-IVR

SUS scores averaged $79.79 (SD = 11.66)$, therefore usability of the MDASI-IVR was considered “good,” indicating that the system was acceptable to use (Bangor et al., 2008). As expected, the mean dropped for individuals who took more than one attempt to complete the IVR ($n = 32, M = 76.88, SD = 12.35$) compared to those who finished in one try ($n = 255, M = 80.16, SD = 11.55$), however these two groups did not reliably differ from each other, $t(285) = 1.50, p = .134$. Excluding users who took more than one attempt to complete the MDASI-IVR, there was no difference in SUS score for those who took the IVR before or after the web-based version, $t(253) = 1.49, p = .138$.

Interestingly, there was an effect of IVR gender on SUS scores, whereby users rated usability of the MDASI-IVR higher if they had listened to a male voice ($M = 81.56, SD = 10.92$) than to a female voice ($M = 78.01, SD = 12.14$), $t(285) = 2.61, p = .01, d = .31$.

Experiment 2 Discussion

In Experiment 2, there were no rating differences based on the system voice heard for the MDASI-IVR. This implies that disclosure is not influenced by a particular voice personality or gender for the medical IVR investigated. Additionally, there were no differences in rating by user gender, indicating that customizing IVRs to users based on their gender would be an unnecessary step. Interestingly, there were no rating differences based on trust – in Experiment 2, all six voices were highly trusted, whereas the upbeat voices were not considered trustworthy in Experiment 1. One must question if disclosure would have been influenced had a voice been rated as untrustworthy.

The positive correlations between the IVR and the CDC-HRQOL imply that users were rating their health consistently across the two health surveys. Users additionally filled out a web-based MDASI, and their responses on this format were reliably lower than they were on the MDASI-IVR. This may have indicated mild underreporting of symptoms on the web-based format. As users completed the experiment away from the lab, a shared dorm room or a public computer lab may have been utilized. In this case, privacy may have been compromised on the web portion as others could have been able to monitor users' responses. On the other hand, the DTMF input of the IVR would have maintained privacy regardless of location.

Finally, the acceptable rating on the SUS indicates that the usability of the MDASI-IVR was not in question. The SUS score was influenced by the system gender, such that users listening to the male voice rated the system as more usable than those listening to the female voice. However, while the scores for both the male-voiced system and female-voiced system still fell into the “good system usability” category (Bangor et

al., 2008), it does raise questions about how the system voice may be able to influence system usability.

4.0 General Discussion

It is clear from these experiments that separable exemplars can be captured. Additionally, disclosure is not influenced by a particular voice personality or gender for the medical IVR investigated. This stability in rating indicates that IVR designers do not have to concern themselves with locating a particular voice talent when seeking a voice for an IVR that may involve disclosure. Therefore, tailoring the IVR to male or female user groups would be an unnecessary step.

The IVR utilized does involve disclosure, as the MDASI-IVR inquires about symptom severity and interference related to numerous private topics (i.e., nausea, mood, and enjoyment of life). However, there are medical IVRs that require disclosure of more private topics such as alcohol and drug use or stigmatizing diseases such as AIDS or STDs. It is possible that the degree of disclosure in the MDASI-IVR would not compare to these other surveys. If this is the case, voice personality and/or gender may only have an effect on item response when the patient is greatly affected by disclosure.

This study was limited because the medical IVR focused on healthy users who disclosed symptoms related to physical and mental pain. Utilizing an unhealthy population such as those with a chronic illness could have given results that would be more representative of medical IVRs being used in the field. In addition, the sample consisted of undergraduates, which does not lend to the ability to validate across all age ranges of patients who may utilize an IVR. Older patients may report differently because of their satisfaction with the IVR itself. Dulude (2002) found that older users had

usability issues which affected their performance compared to younger users. A pleasant voice or a voice that is easier to comprehend by older users may increase user satisfaction via perceived usability. Lines and Hone (2002) found that elderly listeners had a preference for a male voice, but this does not imply that these individuals would alter their response rates depending on the gender heard over an IVR.

There is always a possibility that the users did not respond honestly because they knew they were in an experiment, despite the removal from an experimental setting. While testing the impact of system voice in a medical IVR in a natural setting could remove this possibility, it could impact patient health. One could also implement a voice that should drive symptom underreporting into the experimental setting. A rude or annoying voice, for example, could lead the user to hang up before the call completes or to enter erroneous data (e.g. enter all "1"s using barge-through to avoid hearing the system voice). Indeed, Rolandi (2007) noted that observation would reveal such behaviors in response to a particular persona. Rolandi was specifically commenting on "over-the-top" personae such as an extremely upbeat voice.

Finally, all of the voices utilized in Experiment 2 were rated highly on the degree to which the user trusted the voice, despite the fact that some (i.e., the upbeat voices) were rated lower overall in Experiment 1. It is unclear why female listeners did not trust the upbeat female voice in Experiment 1 but did trust the same voice in Experiment 2. It could be interesting to see if disclosure is impacted when users do not trust the IVR voice to which they are responding.

While these results were limited to a medical IVR, it is possible that there could be different findings with a more commercial IVR or product. For example,

customization is often utilized on GPS devices, where interaction with the product's voice is frequent. While the voice in a non-medical IVR will not affect disclosure of medical information, it has the potential to affect satisfaction with the product or the amount of time the user remains on the line and, in turn, product sales.

References

- Abu-Hasaballah, K., James, A., & Aseltine Jr., R. H. (2007). Lessons and pitfalls of interactive voice response in medical research. *Contemporary Clinical Trials, 28*, 593–602.
- Agel, J., Rockwood, T., Mundt, J. C., Greist, J. H., & Swiontkowski, M. (2001). Comparison of interactive voice response and written self-administered patient surveys for clinical research. *Orthopedics, 24*, 1155–1157.
- Alemagno, S. A., Frank, S., Mosavel, M., & Butts, J. (1998). Screening adolescents for health risks using interactive voice response technology: An evaluation. *Computers in Human Services, 15*, 27–37.
- Andresen, E. M., Fouts, B. S., Romeis, J. C., & Brownson, C. A. (1999). Performance of health-related quality-of-life instruments in a spinal cord injured population. *Archives of Physical Medicine and Rehabilitation, 80*, 877–884.
- Andresen, E. M., Catlin, T. K., Wyrwich, K. W., & Jackson-Thompson, J. (2003). Retest reliability of surveillance questions on health related quality of life. *Journal of Epidemiological Community Health, 57*, 339–343.
- Aronovitch, C. D. (1976). The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker. *The Journal of Social Psychology, 99*, 207–220.
- Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction, 24*, 574–594.

- Beck, R. S., Daughtridge, R., & Sloane, P. D. (2002). Physician-patient communication in the primary care office: A systematic review. *Journal of the American Board of Family Practice, 15*, 25–38.
- Bendapudi, N. M., Berry, L. L., Frey, K. A., Parish, J. T., & Rayburn, W. L. (2006). Patients' perspectives on ideal physician behaviors. *Mayo Clinic Proceedings, 81*, 338–344.
- Brooke, J. (1996). SUS: A “quick and dirty” usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland (Eds.), *Usability Evaluation in Industry* (pp. 189–194). London: Taylor and Francis.
- Bushey, R. R., Martin, J. M., & Joseph, K. M. (2001). Using the customer-centric approach to design interactive voice response systems. *Proceedings of the Human Factors and Ergonomics Society, Minneapolis/St. Paul, MN, 45*, 547–551.
- Christen, R. N., Alder, J., & Bitzer, J. (2008). Gender differences in physicians' communicative skills and their influence on patient satisfaction in gynaecological outpatient consultations. *Social Science & Medicine, 66*, 1474–1483.
- Cleeland, C. S., Mendoza, T. R., Wang, X. S., Chou, C., Harle, M. T., Morrissey, M., & Engstrom, M. C. (2000). Assessing symptom distress in cancer patients: The M. D. Anderson Symptom Inventory. *Cancer, 89*, 1634–1646.
- Corkrey, R., & Parkinson, L. (2002). A comparison of four computer-based telephone interviewing methods: Getting answers to sensitive questions. *Behavior Research Methods, Instruments, & Computers, 34*, 354–363.
- Couper, M. P., Singer, E., & Tourangeau, R. (2004). Does voice matter? An interactive voice response (IVR) experiment. *Journal of Official Statistics, 20*, 551–570.

- Dryer, D. C. (1999). Getting personal with computers: How to design personalities for agents. *Applied Artificial Intelligence, 13*, 273–295.
- Dulude, L. (2002). Automated telephone answering systems and aging. *Behavior and Information Technology, 21*, 171–184.
- Fendrich, M., & Vaughn, C. M. (1994). Diminished lifetime substance use over time: An inquiry into differential underreporting. *The Public Opinion Quarterly, 58*, 96–123.
- Forster, A. J., LaBranche, R., McKim, R., Faight, J. W., Feasby, T. E., Janes-Kelley, S., Shojanian, K. G., & van Walraven, C. (2008). Automated patient assessments after outpatient surgery using an interactive voice response system. *The American Journal of Managed Care, 14*, 429–436.
- Frank, A. P., Wandell, M. G., Headings, M. D., Conant, M. A., Woody, G. E., & Michel, C. (1997). Anonymous HIV testing using home collection and telemedicine counselling: A multicenter evaluation. *Archives of Internal Medicine, 157*, 309–315.
- Hura, S. L. (2008). Designing usable voice user interfaces. In P. Kortum (Ed.), *HCI Beyond the GUI: The Human Factors of Non-Traditional Interfaces*. Burlington, MA: Morgan Kaufman.
- 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. *Behavior and Information Technology, 16*, 125–144.

- Kim, H., Bracha, Y., & Tipnis, A. (2007). Automated depression screening in disadvantaged pregnant women in an urban obstetric clinic. *Archives of Women's Mental Health, 10*, 163–169.
- Klie, L. (2007). It's a persona, not a personality. *Speech Technology, 12*, 22–26.
- Knott, B. A., & Kortum, P. (2006). Personification of voice user interfaces: Impacts on user performance. *Proceedings of the Human Factors and Ergonomics Society, San Francisco, CA, 50*, 599–603.
- Krysan, M. (1998). Privacy and the expression of white racial attitudes: A comparison across three contexts. *The Public Opinion Quarterly, 62*, 506–544.
- Lee, E-J., Nass, C., & Brave S. (2000). Can computer-generated speech have gender? An experimental test of gender stereotype. *CHI '00 Extended Abstracts on Human Factors in Computing Systems, The Hague, Netherlands*, 289–290.
- Lines, L. & Hone, K. S. (2002). Older adults' evaluations of speech output. *Assets, Edinburgh, Scotland*, 170–177.
- Mast, M. S., Hall, J. A., & Roter, D. L. (2007). Disentangling physician sex and physician communication style: Their effects on patient satisfaction in a virtual medical visit. *Patient Education and Counseling, 68*, 16–22.
- Mielenz, T., Jackson, E., Currey, S., DeVellis, R., & Callahan, L. F. (2006). Psychometric properties of the Centers for Disease Control and Prevention Health-Related Quality of Life (CDC HRQOL) items in adults with arthritis. *Health and Quality of Life Outcomes, 4*, 66.
- Moeller, M. & Bort, J. (1993). VM gets the message. *Communications International, 20*, 14–15.

- Moriarty, D. G., Zack, M. M., & Kobau, R. (2003). The Centers for Disease Control and Prevention's Healthy Days Measures – Population tracking of perceived physical and mental health over time. *Health and Quality of Life Outcomes, 1*, 37.
- Moskowitz, J. M. (2004). Assessment of cigarette smoking and smoking susceptibility among youth: Telephone computer-assisted self-interviews versus computer-assisted telephone interviews. *Public Opinion Quarterly, 68*, 565–587.
- Nass, C., Foehr, U. G., & Somoza, M. (2001). The effects of emotion of voice in synthesized and recorded speech. *Proceedings of the AAAI Symposium Emotional and Intelligent II: The Tangled Knot of Social Cognition. North Falmouth, MA.*
- Nass, C. I., Moon, Y., Morkes, J., Kim, E-Y., & Fogg, B. J. (1997). Computers are social actors: A review of current research. In B. Friedman (Ed.), *Human Values and the Design of Computer Technology* (pp. 137–162). New York, NY: Cambridge University Press.
- Nass, C., Robles, E., Heenan, C., Bienstock, H., & Trienen M. (2003). Speech-based disclosure systems: Effects of modality, gender of prompt, and gender of user. *International Journal of Speech Technology, 6*, 113–121.
- Oksenberg, L., Coleman, L., & Cannell, C. F. (1986). Interviewers' voices and refusal rates in telephone surveys. *Public Opinion Quarterly, 50*, 97–111.
- Olive, J. P. (1999). The voice user interface. *IEEE Global Telecommunications Conference, 4*, 2051–2055.
- Prevalence Data (n.d.). Retrieved December 29, 2008, from <http://apps.nccd.cdc.gov/HRQOL/index.asp>

- Quirk, M., Mazor, K., Haley, H-L, Philbin, M., Fischer, M., Sullivan, K., & Hatem, D. (2008). How patients perceive a doctor's caring attitude. *Patient Education and Counseling, 72*, 359–366.
- Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. New York, NY: Cambridge University Press.
- Richardson, L. C., Wingo, P. A., Zack, M. M., Zahran, H. S., & King, J. B. (2008). Health-related quality of life in cancer survivors between ages 20 and 64 years: Population-based estimates from the Behavioral Risk Factor Surveillance System. *Cancer, 112*, 1380–1389.
- Reidel, K., Tamblyn, R., Patel, V., & Huang, A. (2008). Pilot study of an interactive voice response system to improve medication refill compliance. *BMC Medical Informatics and Decision Making, 8*, 46.
- Rolandi, W. (2007). The persona craze nears an end. *Speech Technology, 12*, 9.
- Schattner, A., Rudin, D., & Jellin, N. (2004). Good physicians from the perspectives of their patients. *BMC Health Services Research, 4*, 26.
- Simpson, T. L., Kivlahan, D. R., Bush, K. R., & McFall, M. E. (2005). Telephone self-monitoring among alcohol use disorder patients in early recovery: a randomized study of feasibility and measurement reactivity. *Drug and Alcohol Dependence, 79*, 241–250.
- Stentiford, F. W. M., & Popay, P. A. (1999). The design and evaluation of dialogues for interactive voice response systems. *BT Technology Journal, 17*, 142–148.

- Tanke, E. D., & Leirer, V. O. (1994). Automated telephone reminders in tuberculosis care. *Medical Care*, *32*, 380–389.
- Tannen, D. (1994). *Gender and Discourse*. New York: Oxford University Press.
- Toll, B. A., Cooney, N. L., McKee, S. A., & O'Malley, S. S. (2005). Do daily interactive voice response reports of smoking behavior correspond with retrospective reports? *Psychology of Addictive Behaviors*, *19*, 291–295.
- Tourangeau, R., Couper, M. P., & Steiger, D. M. (2003). Humanizing self-administered surveys: Experiments on social presence in web and IVR surveys. *Computers in Human Behavior*, *19*, 1–24.
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*, 859–883.
- United States Department of Health & Human Services. (2003). *OCR Privacy Brief: Summary of the HIPAA Privacy Rule*. Retrieved December 20, 2008, from <http://www.hhs.gov/ocr/privacysummary.pdf>
- van Mulken, S., André, E., & Müller, J. (1998). The persona effect: How substantial is it? *Proceedings of HCI on People and Computers XIII*, 53–66.
- Wagner, J. J., Van der Loos, H. F. M., & Leifer, L. J. (2000). Construction of social relationships between user and robot. *Robotics and Autonomous Systems*, *31*, 185–191.
- Wang, P. S., Beck, A. L., McKenas, D. K., Meneades, L. M., Pronk, N. P., Saylor, J. S., Simon, G. E., Walters, E. E., & Kessler, R. C. (2002). Effects of efforts to increase response rates on a workplace chronic condition screening survey. *Medical Care*, *40*, 752–760.

- Ware Jr., J. E., Kosinski, M., & Keller, S. D. (1996). A 12-item short-form health survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care, 34*, 220–233.
- Witten, I. H., & Madams, P. H. C. (1977). The telephone enquiry service: a man-machine system using synthetic speech. *International Journal of Man-Machine Studies, 9*, 449–464.
- Wright, E. B., Holcombe, C., & Salmon, P. (2004). Doctors' communication of trust, care, and respect in breast cancer: Qualitative study. *British Medical Journal, 328*, 864.
- Zullig, K. J., Valois, R. F., Huebner, E. S., & Drane, J. W. (2004). Evaluating the performance of the Centers for Disease Control and Prevention Core Health-Related Quality of Life Scale with adolescents. *Public Health Reports, 119*, 577–584.
- Zullig, K. J. (2005). Using CDC's health-related quality of life scale on a college campus. *American Journal of Health Behavior, 29*, 569–578.

Appendix B

Sample Script

"For this survey, we will ask you to rate your symptoms in the last 24 hours. Rate each symptom from zero, meaning you did NOT have any symptom AT ALL, to 10, meaning the symptom was AS BAD as you can imagine it could be.

We will then ask you to rate how much your symptoms have interfered in the last 24 hours.

From zero to 10, rate your pain at its worst in the last 24 hours.

From zero to 10, rate your nausea at its worst in the last 24 hours."

Instructions for Voice Actors

As you do your voice, it is very important that the voices be perceived as "real", not contrived or fake. For example, if asked to do a newscaster, a "Ted Knight" voice would be inappropriate since it is really just a caricature of the real news voice presented by a professional like Walter Cronkite or Dan Rather.

No accents can be used, as this significantly complicates our understanding of the phenomenon we are studying.

Voice 1: Lively

Happy, outgoing, caring, interested, perky, optimistic, passionate about life
NOT: silly, teasing, sexy, excited

Voice 2: Professional

Even-keeled, dispassionate, matter-of-fact, somber, trustworthy
NOT: sarcastic, angry, depressed, unhappy, cold

Voice 3: Sympathetic

Sympathetic, compassionate, sincere, kind, warm, trustful
NOT: forced, sarcastic, excited, eager, whiny

Appendix C

Personality Scale

Please rate the voice you just heard along the dimensions below from 1 (Not at All) to 7 (Very Much)

	Not at All 1	2	3	4	5	6	Very Much 7
1. Professionalism	1	2	3	4	5	6	7
2. Perkiness	1	2	3	4	5	6	7
3. Sympathy	1	2	3	4	5	6	7
4. Happiness	1	2	3	4	5	6	7
5. Calmness	1	2	3	4	5	6	7
6. Compassion	1	2	3	4	5	6	7
7. Optimism	1	2	3	4	5	6	7
8. Enthusiasm	1	2	3	4	5	6	7
9. Sincerity	1	2	3	4	5	6	7
10. Extraversion	1	2	3	4	5	6	7
11. Trust	1	2	3	4	5	6	7
12. Cheerfulness	1	2	3	4	5	6	7
13. Simplicity	1	2	3	4	5	6	7
14. Kindness	1	2	3	4	5	6	7
<hr/>							
15. Overall, how much did you like this voice?	1	2	3	4	5	6	7

Appendix D

MDASI-IVR Script

Prompt 1¹: Please enter your participant number, followed by the pound key.

Prompt 2: How you feel is very important to us. The purpose of the Symptom Monitor is to help you report your symptoms to your care team at any time. During each call, we will ask you about symptoms you may have experienced in the last 24 hours.

Prompt 3: After listening to each question, use the numbers on your telephone to enter your answers. For example, to enter the number 1, press 1 on your telephone. To enter 2, press 2. To enter 10, press the 1 and then the 0. And so on. The spoken instructions will tell you what type of response is expected.

Prompt 4: First, we will ask you to rate your symptoms in the last 24 hours. Rate each symptom from 0, meaning you did NOT have the symptom AT ALL, to 10, meaning the symptom was AS BAD as you can imagine it could be.

Prompt 5: We will then ask you to rate how much your symptoms have interfered in the last 24 hours. Answer each interference question in the same way you rated each symptom – from 0, meaning your symptoms did NOT interfere AT ALL, to 10, meaning your symptoms interfered COMPLETELY.

Prompt 6: The system will give you plenty of time to enter your responses, so you don't need to rush. If you want to listen to a question again, wait at the end of the question and

¹ Prompt 1 asks patients for their patient ID and date of birth in the Original MDASI-IVR. It was changed because this degree of information was unnecessary and the vocabulary of "participant ID" was more familiar to the users in this study. In addition, the prompt following Prompt 1, "Welcome to the Community Cancer Care Symptom Monitor. If you are experiencing severe symptoms of any kind, please hang up and call your doctor or an emergency room immediately.", for two reasons. Users were not cancer patients, and we did not want to concern them with believing we were screening for cancer. Second, we were not in the position to give medical advice, which might have been perceived with the second portion of the prompt.

it will repeat automatically. Or, you may press the “STAR” key to repeat the current question.

Prompt 7: If you are ready to begin the questions, press 1 now. If you would like to listen to the instructions again, press 2.

Prompt 8²: People who do not feel well frequently have symptoms that are caused by being sick or injured. Now we will ask you to rate how severe your symptoms have been in the last 24 hours. Rate each symptom from 0 (meaning you did NOT have the symptom at all) to 10 (meaning the symptom was AS BAD as you can imagine it could be).

Prompt 9: From 0 to 10, rate your PAIN at its worst in the last 24 hours.

Prompt 10: From 0 to 10, rate your FATIGUE, or TIREDNESS at its worst in the last 24 hours.

Prompt 11: Rate your NAUSEA at its worst in the last 24 hours.

Prompt 12: Your DISTURBED SLEEP, at its worst.

Prompt 13: Your feelings of being DISTRESSED or UPSET, at its worst.

Prompt 14: Your SHORTNESS OF BREATH, at its worst.

Prompt 15: Your problem with REMEMBERING THINGS, at its worst.

Prompt 16: Your problem with LACK OF APPETITE, at its worst.

Prompt 17: Your feeling DROWSY or SLEEPY, at its worst.

Prompt 18: Your having a DRY MOUTH, at its worst.

Prompt 19: Your feeling SAD, at its worst.

Prompt 20: Your VOMITING, at its worst.

² The first sentence of Prompt 8 was altered. The original reads “People with cancer frequently have symptoms that are caused by their disease or by their treatment”. Because the users were not a cancer population, the vocabulary was altered.

Prompt 21: Your NUMBNESS or TINGLING, at its worst.

Prompt 22: Your MOUTH PAIN, at its worst.

Prompt 23: Your THROAT PAIN, at its worst.

Prompt 24: Now we will ask you to rate how much your symptoms have interfered in the last 24 hours. Rate each item from 0 (meaning your symptoms DID NOT INTERFERE AT ALL) to 10 (meaning your symptoms INTERFERED COMPLETELY).

Prompt 25: From 0 to 10, rate how much your symptoms have interfered with your GENERAL ACTIVITY in the last 24 hours.

Prompt 26: With your MOOD in the past 24 hours.

Prompt 27: With your WORK, INCLUDING WORK AROUND THE HOUSE.

Prompt 28: With your RELATIONS WITH OTHER PEOPLE.

Prompt 29: With your WALKING.

Prompt 30: With your ENJOYMENT OF LIFE.

Prompt 31³: This completes today's call. Thank you for reporting your symptoms.

Prompt 32: Thank you and good bye.

³ Prompt 31 had a sentence after the existing script which read, "Remember, if you are experiencing severe symptoms of any kind, call your doctor or an emergency room immediately." Once again, this portion was removed as the survey was not being overseen by physicians and we were not administering medical advice.

Appendix E

System Usability Scale

© Digital Equipment Corporation, 1986.

	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	1	2	3	4	5
2. I found the system unnecessarily complex	1	2	3	4	5
3. I thought the system was easy to use	1	2	3	4	5
4. I think that I would need the support of a technical person to be able to use this system	1	2	3	4	5
5. I found the various functions in this system were well integrated	1	2	3	4	5
6. I thought there was too much inconsistency in this system	1	2	3	4	5
7. I would imagine that most people would learn to use this system very quickly	1	2	3	4	5
8. I found the system very cumbersome to use	1	2	3	4	5
9. I felt very confident using the system	1	2	3	4	5
10. I needed to learn a lot of things before I could get going with this system	1	2	3	4	5

Appendix F

CDC-HRQOL-9

Healthy Days Core Module (CDC HRQOL-4)

1. Would you say that in general your health is:
 - a. Excellent
 - b. Very good
 - c. Good
 - d. Fair
 - e. Poor

2. Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?

3. Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?

4. During the past 30 days, for about how many days did poor physical or mental health keep you from doing your usual activities, such as self-care, work, or recreation?

Healthy Days Symptoms Module

1. During the past 30 days, for about how many days did PAIN make it hard for you to do your usual activities, such as self-care, work, or recreation?

2. During the past 30 days, for about how many days have you felt SAD, BLUE, or DEPRESSED?

3. During the past 30 days, for about how many days have you felt WORRIED, TENSE, or ANXIOUS?

4. During the past 30 days, for about how many days have you felt you did NOT get ENOUGH REST or SLEEP?

5. During the past 30 days, for about how many days have you felt VERY HEALTHY AND FULL OF ENERGY?