A Comparison of Broad Versus Deep Auditory Menu Structures

Patrick M. Commarford, IBM, Software Group, Louisville, Kentucky, **James R. Lewis,** IBM, Software Group, Boca Raton, Florida, and **Janan Al-Awar Smither** and **Marc D. Gentzler,** University of Central Florida, Orlando, Florida

Objective: The primary purpose of this experiment was to gain a greater understanding of the utilization of working memory when interacting with a speech-enabled interactive voice response (IVR) system. **Background:** A widely promoted guideline advises limiting IVR menus to five or fewer items because of constraints of the human memory system, commonly citing Miller's (1956) paper. The authors argue that Miller's paper does not, in fact, support this guideline. Furthermore, applying modern theories of working memory leads to the opposite conclusion – that reducing menu length by creating a deeper structure is actually more demanding of users' working memories and leads to poorer performance and satisfaction. **Method:** Participants took a working memory capacity test and then attempted to complete a series of e-mail tasks using one of two IVR designs (functionally equivalent, but one with a broad menu structure and the other with a deep structure). **Results:** Users of the broad-structure IVR performed better and were more satisfied than users of the deep-structure IVR. Furthermore, this effect was more pronounced for those with low working memory capacity. **Conclusion:** Results indicate that creating a deeper structure is more demanding of working memory resource than the alternative of longer, shallower menus. **Application:** This experiment has important practical implications for all systems with auditory menus (particularly IVRs) because it provides empirical evidence refuting a widely promoted design practice.

INTRODUCTION

Interactive Voice Response Systems

Automated phone-based user interfacing systems known as *interactive voice response* (IVR) systems enable users to accomplish many goals without the help of a human representative. User acceptance of IVRs, however, has been slow; many hold negative attitudes toward the technology, characterize IVRs as difficult to use, and resent being routed to a machine rather than a human. To be a successful replacement for a human operator, an IVR must be designed to allow users to accomplish their goals effectively and efficiently.

IVR input methods. There are two primary methods of input for IVRs: speech and dual-tone multiple frequency (DTMF; also referred to as *touch-tone* or *keypad input*). These two types of interfaces impose different requirements on users' cognitive resources because speech-enabled IVRs often offer predictable inputs, require users to store half the information (no pairing of a function to a key), and have stimulus-response compatibility advantages (Brainard, Irby, Fitts, & Alluisi, 1962; Wickens, Sandry, & Vidulich, 1983; Wickens, Vidulich, & Sandry-Garza, 1984). *Barge-in* refers to a widely adopted, speech-enabled IVR setting that enables users to interrupt the system and provide speech input at any time. In this article, we focus on cognitive resources and strategies associated with speech-enabled, barge-in-enabled IVR use.

Interactive voice response menu length – current design guidelines. Researchers and speech user interface designers (e.g., Balentine & Morgan, 2001; Cohen, Giangola, & Balogh, 2004; Gardner-Bonneau, 1992; Schumacher, Hardzinski, & Schwartz, 1995) advise that IVR menus must

Address correspondence to Patrick M. Commarford, 1019 Willow Springs Dr., Louisville, KY 40242; commarfo@us.ibm.com.
HUMAN FACTORS, Vol. 50, No. 1, February 2008, pp. 77–89. DOI 10.1518/001872008X250665.

be relatively short because of constraints of the human memory system. These individuals generally cite Miller's (1956) paper to support their claims, stating that humans simply cannot remember more than 7 ± 2 items. For example, Cohen et al. (2004) suggested limiting menus to three or four items. In agreement, Gould, Boies, Levy, Richards, and Schoonard (1987) as well as the Voice Messaging User Interface Forum (1990) advocated no more than four options per menu. Marics and Engelbeck (1997) also stated that menus should be limited to four or fewer items but advised that items such as Help and Exit should be excluded from this count. Although some (e.g., Gardner-Bonneau, 1999; Schumacher et al., 1995) have recognized lists (e.g., a movie list) as exceptions, there is widespread agreement in the speech user interface community that IVR menus should be short to avoid overtaxing users' memories.

Note that it is the designer's role to determine how best to allow users access to a specified set of options. Although it is an important factor, menu length is not the only factor. It is also important for designers to provide unambiguous menu labels and to place menu items into logical groups to meet user expectations and avoid confusion. If the items fall nicely into groups of four or fewer, it is reasonable to organize them in this manner. The IVR design question under consideration (whether long menus have user performance advantages compared with sets of shorter menus) becomes critical when more than a few items are relevant (potentially useful) at a particular point in the user interface flow.

Working Memory Theory

Miller (1956) cited experiments by Hayes (1952) and Pollack (1953) that indicate that the amount of information available for immediate recall increases substantially as the amount of information per item increases. As the amount of information per item increases, the number of items available for immediate recall is attenuated, but not linearly. Miller's paper had great influence in that it pointed out that memory span is, on average, between five and nine items for most "chunks" of information. Waugh and Norman (1965) theorized a two-component memory system, and Atkinson and Shiffrin (1968, 1971) described a three-component system; each pair of researchers described a component that can store a limited amount of information for short periods.

Baddeley and Hitch (1974) proposed a working memory system that is far more active and complex than those described by their predecessors. Their original model proposed a central executive that controls and monitors one's attention, as well as two slave systems used to store and rehearse verbal and visuospatial information. Baddeley and Hitch's (1974) original model was revolutionary in that it stipulated that the working memory system is responsible for the storage and processing of information.

Just and Carpenter's (1992) capacity theory also views working memory as an active system that is responsible for more than simple storage. They conceptualize *activation* as a single commodity that allows storage, retrieval, and computing and *capacity* as the amount of activation available. This model differs from Baddeley's (2000, 2001) model in that it does not propose discrete systems and components but instead favors an activation pool that can be used for such processes as information storage and computation.

Working Memory Measurement

Traditional short-term memory capacity tests require participants to perceive a series of digits or words and then repeat these items once the stimulus is no longer present (Reisberg, 1997). As a participant completes each set, the experimenter presents larger and larger sets until the participant starts making mistakes. This method, referred to as a *short-term memory span task,* has provided evidence that, in general, users can hold five to nine items in working memory (e.g., Hayes, 1952; Keller, Cowan, & Saults, 1995; Pollack, 1953; Smyth, Pearson, & Pendleton, 1988; Watkins, 1977).

More recent theories (e.g., Baddeley & Hitch, 1974) view working memory as an active system that is responsible for more than simple storage. The working memory system is now assumed to be responsible for directing attention resources and processing information. Many activities in which humans engage require both storage and processing capabilities. Traditional tests of memory capacity (memory span tests) measure storage capacity only and do not account for these types of memory processes. Therefore, as theory developed, researchers began developing measures of working memory capacity (WMC), which take these processes into account.

Daneman and Carpenter (1980) introduced the first complex span task, the *reading span task.* This task combines a traditional memory span task with additional processing demands – requiring participants to read a set of unrelated sentences and then attempt to recall the final word in each sentence. Engle, Nations, and Cantor (1990) developed another complex span task, the *operation span task,* which required participants to calculate math equations while storing a set of words. Unsworth, Schrock, Heitz, and Engle (2003) developed an automated, computer-administered version of the operation span task. Barret, Tugade, and Engle (2004) found that individuals perform consistently across a variety of complex span tasks that require different types of computations to be made. These types of complex tasks are more valid than traditional span tasks because they are closer in nature to tasks performed by working memory systems on a daily basis (Reisberg, 1997). Furthermore, these tasks correlate with reading comprehension scores on standardized tests (Baddeley, Logie, & Nimmo-Smith, 1985; Reisberg, 1997; Shah & Miyake, 1996; Turner & Engle, 1989) as well as measures of fluid intelligence (Engle, Tuholski, Laughlin, & Conway, 1999) and attentional control (Kane, Bleckley, Conway, & Engle, 2001).

Working Memory Theory and Measurement Techniques – Conclusions

Although there is debate regarding the specific underlying mechanisms involved, most researchers agree that there is an upper limit to the amount of information an individual can hold in short-term or working memory. Depending on context, five to nine items is a good estimate of the upper boundaries of simple memory span. However, modern theories of working memory (e.g., Baddeley, 2001; Just & Carpenter, 1992) support a system that goes beyond simple short-term information storage, suggesting that working memory capacity should be measured by the ability to both store and process information. When attempting to simultaneously operate on information, individuals can store a smaller amount.

Reducing the Number of Menu Items

Often, more than a few options are relevant at a particular point in an IVR system. Providing all these items at such a point helps to ensure that the system matches users' mental models, allowing for predictability of items and reducing confusion that could result from unfulfilled expectations. However, as described previously, the traditional belief is that there is a need to limit the number of items to avoid overtaxing users' working memory systems. The primary method of splitting a long menu into sets of shorter menus is to divide the items into two or more groups and provide access to these groups via higher-level menu items (Paap & Cooke, 1997). This method reduces the number of items in each menu but makes the menu deeper hierarchically.

Working Memory Theory and IVR Use

Many believe that the ability to immediately recall all items in each menu is key to speech interface use. However, when analyzing the requirements for selecting a menu item from a list, recall of all menu items does not seem essential. Instead, it seems perfectly reasonable for users to discard menu items that do not match their goals as these items are presented. If it is necessary, however, for users to hold all menu items in working memory, then as the number of items in an IVR menu increases, performance and user satisfaction ratings should decrease.

We propose that working memory is used during IVR tasks in a manner such that the use of a broad menu as opposed to a deep structure will not increase the demand on or overtax one's working memory. Users access an IVR with a goal and then search for the target options that will help them accomplish that goal. Users do not need to recall all items in a menu; they need only maintain one or two items in working memory before making a selection, regardless of menu length. Furthermore, separating the items into smaller sets of menus does not reduce the total number of options; it just increases complexity.

We propose that users select a "best of" item and hold this item in working memory. They process each new item as it is presented and either discard it or replace the "best of" item with the new item. They do this until they are confident that the current "best of" item will help them accomplish their goal or until the menu ends. When either of these conditions occurs, the user makes a selection. This strategy requires that users hold up to two items in working memory at any given moment. More accurately, it requires that the user hold one item (the current "best of"

item) and process information about another (the menu item under evaluation).

Figure 1 graphically depicts this theory of user interaction with interactive voice response systems. There may be instances in which users hold two approximately equal candidate items in working memory while they process the additional items. In these cases, when users reach the end of the list, they may return to reexamine the candidates before making a selection. However, in general, users will not carry two candidates; instead, they will make a quick comparison as each item is presented and drop the less attractive item.

Note that our model is largely consistent with MacGregor, Lee, and Lam's (1986) criterion-based decision model. Their model explains user selection behavior for serially presented lists such that individuals create low and high criterion levels. Any option that falls beneath the low criterion is immediately rejected. Any option that falls above the low criterion but below the high is considered a candidate and is maintained in working memory. Finally, any item that falls above the high criterion is immediately selected, resulting in a self-terminating search. If, after all items are presented, only a single candidate exists, the individual will choose that item, resulting in an exhaustive search.

When an individual must choose a single candidate and has encountered multiple items that fall between the low and high criteria, he or she will often revisit the candidates before making a selection (a redundant search). Paap and Cooke (1997) suggested that, alternatively, the individual might simply select the best of the candidates without revisiting them (a second route to an exhaustive search).

The main differences between our model and the criterion-based decision model are in terms of user behavior for items that are not selected immediately upon presentation. Our model assumes that all other items are candidates until supplanted by a better match and that, upon presentation of a stronger candidate, inferior items are discarded even if they surpass a minimum threshold. Therefore, our model does not allow the list of candidate items to grow beyond two and thus has different implications for working memory demand.

Given that holding one or two menu items while processing another is clearly within the capabilities of the human working memory system (Daneman & Carpenter, 1980; Hayes, 1952; Kane et al., 2004; Pollack, 1953), our theory of IVR menu use proposes that selecting an item from a single auditory menu should not overtax users' working memory systems. If users discard items

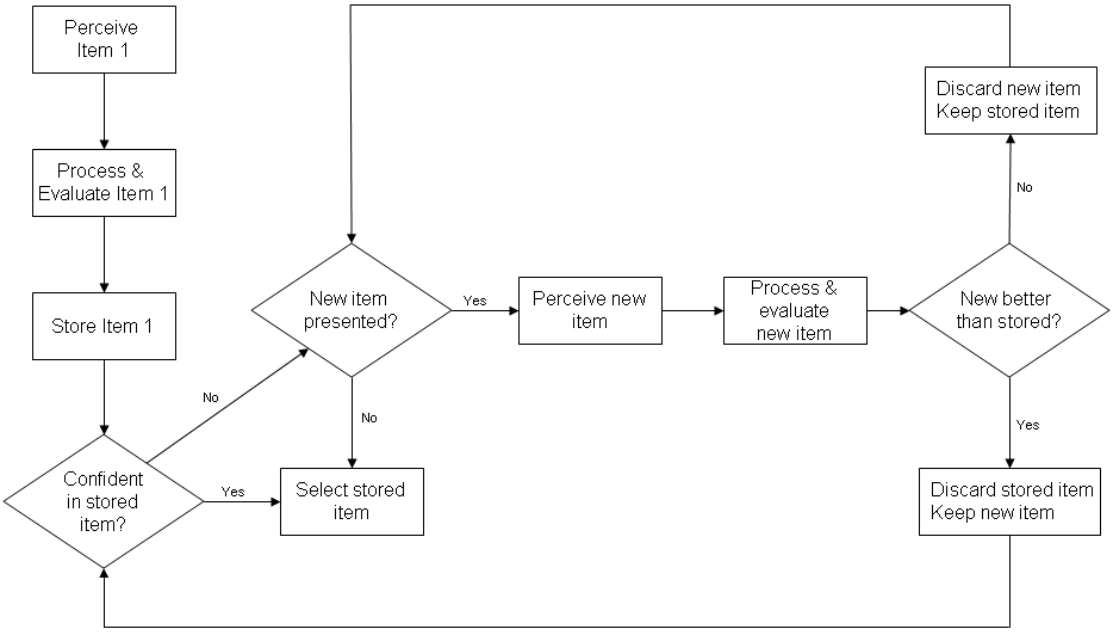

Figure 1. User flow for single-item selection from a single menu. The user holds one item (the current "best of" item) and processes information about another (the menu item under evaluation).

one by one, then increasing the number of menu items should not decrease performance or user satisfaction. In fact, as we will argue, splitting and artificially shortening IVR menus is more likely to decrease performance.

Our current knowledge of the human working memory system suggests that users will be more effective and more satisfied when using an IVR system that has a single long menu of appropriately grouped items than a system that artificially separates these items into multiple short menus with a deeper structure. There are several reasons for this. For example, the split scheme may be inconsistent with users' mental models, leading to unfulfilled expectations and mistakes. Also, splitting menus in this manner does not actually reduce the total number of menu items to be evaluated; the deep menu just makes it more difficult to access each of these items. In addition, the added need to navigate the user interface to access all options will decrease users' ability to maintain information in working memory because it requires them to engage in wayfinding activities, speak, listen to speech, and process information.

Furthermore, when users decide to select an item, they will often need to recall where the item resides as well as the command itself. Finally, deeper structures require additional user-system speech turns, which would increase time on task even if the user and system were perfect; because users commit speech errors and systems do not exhibit perfect recognition, the problem is exacerbated. The performance and satisfaction degradation experienced by users of the deeper structure should be greater for those with low working memory capacity because any system that stresses the limits of working memory will exceed the limits more often and to a greater degree for low-capacity individuals.

The Role of Target Location

The position of the target item within a menu is another important variable. When users are unable to predict (or remember from past experience) the appropriate speech input for the desired menu item, it takes longer, for example, to listen through a full, 10-item list than to listen only to the first two items. The common belief that long menus are too taxing of users' working memory either does not take target location into account or assumes that users must attempt to recall all items regardless of the position of the desired item.

Paap and Cooke (1997) advised the general use of broad rather than deep menu visual displays but cited three reasons why one might want to use a deeper structure. The first two reasons are not relevant to this topic, but the final reason – funneling – is relevant. Funneling allows the designer to provide shorter pathways to certain items (in general, those items that would appear in a late position of the broad menu). Though funneling provides an advantage such that pathways are shorter for late-position items, this design strategy also requires more user interaction to access these same items. Because increased user interaction can lead to an increase in time and errors (Paap & Cooke, 1997; Snowberry, Parkinson, & Sisson, 1983), the trade-off requires careful consideration.

Depending on a number of factors, a deeper structure could save or cost variable amounts of time for each menu item, given the item's position in a competing single, long menu. For these reasons, any investigation of the effect of menu structure on IVR usability should control for target item location. Good design places commonly chosen items near the front of the list and rarely chosen items at the end. Therefore, it is often the case that substantially fewer users will be affected by costs or savings associated with late-position items than with early-position items.

Individual Differences in Working Memory Capacity

The degree to which working memory is taxed by an IVR system's structure will depend, in part, on individual differences in WMC. Our experiment focuses on two competing theories of working memory involvement for speech-enabled IVR use; therefore, it is important to determine if participants with high WMC are affected by menu length and structure differently from those with low WMC. Specifically, we expect participants with lower WMC to be more sensitive to differences in IVR systems that increase the demands on users' working memories.

All IVR systems impose some demand on working memory, so users with low WMC should have more difficulty with an IVR system than those with high WMC – unless the system demands so little working memory resource that it does not approach the threshold of those on the low end. The traditional view suggests that achieving this low-demand criterion can be accomplished by

limiting the number of items in all interface menus to fewer than five (the low end of simple memory span). We propose that increasing menu length will not increase the load on users' memory. Therefore, if those with lower WMC have difficulty with an IVR, these difficulties will not be exacerbated by increased menu length. In fact, we hypothesize that complications associated with a deeper structure will strain working memory and degrade performance and satisfaction to a greater degree for users with low WMC.

IVR Menu Length – Research to Date

Huguenard, Lerch, Junker, Patz, and Kass (1997) investigated the effect of using a deeper versus a broader menu structure for touch-tone IVR systems. These authors determined that reducing the number of items per menu to three or fewer does not result in fewer errors.

Virzi and Huitema (1997) investigated selection times associated with broad versus deep menu structures for touch-tone IVR applications. Specifically, they tested touch-tone IVR systems with a single, eight-item top-level menu against identical systems with the top menu split such that the fifth item in the first set provided access to the final four items. They found that it took participants longer to make selections when the menu was split in this manner.

Using speech-enabled systems, Vanhoucke, Neeley, Mortati, Sloan, and Nass (2001) focused on determining whether certain prompting styles are better suited for broad versus deep menu structures. These studies addressed related topics, but to our knowledge, no researcher has empirically investigated the effects of implementing a broad menu design compared with sets of shorter menus with a deeper structure in a speech-enabled IVR system.

PHASE 1

Background and Purpose

We previously designed a voice portlet that allows users to access and manage their in-box by phone. After listening to each mail message, users are presented with a set of 8 to 11 e-mail navigation and management options. These options are listed below. Items in brackets may or may not be present in the menu, depending on the context. For example, "Next" does not play when the user has reached the last item in the list, and "Reply to all" does not play when the user is the only message recipient.

1. [Next]
2. [Previous]
3. Repeat
4. Delete
5. Reply
6. List recipients
7. [Reply to all]
8. Forward
9. Mark unread
10. Add sender
11. Time and date

In its current form, this menu (the most commonly encountered menu in the system) violates the commonly cited recommendation that speech IVR menus should have five or fewer items. The purpose of Phase 1 was to determine the most appropriate way to split this menu into separate menus, each containing five or fewer items.

Method

Participants. Twenty-six individuals with at least 3 months' experience using e-mail participated in this experiment. Most of the participants were employed at IBM® or at a local recruiting agency in South Florida. All participated on a volunteer basis.

Materials. We employed the automatic card-sorting and cluster analysis programs (respectively, Usort and EZCalc) developed by the user-centered design group at IBM. Each program ran on an IBM T41p Thinkpad® running Windows XP Professional®.

Procedure. We explained to participants that the purpose of the research was to determine the best way to organize menu items for a speech-based e-mail system. Participants read a brief description that clearly explained the action that occurs when each menu item is selected and then used the automatic card-sorting tool to place the 11 e-mail commands into groups of five or fewer. The Usort card sort tool provided each participant with a new, randomized order of item presentation.

Results

We analyzed the data using the average linkage algorithm provided by the EZCalc cluster analysis program. The program provides a visual representation of the participants' aggregated mental models produced via the analysis of the item distance matrix (see the appendix). The output

indicated that the new, higher level menu should be composed of four groups:

1. Delete, Forward, Reply, and Reply to All
2. Repeat, Next, and Previous
3. Mark Unread and Time and Date
4. List Recipients and Add Sender

PHASE 2

Phase 1 provided evidence that the best way to organize a speech-enabled e-mail system into a deeper structure with shorter menus was to group the 11 options into a new four-item menu. In Phase 2, we sought to determine the most appropriate labels to use in the new higher level menu. To accomplish this, we conducted a set of two Web-based user surveys. The first produced the label candidates for each group, and the second determined the final labels.

Method

Participants. We invited 1,000 IBM employees (all Lotus Notes® e-mail client users) to participate in Survey 1 and a separate group of 1,000 employees to participate in Survey 2. We received 101 sets of responses for Survey 1 and 155 for Survey 2.

Procedure. Survey 1 participants read a description of the function of each menu item, examined each menu item group, and suggested a label for each of the menus. Survey 2 participants read the description of the menu items, examined each menu item group, and selected, via multiple-choice format, the most appropriate label from the list of most common suggestions generated in Survey 1.

Results

Table 1 provides the most commonly suggested labels from Survey 1 for each of the menu groups, as well as the number of participants in Survey 2 who selected each as the best label.

The surveys yielded the following menu labels: "Listen to Messages," "Respond," "Distribution," and "Message Details." Figure 2 illustrates the final design derived from Phases 1 and 2.

PHASE 3

Method

Participants. All participants were undergraduate students at the University of Central Florida with at least 3 months of e-mail experience. The mean participant age was 20.3 years, and the range was 18 to 40. All received course credit for their participation in the study. No participants reported having a speech or hearing deficit.

Materials. In this experiment we employed two versions of a speech-enabled, barge-in-enabled e-mail voice application. Each of these voice applications was prepopulated with a set of e-mail messages for participants to access and act upon. After a message was played, each IVR offered 8 to 11 (depending on context) e-mail-related menu options. The broad version played all options immediately following each message in a single menu (Figure 3 illustrates this general design). The deep version split the menu items into four menus; the higher-level menu played immediately following a message (see Figure 2). From this

TABLE 1: Most Commonly Suggested Labels From Survey 1 and the Number (%) of Participants Who Selected Each Label in Survey 2

Item in Menu Group	Survey 1 Suggested Labels (#)	Survey 2 Number Selected
Next, Repeat, Previous	Navigate (22)	25 (17%)
	Listen to Messages (15)	130 (83%)
Reply, Reply to All, Forward, Delete	Action (22)	53 (34%)
	Respond (19)	102 (66%)
Add Sender, List Recipients	Address Book (9)	56 (36%)
	Distribution (9)	99 (64%)
Mark Unread, Time and Date	Message Details (10)	62 (40%)
	Status (10)	38 (25%)
	Miscellaneous (5)	15 (10%)
	Options (5)	40 (26%)

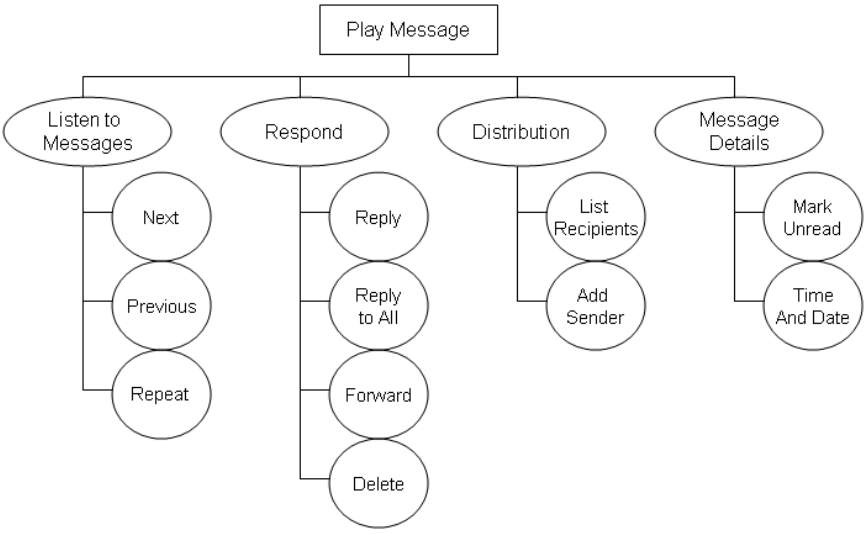

Figure 2. Deep, short menu interactive voice response (IVR) design with all options.

point on, these two IVRs will be referred to as the "broad" and "deep" versions, respectively. Each design, in general, ordered items from most frequently to least frequently selected, keeping highly related items in close proximity. The results of the cluster analysis conducted in Phase 1 made it possible for neighboring items to remain together and allowed the designs to present all 11 menu items in the same order with each design. Each system employed a female concatenative text-to-speech voice.

Participants also completed the automated operation span test created by Unsworth et al. (2003) to assess WMC. The authors demonstrated that this version of the operation span test is reliable and valid, correlating moderately with Turner and Engle's (1989) operation span test and showing high test-retest reliability. The test presents participants with an arithmetic operation (e.g., 2 * 3 – 1) followed by a proposed answer that may or may not be correct. The participant selects True or False, and then the system presents a letter to be retained for recall at the end of the trial. The participant answers a series of three to seven math problems followed by a letter and then attempts to recall the letters in their order of presentation. The test consists of a total of 75 combinations. The automated operation span score is the total number of letters that were recalled in their respective positions for perfectly recalled trials.

We also employed the Post-Study System Usability Questionnaire (PSSUQ; see Lewis, 1995, 2002). The revised version of the PSSUQ (Lewis, 2002) contains 16 items for which participants indicate their level of agreement on a 7-point Likert-type scale, with lower ratings indicating greater user satisfaction. We used audio editing software and a phone tap to record the experimental sessions. The IVR system was hosted at an IBM facility, and participants interacted with the system

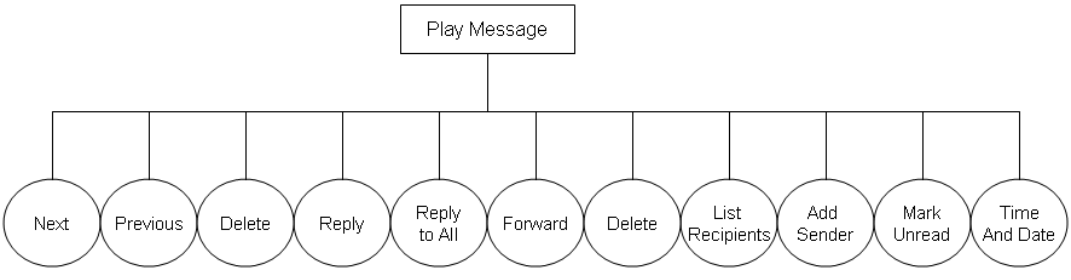

Figure 3. Long, broad menu interactive voice response (IVR) design with all options.

using a standard telephone handset. An audio booster enhanced the volume for recording purposes and was connected to a speaker, allowing the experimenter to hear the IVR system throughout each session.

Experimental design. We employed a 2 × 2 between-subjects design with three dependent variables. The independent variables were menu design (broad, deep) and WMC (low, high). The dependent variables were time, tasks completed, and PSSUQ scores. We categorized those who scored in the upper quartile on the WMC test as *high* and those who scored in the lower quartile as *low* (final determination of high and low criterions is described below). The experimenter assigned each participant to either the deep or broad menu group and to one of four task orders via a randomization process with constraints, following a predetermined sequence. This procedure ensured that approximately an equal number of participants from each WMC group would use the deep system as the broad system and that approximately the same number in each group worked with each of four task sequences. We created the four task orders using a random number table.

Procedure. First, participants completed the automated operation span task, which we used as a screening tool to eliminate those whose score fell into the middle third of the expected distribution. This allowed us to dismiss participants who clearly would not be in the top or bottom quartile of the final data set. We calculated the expected distribution based on data reported by Unsworth et al. (2003). Then, again using Unsworth et al. data, we dismissed all participants whose automated operation span score fell between 33 and 46. We also dismissed all participants who committed more than 20% math errors, because it was likely that they spent too much resource rehearsing letters at the expense of the operations tasks. In total, 121 participants completed the WMC test and 80 continued with the remainder of the experiment, as assigned per the prepared sequence.

Once qualifying participants completed the working memory span test, they began the IVR tasks (see Table 2). The experimenter explained that they would attempt to accomplish a set of e-mail management tasks using an automated speech-enabled phone application. The experimenter further instructed that they would have a maximum of 5 min to complete each task. Prior to beginning each of the seven IVR tasks, participants read the task and asked questions if they required clarification. After participants indicated that they understood the upcoming task, they dialed a number and accessed an in-box that they were to imagine was their own.

Completion of all tasks required utilization of all 11 menu items at least once. Therefore, to complete these tasks, users had to make early, middle, and late selections from the broad menu and selections from all positions within each menu of the deep structure. As previously mentioned, deeper structures require additional system-user speech turns. Because of the inclusion of the higher-level

TABLE 2: IVR Tasks (Participants Each Attempted in One of Four Randomly Generated Orders)

Task #	Task Description
1	Craig Marshall sent a note earlier asking if you had lunch plans. You've been waiting to see how long your 11:00 meeting was going to last and now you see that you won't be able to meet Craig for lunch. Access Craig's message and reply to him appropriately.
2	Access the message from Laura Harrington and mark it as an unread message so it will catch your attention when you access the system next from your PC.
3	Find the message from Joe Jacobs and then add him to your address book.
4	Your coworker, Ken Jeffries, sent you a message inviting you to a party at his place. You're interested in finding out which of your coworkers he also invited. Find the message and then check to see who else Ken sent the invite to.
5	Check your mail for any messages from Jon Cardo. What time did he send the note? Follow through with his request.
6	Find and delete the message from Mark Riverside.
7	Find the message with subject "Company Picnic" and reply to the sender and all recipients with the following message: "Count me in. I'll be there. See everyone this weekend."

menu, the minimum required number of user utterances to complete all tasks is nearly double for the deep structure (62 utterances) than for the shallow structure (33 utterances). Once participants indicated that they had completed each task or once the 5-min time interval had elapsed, they hung up the phone. After finishing all tasks, participants completed the PSSUQ.

Results

Working memory capacity. We removed 5 participants from the original data set of 121 before conducting the analyses. Two participants scored a zero, which indicates that they did not attempt to do well, and 3 participants committed too many math errors. After removal of these scores, the final data set included 116 scores. The mean score was 42.3, the median was 42.5, and the standard deviation was 14.96. The top quartile included participants who scored 52 and above and the bottom quartile included participants who scored 33 and below. This yielded 29 low- and 29 high-WMC participants and yielded 31 deep system users and 27 broad system users. Table 3 displays the sample size for each of the four cells.

Performance and satisfaction. We conducted a set of three two-factor analyses of variance (ANOVAs). The independent variables were menu and WMC, and the dependent variables were total time (total time to complete all tasks), complete (number of tasks successfully completed), and PSSUQ score. The analyses indicated that there was a main effect of menu for total time, $F(1, 54) = 67.55$, $p < .0005$, such that it took participants significantly longer to complete all tasks when using the deep system ($M = 1{,}358$ s) than the broad system ($M = 917$ s). There was also a main effect of menu for complete, $F(1, 54) = 35.14$, $p < .0005$, such that those using the deep system completed significantly fewer tasks ($M = 5.26$) than those using the broad system ($M = 6.70$).

TABLE 3: Distribution of the Number of Participants Among the Four Conditions

	Working Memory Capacity	
Menu Design	High	Low
Broad	13	14
Deep	16	15

The ANOVA revealed a final main effect of menu on PSSUQ, $F(1, 54) = 19.85$, $p < .0005$, such that those using the deep system indicated that they were significantly less satisfied with the system ($M = 4.17$) than were those using the broad system ($M = 2.64$).

There was a main effect of WMC: Those with high working memory capacity completed significantly more tasks, $F(1, 54) = 4.22$, $p = .045$ ($M = 6.17$), than did low-WMC participants ($M = 5.69$). There were no significant differences between high- and low-WMC users in terms of time to complete all tasks or satisfaction.

The ANOVA revealed a significant interaction between WMC and menu, $F(1, 54) = 4.53$, $p = .038$, for total time. Participants with high and low WMC completed tasks at approximately the same rate when using the broad system, $M_{diff} = 53$ s; 7.6 s per task, $t(25) = 0.64$; $p = .526$, but high-WMC participants were significantly faster when using the deep system, $M_{diff} = 176$ s; 25.1 s per task, $t(29) = 2.51$; $p = .018$. Figure 4 depicts this interaction. There was not a significant interaction for task completion rate ($p = .129$) or subjective ratings ($p = .459$).

GENERAL DISCUSSION

Citing Miller (1956), many authors warn that IVR menus should never contain more than five items. The reasoning is that because most individuals can serially recall an average of five to nine newly presented items, menus containing greater than five items will tax users' memories. The implied assumption is that it is necessary to remember all menu items in their presentation order to effectively use an IVR. We have argued that this is not, in fact, necessary to work efficiently with an IVR. Based on modern theories of working memory, we presented a model of IVR use that predicts that, regardless of menu length, users need store only one or two items while perceiving and evaluating another.

This article argues that when a group of items are all applicable at a particular point in an IVR system, splitting these items and creating a deeper menu structure will demand additional working memory resource. The results of this experiment support most of our hypotheses: Participants who used the broad menu structure significantly outperformed participants who used the deep structure and indicated significantly higher levels of

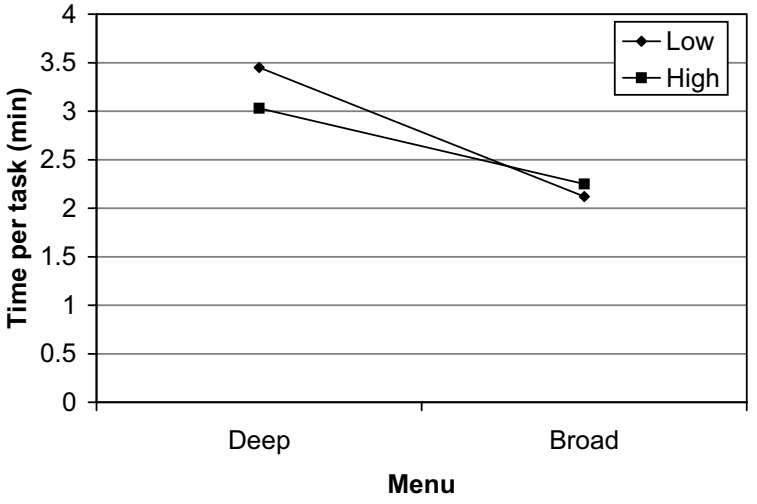

Figure 4. Time per task for high- and low-working memory capacity (WMC) participants using each system.

satisfaction. Furthermore, participants with low WMC were more negatively affected by the deep menu system than were participants with high WMC on total time to complete all tasks. These findings suggest that deep menu structures, rather than broad, are more demanding on users' working memory resource.

In addition to increased strain on working memory capacity, there are likely other attributes of an IVR system that has been artificially deepened and shortened that are detrimental to performance. For example, increasing depth increases the number of speech utterances a user must provide to navigate the IVR. This increases time on task and allows for additional error opportunities. Also, additional menu layers require participants to engage in more class inclusion searches. In this experiment, participants recognized most of the target commands in the broad system but often had trouble identifying which higher-level menu would provide access to these commands in the deep system.

Considering that other arguments can be made for why individuals using the broad system outperformed individuals using the deep system, the evidence that they were affected differentially based on their working memory capacity is particularly compelling. With the broad system, high-WMC participants did not outperform low-WMC users (they were actually 7.6 s per task slower); however, with the deep system, high-WMC participants did significantly outperform low-WMC participants (a difference of 25.1 s per task). This provides strong evidence that the deep menu significantly taxed users' working memories. As users build familiarity and expertise with an IVR, working memory demands will shrink; however, the broad menu design will still be superior because of the smaller number of required user-system interactions. Therefore, the broad design is superior to the deep design for both novice and expert users.

A potential limitation to this experiment is that we tested users only with mail navigation and management tasks. The high degree of familiarity of this domain to all participants set up a test scenario such that most menu options should be extremely clear and even predictable. It would be interesting to replicate this experiment using a domain with which participants are less familiar. Also, because all participants were drawn from the same university, the range of WMC scores is less variable than what would be expected from the entire IVR user population. If practical, it would be informative to replicate this experiment with a population that has greater variance in WMC. It might also be reasonable to consider including the entire WMC range in the study sample, rather than dismissing the middle quartiles, to ensure that the results can be generalized to the entire population under study.

Replicating this experiment with the older adult population, which is characterized by, among other limitations, reduced working memory capacity, would also be informative. This characteristic of the older population is indicative that applying a

broad design, rather than a deep structure, will be particularly beneficial to older users. However, other limitations characteristic of the older population (e.g., decreased speed of processing and hearing deficits) suggest that broad menus alone will likely be insufficient to ensure that an IVR is usable by older adults.

Note also that in this study, we did not explicitly test menus that are longer than 8 to 11 items and did not test barge-in-disabled systems. We have no reason to believe that increasing the number of items beyond 11 would yield different results (assuming the application of other leading design principles). Without having tested this, however, we have to be cautious in extrapolating our results. Systems that disable barge-in capability prevent users from selecting an option until the menu is complete. Disabling barge-in would not be likely to affect working memory demand but would create a frustrating experience for broad menu systems and would negate many of the previously discussed benefits of broad design. Therefore, we question the ability to generalize the results to barge-in-disabled systems without having empirically tested such systems.

Conclusion

This experiment provided evidence that, contrary to common belief, it can be advantageous to design an IVR system to use a broader structure with fewer long menus as opposed to a deeper structure with a greater number of shorter menus. These findings are consistent with predictions based on examination of modern theories of working memory and detailed analyses of phone-based tasks.

The experiment further provided evidence that intensive demand on working memory resource is one of the contributing factors to the performance detriment associated with a design that employs a hierarchical set of menus containing five or fewer items. This argument is supported by the interaction such that low-WMC participants expended similar amounts of time compared with high-WMC participants when using the broad menu system but expended significantly more time than high-WMC participants when using the deep menu system. This experiment has very important practical implications for all systems with auditory menus, and particularly for IVRs, because it provides empirical evidence that contradicts a widely promoted design practice.

APPENDIX

Menu Item Distance Matrix

Item	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Next (1)	.00	.00	.01	.37	.50	.50	.50	.44	.42	.50	.46
Previous (2)	.00	.00	.01	.37	.50	.50	.50	.44	.42	.50	.46
Repeat (3)	.01	.01	.00	.35	.44	.48	.44	.37	.44	.50	.44
Delete (4)	.37	.37	.35	.00	.46	.48	.27	.25	.38	.46	.48
Reply (5)	.50	.50	.44	.27	.00	.46	.00	.12	.44	.48	.48
List Recipients (6)	.50	.50	.48	.48	.46	.00	.46	.46	.40	.12	.27
Reply to All (7)	.50	.50	.44	.27	.00	.46	.00	.12	.44	.48	.48
Forward (8)	.44	.44	.37	.25	.12	.46	.12	.00	.44	.46	.50
Mark Unread (9)	.42	.42	.44	.38	.44	.40	.44	.44	.00	.40	.17
Add Sender (10)	.50	.50	.50	.46	.48	.12	.48	.46	.43	.00	.35
Time and Date (11)	.46	.46	.44	.48	.48	.27	.48	.50	.17	.35	.00

REFERENCES

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89–105). New York: Academic Press.

Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American*, *225*, 82–90.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*, 417–423.

Baddeley, A. D. (2001). Is working memory still working? *American Psychologist*, *56*, 851–864.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *Recent advances in learning and motivation* (Vol. 8, pp. 47–90). New York: Academic Press.

Baddeley, A., Logie, R., & Nimmo-Smith, I. (1985). Components of fluent reading. *Journal of Memory and Language*, *24*, 119–131.

Balentine, B., & Morgan, D. P. (2001). *How to build a speech recognition application: A style guide for telephony dialogs* (2nd ed.). San Ramon, CA: Enterprise Integration Group.

Barret, L. S., Tugade, M. M., & Engle, R. W. (2004). Individual differences in working memory capacity and dual-process theories of the mind. *Psychological Bulletin*, *130*, 553–573.

Brainard, R. W., Irby, T. S., Fitts, P. M., & Alluisi, E. (1962). Some variables influencing the rate of gain of information. *Journal of Experimental Psychology*, *63*, 105–110.

Cohen, M. H., Giangola, J. P., & Balogh, J. (2004). *Voice user interface design*. Boston: Addison-Wesley.

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

Engle, R. W., Nations, J. K., & Cantor, J. (1990). Is "working memory capacity" just another name for word knowledge? *Journal of Educational Psychology*, *82*, 799–804.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent variable approach. *Journal of Experimental Psychology: General*, *128*, 309–331.

Gardner-Bonneau, D. (1999). Guidelines for speech-enabled IVR application design. In D. Gardner-Bonneau (Ed.), *Human factors and voice interactive systems* (pp. 147–162). Boston: Kluwer Academic.

Gardner-Bonneau, D. J. (1992). Human factors in interactive voice response applications: "Common sense" is an uncommon commodity. *Journal of the American Voice I/O Society*, *12*, 1–12.

Gould, J. D., Boies, S. J., Levy, S., Richards, J. T., & Schoonard, J. (1987). The 1984 Olympics message system: A test of behavioral principles of system design. *Communications of the ACM*, *30*, 758–769.

Hayes, J. M. (1952, January–June). Memory span for several vocabularies as a function of vocabulary size. In *Quarterly Progress Report*. Cambridge: Massachusetts Institute of Technology, Acoustics Laboratory.

Huguenard, B. R., Lerch, F. J., Junker, B. W., Patz, R. J., & Kass, R. E. (1997). Working-memory failure in phone-based interaction. *ACM Transaction on Human-Computer Interaction*, *4*, 67–102.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122–149.

Kane, M. J., Bleckley, M. K., Conway, A. R., & Engle, R. W. (2001). A controlled-attention view of working memory capacity: Individual differences in memory span and the control of visual orienting. *Journal of Experimental Psychology: General*, *130*, 169–183.

Kane, M. J., Hambrick, D. Z., Tulholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*, 189–217.

Keller, T. A., Cowan, N., & Saults, J. S. (1995). Can auditory memory for tone pitch be rehearsed? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*, 635–645.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, *7*, 57–78.

Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction*, *14*, 463–488.

MacGregor, J., Lee, E., & Lam, N. (1986). Optimizing the structure of database menu indexes: A decision model of menu search. *Human Factors*, *28*, 387–399.

Marics, M. A., & Engelbeck, G. (1997). Designing voice menu applications for telephones. In M. G. Helander, T. K. Landauer, & P. V. Prabhu (Eds.), *Handbook of human-computer interaction* (pp. 1085–1102). Amsterdam: Elsevier.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, *63*, 81–97.

Paap, K., & Cooke, N. (1997). Design of menus. In M. G. Helander, T. K. Landauer, & P. V. Prabhu (Eds.), *Handbook of human-computer interaction* (pp. 533–572). Amsterdam: Elsevier.

Pollack, I. (1953). The assimilation of sequentially encoded information. *American Journal of Psychology*, *66*, 421–435.

Reisberg, D. (1997). *Cognition: Exploring the science of the mind*. New York: W. W. Norton.

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

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General*, *125*, 4–27.

Smyth, M. M., Pearson, N. A., & Pendleton, L. R. (1988). Movement and working memory: Patterns and positions in space. *Quarterly Journal of Experimental Psychology*, *40*, 497–514.

Snowberry, K., Parkinson, S., & Sisson, N. (1983). Computer display menus. *Ergonomics*, *26*, 699–712.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*, 127–154.

Unsworth, N., Schrock, J. C., Heitz, R. P., & Engle, R. W. (2003, May 29–June 1). *An automatic version of Ospan*. Poster presented at the annual meeting of the American Psychological Society, Atlanta, GA.

Vanhoucke, V., Neeley, W. L., Mortati, M., Sloan, M., & Nass, C. (2001). Effects of prompt style when navigating through structured data. In M. Hirose (Ed.), *Proceedings of INTERACT 2001: 8th TC13 IFIP International Conference on Human-Computer Interaction* (pp. 530–536). Amsterdam: IOS Press.

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

Voice Messaging User Interface Forum (1990). *Specification document*. Cedar Knolls, NJ: Probe Research.

Watkins, M. J. (1977). The intricacy of memory span. *Memory and Cognition*, *5*, 529–534.

Waugh, N. C., & Norman, D. A. (1965). Primary memory. *Psychological Review*, *72*, 89–104.

Wickens, C. D., Sandry, D., & Vidulich, M. (1983). Compatibility and resource competition between modalities of input, central processing, and output: Testing a model of complex task performance. *Human Factors*, *25*, 227–248.

Wickens, C. D., Vidulich, M., & Sandry-Garza, D. (1984). Principles of S-C-R compatibility with spatial and verbal tasks: The role of display-control location and voice-interactive display-control interfacing. *Human Factors*, *26*, 533–543.

Patrick M. Commarford is a user experience engineer at IBM, Software Group in Louisville, Kentucky. He received his Ph.D. in applied experimental psychology and human factors from the University of Central Florida in 2006.

James R. (Jim) Lewis is a consultant in the design of interactive voice response systems at IBM in Boca Raton, Florida. He received his Ph.D. in psycholinguistics from Florida Atlantic University in 1996.

Janan Al-Awar Smither is an associate professor of psychology at the University of Central Florida. She received her Ph.D. from Johns Hopkins University in 1985.

Marc D. Gentzler is a doctoral student in the applied experimental psychology and human factors program at the University of Central Florida. He received his master's degree with a concentration in organizational psychology from Claremont Graduate University in 2007.

Date received: May 22, 2007
Date accepted: October 1, 2007