Towards Accessible Automatically Generated Interfaces Part 2: Study with Model-Based Self-voicing Interfaces

J. Bern Jordan(✉) and Gregg C. Vanderheiden

University of Maryland, College Park, Md., USA
{jbjordan,greggvan}@umd.edu

Abstract. The automatic generation of personalized user interfaces is a potential strategy to enable people with disabilities to enjoy wider access to devices and systems. The Functionality Input Needs and User Sensible Input (FIN-USI) model was created to make it easier to both model the devices and functionality to be controlled and for the automatic generation of personalized, accessible user interfaces. In order to study the feasibility of the model, two basic interface generators were created that used the FIN-USI model and users' interactor preferences to generate self-voicing, mobile-device interfaces intended for people who are blind. The efficacy of the FIN-USI model and generators was then tested using 12 blind and 12 inexperienced, blindfolded participants. In the user study, participants' performance, preference, and satisfaction were measured and compared on four interfaces: two interfaces that were manufacturer created and two that were automatically generated from each user's preferences. All usability measures in the study were better significantly for the automatically-generated interfaces compared to the manufacturer-created ones, including a manufacturer-created interface specifically designed for people who are blind.

Keywords: Automatic UI generation · Personalization · User interface model · Screen reader · Self-voicing interfaces

1 Introduction

People with disabilities often have difficulty using mainstream user interfaces (UIs) because of a poor fit with their individual needs and constraints. The automatic generation of user interfaces is a potential solution to this problem [1]. With auto-generated interfaces, alternative user interfaces can be generated in a one-size-fits-one manner to fit individual needs and preferences that cannot be easily met through other strategies. The auto-generation of user interfaces is based on an underlying model of the interface or functionality to be controlled.

Many interface models have been created that could be used for generating interfaces, but there has been little practical progress towards real-world applications suitable for people with disabilities. Many of the models that have been developed are too complex for their potential benefit [2, 3]. Other models, such as those used in [1, 4] are simpler, but have been focused on limited types of interaction and input. In order to be useful and usable to industry and to make it beyond research to application, a model

J. Zhou and G. Salvendy (Eds.): ITAP 2017, Part I, LNCS 10297, pp. 110–128, 2017.
DOI: 10.1007/978-3-319-58530-7_8

needs to be both robust and easy-to-apply to many different products and functionality. It also should be validated for improved accessibility and shown to be useful to people with disabilities and others who need alternative user interfaces.

The FIN-USI model (detailed in Part 1, this volume) was developed to address the shortcomings of currently available models. It is intended to be simpler to apply to products and functionality than many of the current models. This would potentially make it easier for industry to adopt and build into their own products. The FIN-USI model also covers a wider range of input and interaction styles than other current models, which may allow it to be used for more devices and functionality.

Much of the user research to date has been around gathering user requirements and preferences that might be used in generating or adapting user interfaces (for example [5, 6]). However, relatively few user studies have been conducted to verify that personalized, auto-generated interfaces improve access for people with disabilities over the interfaces and accessibility strategies that are employed by industry today. User testing in the Personal Universal Controller (PUC) project showed a reduction of errors and an increase in user speed with automatically generated interfaces on touchscreen personal digital assistants (PDAs) for copiers, especially for interfaces that were generated to match an interface on which a person was previously trained [7]. The PUC graphical user interface generation system did not account for users' preferences, characteristics, or abilities. In the SUPPLE project, graphical user interfaces could be generated to fit varying screen sizes and users' pointing performance. User testing with people with physical disabilities showed that participants generally performed better with the SUPPLE system than with other graphical user interfaces [1]. However, to reduce performance variance, participants in the SUPPLE study were explicitly led through the interface and told where to point and click. Such an experimental design is helpful for teasing out differences in pointing performance, but cannot be generalized more broadly to the actual usage and usability of auto-generated systems.

This study explores the use of auto-generated interfaces for people with functional limitations. In this study, the FIN-USI model was used as the basis for the auto-generation of two self-voicing interfaces. Both self-voicing interfaces accepted gesture input on a mobile device touchscreen.

2 Experimental Design

The main purpose of the experiment was to compare auto-generated interfaces to manufacturer interfaces. Specifically, blind users' performance, preference, and satisfaction were compared for manufacturer-created interfaces, including one interface designed for blind users by the manufacturer, and automatically-generated interfaces built using participants' interactor preferences. The auto-generated interfaces in this experiment were created on the base of the FIN-USI model (described in Part 1 in this volume), which aims to be a simple model that can cover a broad range of applications and functionality. With the FIN-USI model, each input element of an interface can be modeled as type of input (data type) and characteristics (input cardinality, time dependence, and validity characteristics) that are applied to that input. In the model, inputs may be further

grouped in logical or functional groupings. From the model of each target device, the interface generators created self-voicing, mobile device-based interfaces suitable for people who were blind or who had very low vision.

The research questions this study was designed to answer are:

Primary research question: Is there evidence of improved usability (performance, preference, and satisfaction) over manufacturer-created interfaces of either or both FIN-USI auto-generated interfaces that use each person's preferred interactors?

Secondary research questions: Are any usability differences moderated by the user group (which have different levels of self-voicing interface experience) or the target device? Even if usability is moderated by interacting factors, do the main effects of interface still stand?

While the study was not designed to specifically answer a further research question, we were also interested to see if there was any evidence that a novel loop-navigation interface (one of the auto-generated interfaces) showed improved usability over a more typical (but also auto-generated) interface layout.

To answer the research questions, the experiment was conducted as a 4-Interface × 2-Device × 5-Task(Device) within-subjects design with 2 groups of participants (between subjects).

2.1 Participants

Both participants with sight and participants who were blind were invited to be a part of the study of the screen reader and self-voicing interfaces. Blind participants who regularly use screen readers on computers or mobile devices are the primary target of such interface generators and thus were included in the study. While it is recognized that it is best to study representative users in controlled experiments [8], we were also interested in exploring a greater diversity of experience levels. Screen reader users tend to be technically savvy because current screen readers are complex—effectively using a screen reader requires one to memorize many keyboard shortcuts or gestures and switch between modes of operation. Could automatically-generated, preference-built interfaces be useful to people who had no prior experience with screen readers or other strategies that technically-savvy blind people might use? To have a greater diversity of technology experience, sighted participants were invited to participate in the study while wearing blindfolds. The recruitment target was for 12 people who were blind and 12 people who were sighted. Participants were paid for travel expenses and at a rate of $15 per hour for their participation.

Blind participants were recruited by making an announcement at a state blind convention, sending study advertisements to various local and statewide email lists that were of interest to people who are blind, and by having an e-mail message sent by a local vocational rehabilitation counselor to clients. Sighted participants were recruited through posters that were placed in stores and libraries in the local community. All participants completed a short set of screening questions over the phone. Participants were screened out of the study if they reported significant physical difficulty that would

hinder their ability to use a smartphone or if sighted participants felt that they could not wear a blindfold for at least 20 min at a time.

In total 12 people who were blind and 15 people who were sighted were initially recruited into the study. Three sighted individuals were dropped from the main study. The first two sighted individuals were dropped as run-in participants as the researchers identified issues with the experimental methods. One sighted participant was also screened out of this study because of observed difficulty in making the required interface gestures and completing the tasks during Phase 1, which took him 75 min (42%) longer than the next longest Phase 1 session with increased errors. This participant was invited back later to be part of a case study which had additional interfaces available for testing, including manufacturer-created and auto-generated interfaces that did not require gestures on a touchscreen. From the case study with this participant, no evidence was found that would contradict the findings of this study.

A total of 12 blind participants (6 female) with an average age of 37 (SD = 12.0) and 12 blindfolded participants (8 female) with an average age of 37 (SD = 15.5) completed both phases of the study. The blindfolded group had more racial diversity (11 white, non-Hispanic and 1 Hispanic blind participants compared to 8 white, 3 black, and 1 Asian blindfolded participants) and educational diversity than the blind group (all 12 blind participants had education beyond high school, whereas 3 of the 12 blindfolded participants had a high school education or less). A few sighted participants reported disabilities that might have had an effect on their performance during the sessions: two participants had ADHD, one of whom also had a diagnosed nonverbal learning disorder. It was also observed during the study that one of the blind participants had hand tremor, with the observed result that the iPod Touch would sometimes misinterpret the participant's gestures.

2.2 Instrumentation

During the experiment, participants frequently interacted with an iPod Touch (Model ME643LL/A, Apple Inc., Cupertino, Calif.) running iOS 8.2. The iPod Touch was connected to a computer and a mixer for recording and to a speaker to allow for louder and clearer audio. The iPod Touch was placed in a modified OtterBox (Fort Collins, Colo.) Defender Series case that had the screen protector removed for better responsiveness and a 1.9-mm-thick plastic piece added to block the top 4.3 mm of the screen with the iOS clock, icons, and notifications bar. For some parts of the experiment, participants used VoiceOver (a screen reader built into iOS) and for other parts of the experiment, the interface was programmed using JavaScript to provide speech output using the same voice as VoiceOver.

Participants also interacted with the physical controls of a smart thermostat (Model CT80, Radio Thermostat Company of America, San Francisco, Calif.) and a multifunction copier (Model X654de, Lexmark International, Inc., Lexington, Ky.), both of which are pictured in Fig. 1. These two devices were chosen as representative of common, every day-use devices in home and office environments that also had applications or web-interfaces that could potentially be used with a screen reader on a smartphone or

other mobile device. The multifunction copier was chosen for study because its web-based interface was specifically designed for use with screen readers. Both devices in the study had their interfaces modeled using the FIN-USI model. The two interface generators were basic: they used the FIN-USI model of the interfaces and then built an interface for the iPod Touch using the interactors that each participant preferred for each modeled input or output. There were no nuances or extra information included in the interface models or generators that could tune interfaces to particular tasks.

Fig. 1. Photographs of the physical interfaces of the copier (left) and thermostat. Both devices had a touchscreen and physical buttons. Rulers in the photos are marked in centimeters.

Participants used both the thermostat and iPod Touch on a table to make it easier to video record their interactions.

2.3 Methods

In this experiment, four general interface types were tested (where the prefix *Mfr-* denotes manufacturer-created interfaces and *Gen-* denotes auto-generated interfaces): Mfr-Physical, Mfr-Item, Gen-List, and Gen-Loop.

- Mfr-Physical: The manufacturer-created physical interfaces on the tested devices (see Fig. 1 above). Both UIs had physical buttons and touchscreens for control. Neither UI had speech output.
- Mfr-Item: Manufacturer-created UIs that were run on the iPod Touch using VoiceOver.
  - The thermostat had a native iOS application, which was not specifically designed for accessibility. Its UI failed two WCAG 2.0 [9] provisions (2.4.2 & 2.4.6), but could be used with VoiceOver if one figured out its idiosyncrasies.
  - The copier was chosen because it had a web-based interface specifically designed to be accessible and usable to people using screen readers on computers. Its UI used links and form fields and met all level-AA WCAG 2.0 provisions.
- Gen-List: Auto-generated, personalized interfaces that were generally laid out in a conventional list layout, with one interactor or element per row. The gestures used on this interface were a subset of the ones in iOS VoiceOver.
- Gen-Loop: Auto-generated, personalized interfaces with a novel interface layout and gestures [10]. In the Gen-Loop interface, all of the interactors and elements were

arranged around the edges of the screen in a clockwise direction. It was designed to be used by dragging around the edges of the screen (although it could also be used with many of the same gestures as the Gen-List interface) to find and activate elements.

When participants were introduced to the interfaces, they were called the Physical, Item, List, and Loop interfaces, respectively. Figure 2 shows example interfaces that were tested.

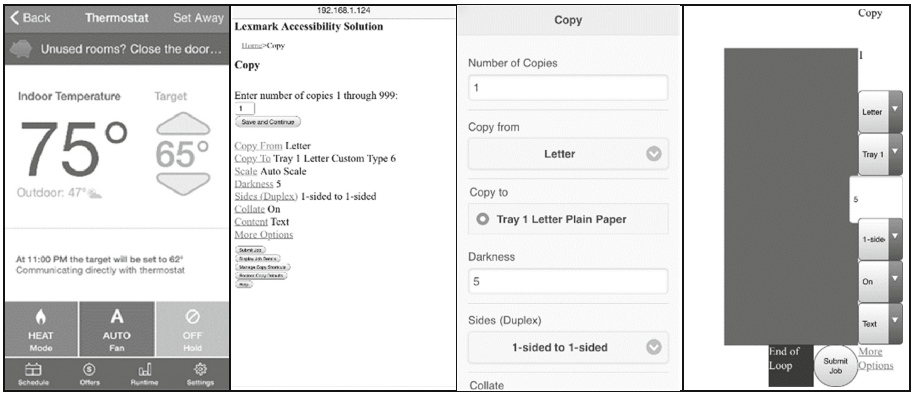

Fig. 2. Screenshots from the iPod Touch of representative interfaces that were tested. The first two were manufacturer-created interfaces. The last two screen shots were auto-generated and took somewhat different forms during the experiment for each user depending on individual preferences.

The experiment was in two phases. In the first phase, each participant's preferences for interactors was elicited. The second phase was the comparative study of the manufacturer-created interfaces and auto-generated interfaces using the participant's Phase 1 preferences. Sighted participants wore a blindfold while training and performing tasks. The interfaces were covered when participants took breaks from the blindfold.

Participants were told at the outset of the study that the researchers were comparing different interfaces for people who were blind. They were told neither the hypotheses of the experiment nor that they were choosing interactors and having customized interfaces generated for them. Furthermore, the experimenters who interacted with the participants during the study sessions were also not told the hypotheses of the experiment. They were familiar only with the procedures for interacting with participants and carrying out the data collection.

Before the experiment was conducted, the significance (α) level of 0.05 was chosen for all statistical tests.

Phase 1: Preference Elicitation. After participants arrived for the first session, a consent form and demographic questionnaire were administered. The volume and rate

of the iPod's text-to-speech output was then adjusted. The speech rate was incrementally slowed until the participant could correctly repeat five consecutive, random phrases that the system spoke from a list of phrases that the system was likely to say during the latter phase of the experiment. Finally, before training, participants double-tapped the screen many times to set their baseline double-tap rate (constrained to be between 250–1000 ms).

Participants then went on with training on the three touchscreen interfaces (Mfr-Item, Gen-List, and Gen-Loop) on the iPod Touch. All three of the introductory training interfaces had a set of on-screen buttons, one for each letter of the alphabet.

- With the Mfr-Item training interface, which utilized iOS's VoiceOver feature, the alphabet buttons were arranged in a grid with 3 buttons per column. VoiceOver would step through them in alphabetical order when the horizontal swiping gestures were used.
- With the Gen-List training interface, the alphabet buttons formed a single column, like a list.
- The Gen-Loop training interface had alphabet buttons arranged around the periphery of the screen.

The order of the three training interfaces was randomly ordered and balanced between participants. Participants were instructed how to navigate (i.e., move the focus around) the self-voicing interfaces by stepping (making short right or left swiping gestures) or by dragging their finger around the interface. Participants were told how to change pages (three-finger swipes up or down on the Mfr-Item and Gen-List interfaces and a continuing looping gesture on the Gen-Loop interface) and how to activate items (double-tapping anywhere on the screen when an element is highlighted). Finally, to ensure that participants had a basic understanding of the three interface styles, they were instructed to find and activate randomly named on-screen buttons until they had activated five buttons correctly in a row. After training on all the iPod self-voicing interfaces, participants ranked them by preference and commented on them.

The last and longest part of the Phase 1 session was eliciting the participants' preferences for particular interactors for specific types of input. Preference elicitation interfaces were constructed using the same two interactors of interest (for example, two dropdown menus or two sets of radio buttons). For each preference elicitation task, participants had to navigate past the first interactor on the screen to get to the second interactor, which they would need to manipulate. Participants would start by trying a random pair of interfaces with the same task and choose their favorite (ties were allowed). Their favorite interface from the prior pair was then paired with a random challenger interface, and they were then to try and rank those two interfaces. This pairwise comparison process was repeated until the favorite interface of all of the available ones was identified. Then the entire process was repeated for the next task and interface style. Participants started with the Gen-List interface and performed three sets of tasks in order: select-one-from-two, followed by select-one-from-seven task, and then numeric entry task. They then repeated the three sets of tasks and interactor comparisons with the Gen-Loop interface.

If there were two or more interactors tied for favorite for a given task, then the system automatically chose the interactor to be used for the auto-generated interfaces. The system's choice was based on the participant's tied interactors and a listing of interactors in a preference order determined before the study by the researchers.

Phase 2: Comparative Interface Testing. The main purpose of Phase 2 of the experiment was to gather performance, preference, and satisfaction data comparing the four interfaces on the two devices. Phase 2 was split into two sessions for the participants who were slower: one session for the initial factorial experiment and the second session for its replication.

Each Phase 2 session started with setting the three personal parameters as in Phase 1: volume, speech rate, and double-tap baseline. Afterwards, participants tried a short gesture practice session to become familiar again with the basic gestures used in the experiment: swiping right and left to navigate, double-tapping to activate, swiping up and down to change values, and three-finger swiping up and down to change pages. In the gesture practice sessions, participants were told a randomly-ordered task they were to accomplish (e.g., "Next", "Activate", or "Earlier Page"), and they would respond by making a gesture. Participants had to get four of each gesture correct the first time they were cued before they could continue to the training.

Participants were also trained on the interfaces that they were to use: the copier's Mfr-Physical interface, the thermostat's Mfr-Physical interface, and the Mfr-Item, Gen-List, and Gen-Loop interfaces on the iPod Touch. For the three iPod Touch interfaces, participants were trained using the preference elicitation interfaces and tasks (from Phase 1) with their favorite interactors (the Mfr-Item training interface used interactors that were like those that participants would use on the copier's Mfr-Item interface). The five interfaces were trained in random order. After completing the training, participants answered a questionnaire verbally before continuing with the factorial experiment.

The bulk of the data was gathered during the factorial experiment where participants tried every combination of Interface and Device with a set of five device-specific tasks (see Table 1). Participants would start with a random device (either the copier or the thermostat) and try each of the four interfaces (Mfr-Physical, Mfr-Item, Gen-Loop, and Gen-List) before moving on to the other device. The order of the interfaces was counterbalanced for each participant using a Latin Square. Participants were told to perform each task "as quickly and as accurately as possible." Participants were given no more than 2 min to complete a task, after which point the participant was told to move on and the task was recorded as a failure. After each task, participants verbally answered the Single Ease Question (SEQ) [11]. After each of the iPod-based interfaces, participants were verbally administered the System Usability Scale (SUS) questionnaire [12] modified to have 7-point Likert-scale items instead of the original 5-point. This modification was made to reduce the confusion observed during pilot testing with users switching between 5- and 7-point scales and because there is evidence that a 7-point SUS scale may provide more accurate measures than the original scale [13]. Participants were asked to rank and comment on all four interfaces after trying all of them for a given device.

Table 1. The device-specific tasks participants attempted on the different interfaces. Simple tasks were chosen so that participants would have higher rates of success. The asterisks denote tasks that would require the use of the (non-voicing) touchscreen on the Mfr-Physical device in order to complete the tasks.

#	Copier tasks	Thermostat tasks
1	Make a single copy	Raise the target temperature 2°
2	Make a darker copy*	Set the target temperature to 70°*
3	Make 5 copies	Turn the fan on*
4	Make a 2-sided copy from a 2-sided original*	Change the mode to cool*
5	Set the Content setting to "Photograph"*	Set the target temperature to 75° and turn hold on*

Depending on the time, participants would replicate the factorial experiment in the same session or come back for a second Phase 2 session. To reduce frustration, participants did not replicate the tasks on the Mfr-Physical interfaces that were a priori considered to be inaccessible because they required the use of non-voicing touchscreens for success. Participants were allowed breaks between devices and whenever they requested one.

3 Results

The mean Phase 1 session length (including breaks) for blind participants was 124 min. (SD = 39) and that for blindfolded participants was 137 min. (SD = 33). The difference between Phase 1 session times between groups was not significant; $t(22) = 0.880$, $p = 0.388$. A total of nine blind participants completed Phase 2 in one session with a mean length of 163 min. (SD = 23); nine blindfolded participants also completed Phase 2 in a single session (M = 184 min., SD = 22). The difference in single Phase 2 session times was not significant ($t(16) = 1.933$, $p = 0.071$). Of the three blind participants who had two Phase 2 sessions, the two sessions averaged 205 min. (SD = 18) and 128 min. (SD = 23) long, respectively. The three blindfolded participants who had two Phase 2 sessions had mean session times of 160 min. (SD = 10) and 132 min. (SD = 35), respectively.

3.1 Preferences for Interactors

To get each participant's favorite interactors for the automatic generation of Gen-List interfaces, participants tried all five available interactors for a Select-1-of-2 task, then all five available interactors for a Select-1-of-7 task, and then all four interactors for a number input task. Ties were allowed in the preferences. Several of the tested Gen-List interface interactors had the same behaviors and interaction as pre-existing interactors that were used in the copier's accessible Mfr-Item interface or the native iOS VoiceOver screen reader. Analyzed with exact binomial tests (see Table 2), participants favored the new interactors created for this study over the pre-existing interactors.

Table 2. Exact binomial tests of new interactors being favored over pre-existing interactors.

Task	P	k	n	Significance
Select 1 of 2	3/5	22	23	<0.001
Select 1 of 7	3/5	23	24	<0.001
Number input	2/4	15	19	0.010

Where P *is the probability of success (i.e., favoring new interactors) due solely to chance,* k *is the number of participants who favored new interactors (i.e., the number of binomial successes), and* n *is the number of participants who did not have tied favorites between new and pre-existing interactors.*

Participants also choose their favorite interactors in a similar manner for Gen-Loop interfaces. However, since the Gen-Loop interface is a novel interface style, there were no pre-existing interactors against which to compare.

3.2 Preferences for Interface Types

Participants were asked to rank the interfaces by preference at seven points during the two phases of the study. For analysis, each separate ranking can be treated as a binomial experiment where a "success" is defined as both automatically-generated interfaces (Gen-List and Gen-Loop) being ranked higher than both of the manufacturer-created interfaces (Mfr-Physical and Mfr-Item). The auto-generated interfaces were not favored with the first ranking ($p = 0.406$, exact binomial test with probability of 1/3, $k = 9$, and $n = 24$) after the Phase 1 training on the iPod Touch interfaces. For each of the subsequent rankings, however, both auto-generated interfaces were strongly favored by participants over both manufacturer-created ones ($p < 0.001$ for all exact binomial tests).

When blind participants were asked to rank the three mobile-device interfaces at the end of the experiment, 8 participants preferred the Gen-Loop interface, 3 preferred Gen-List, 0 preferred Mfr-Item, and 1 participant had a first-place tie between Gen-Loop and Gen-List. When blindfolded participants were asked to rank the interfaces at the end of the experiment, 10 participants preferred the Gen-Loop interface, 2 preferred Gen-List, and 0 preferred Mfr-Item. The difference between the number of blind participants who favored the novel Gen-Loop interface to the Gen-List interface was not significant (two-sided Wilcoxson signed rank test, $z = 1.31$, $p = 0.190$). However, blindfolded participants preferred the Gen-Loop interface over the Gen-List interface (two-sided Wilcoxson signed rank test, $z = 2.02$, $p = 0.043$).

3.3 Comparative Study Results

Four measures were collected as participants performed the various tasks. Success/failure and successful task time were recorded for each task on all four interfaces. The Single Ease Question (SEQ) responses were collected for each task on the three mobile-device interfaces (Mfr-Item, Gen-Loop, & Gen-list). The System Usability Scale (SUS) questionnaire was verbally administered after participants had completed

all the tasks on each of the iPod Touch-based interfaces (Mfr-Item, Gen-List, & Gen-Loop). The Spearman's rank correlations between these measures were significant ($p < 0.001$ for all correlations) and are shown in Table 3.

Table 3. Spearman's rank correlation coefficients for the four Phase 2 study measures.

	Rank-transformed performance	Success/fail	SEQ	SUS
Rank-transformed performance [(b)]	1			
Success/fail [(b)]	–0.902	1		
SEquation [(a)]	–0.760	0.720	1	
SUS [(a)]	–0.616	0.613	0.810	1

[(a)] *Measures were collected for the three iPod Touch interfaces.*
[(b)] *Measures were collected for all four interfaces.*

Each measure was analyzed separately for significance of the Interface factor and any Interface-factor interactions using methods appropriate for each type of data.

Success/Fail Data. The success/fail binary data was analyzed using Generalized Estimating Equations (GEE) with a binomial logit link function. This technique allows for correct inferences of repeated measures binomial data [14]. A full factorial model with Group (2) × Device (2) × Interface (4) × Replication (2) was run and then the most non-significant terms were iteratively removed from subsequent GEE models until a parsimonious model was found with the lowest Corrected Quasi-likelihood Information Criterion (final model QICC = 1656.137). The terms of the model and significance tests of their effects is shown in Table 4.

Table 4. Factors in the final GEE model of the success/fail data and their significance.

Source	Wald Chi-Square	df	Significance
(Intercept)	1.516	1	0.218
Group	9.991	1	0.002
Device	10.589	1	0.001
Interface	214.503	3	<0.001
Replication	18.593	1	<0.001
Group × Device	2.460	1	0.117
Group × Interface	18.918	3	<0.001
Device × Interface	24.228	3	<0.001

All the main effects were significant at a 0.05 α-level. The Group, Device, and Interface factors were also involved in two-way interactions. Pairwise post hoc tests were conducted on the significant interactions using the sequential Sidak procedure.

The Group × Interface interaction (see Fig. 3) was significant. Between the blind and blindfolded groups on the Mfr-Item interfaces, blind participants had a success rate

0.44 points (0.20–0.67 95% Wald difference CI) higher than that of the blindfolded participants ($p < 0.001$). There were no other significant differences between the groups on particular interfaces. For both groups, the Mfr-Physical and Mfr-Item interfaces had significantly lower success rates than the Gen-List and Gen-Loop interfaces (both with $p < 0.001$). For the blindfolded participants, the Gen-List interfaces had a success rate that was 0.60 points (0.43–0.77 difference CI) higher than on the Mfr-Item interfaces. For blind participants, the Gen-List interfaces had a success rate that was 0.36 points (0.21–0.50 difference CI) higher than on the Mfr-Item interfaces. Differences between success rates on the Gen-List and Gen-Loop interfaces within each group were not significant.

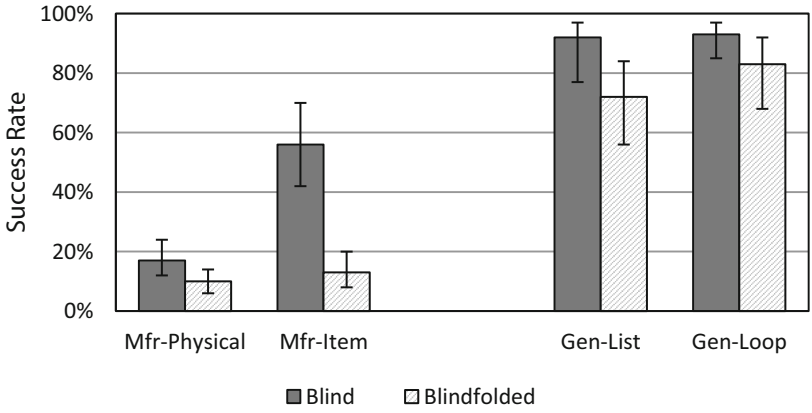

Fig. 3. The estimated marginal mean success rates for the Group × Interface interaction. Error bars show 95% Wald confidence intervals for the marginal estimates.

The Device × Interface interaction was significant. With both devices, participants were more successful on the Gen-List and Gen-Loop interfaces than the Mfr-Physical and Mfr-Item interfaces. On the copier, the Gen-List interface had a success rate 0.42 (0.22–0.63 95% Wald difference CI) points greater than on the Mfr-Item interface ($p < 0.001$). Similarly, participants using the thermostat were successful 0.61 (0.40–0.81 difference CI) points greater when using the Gen-List interface than when using the Mfr-Item interface ($p < 0.001$). Differences between the Gen-List and Gen-Loop interfaces on particular devices were not significant.

Performance Data. The success/fail data is folded into the performance data, because task times were only recorded for successful tasks (see Fig. 4). Because the users failed significantly more often on the manufacturer-created interfaces, the performance differences were only compared between the two auto-generated interfaces. To test the differences in performance, the task time and failure data were converted to a comparable scale, which was done by rank transforming the data (where all failures were given tie rankings). Repeated measures ANOVA was then used to analyze the ranked performance data partitioned by device.

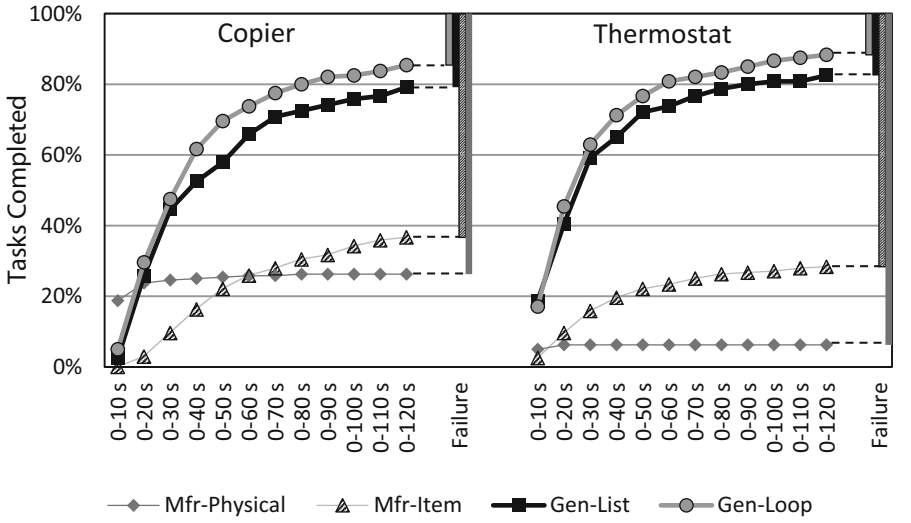

Fig. 4. Cumulative distributions of task performance on the copier and thermostat.

For the copier, the Interface main effect ($F = 3.346$, $df = 1$) was not significant ($p = 0.082$). The only significant interaction involving the Interface factor was the Interface × Task interaction ($F = 3.382$, $df = 4$, $p = 0.013$). Further investigation with sequential simple effect Sidak post hoc tests indicated that participants performed better on the Gen-Loop interface for two of the five tasks (with $p = 0.005$ and $p = 0.015$).

For the thermostat, the Interface main effect ($F = 1.149$, $df = 1$) was not significant ($p = 0.295$). No interactions involving the Interface factor were significant.

Single Ease Question (SEQ) Data. The single ease question (SEQ) on a 7-point scale anchored with 1 = "very difficult" and 7 = "very easy" was asked after each task. A full factorial Type III Sum of Squares GEE model on the Group (2), Device (2), Interface (3), and Replication (2) factors was run on the interval-type SEQ data with a multinomial cumulative logit link function. Only the three iPod interfaces (Mfr-Item, Gen-List, and Gen-Loop) were included in the analysis because the Mfr-Physical interface was low

Table 5. Factors in the final GEE model of the SEQ data and their significance. Only the Mfr-Item, Gen-List, and Gen-Loop interfaces were included in this analysis because they had complete factorial data.

Source	Wald Chi-Square	df	Significance
Group	6.874	1	0.009
Device	0.364	1	0.546
Interface	98.308	2	<0.001
Replication	19.532	1	<0.001
Group × Interface	7.395	2	0.025
Device × Interface	9.626	2	0.025

scoring and did not have complete factorial data for the replication. The GEE model was run iteratively removing 4-, 3-, and 2-way interactions that were obviously non-significant one at a time. The significance data of the resulting final model is shown in Table 5. The two auto-generated interfaces scored significantly higher (Mdn = 7 for Gen-Loop and Mdn = 6 for Gen-List) than the Mfr-Item interface (Mdn = 2, IQR = 4).

As a post hoc comparison of the Gen-List and Gen-Loop interfaces, the data was partitioned to remove the Mfr-Item interface and the full factorial GEE multinomial cumulative logit model was run again with iterative removal of non-significant effects. The factors and significance data of the resulting final model is shown in Table 6. On the SEQ, participants rated the tasks on the Gen-Loop interface as easier (Mdn = 7, IQR = 1) than the Gen-List interface (Mdn = 6, IQR = 2).

Table 6. Factors in the final GEE model of the SEQ data for the two auto-generated interfaces and their significance

Source	Wald Chi-Square	df	Significance
Group	1.463	1	0.227
Device	0.759	1	0.384
Interface	6.290	1	0.012
Replication	21.435	1	<0.001
Group × Replication	3.722	1	0.054

System Usability Scale (SUS) Data. The 10-question Likert SUS questionnaire was answered by participants after completing all five tasks of each replication with the three iPod interfaces (Mfr-Item, Gen-List, and Gen-Loop). A Cronbach's alpha reliability score was calculated for each of the 12 times the SUS was administered. Agreeing with the literature (e.g., [15, 16]), the individual SUS scale items were highly consistent for each administration in this experiment; the average Cronbach's alpha for the items in the SUS in this experiment was 0.931 with values ranging from 0.865–0.973. Because of this high reliability, the individual items were transformed and summed, as is typical with SUS measurements, for the subsequent analysis.

The SUS data was analyzed as a Group (2) × Device (2) × Interface (3) repeated measures ANOVA, as is typical [17]. The main effect of Interface was significant ($F(2) = 77.800$, $p < 0.001$), as was the Interface × Replication interaction ($F(1.525) = 5.148$, $p = 0.018$ with Greenhouse-Geisser correction for significant asphericity of the data), which is shown in Fig. 5.

Sequential simple effect post hoc tests with the sequential Sidak correction were performed on the Interface × Replication interaction. These tests indicated that participants scored the Gen-List interface 8.56 points higher on the replication ($p = 0.015$), the Gen-Loop interface 7.00 points higher on the replication ($p = 0.001$), and no significant SUS score change on the Mfr-Item interface with the replication ($p = 0.273$). Both the Gen-List and Gen-Loop interfaces significantly outscored the Mfr-Item interface at both time points ($p < 0.001$). The SUS scores for the Gen-List and Gen-Loop interfaces for the first block were not significantly different ($p = 0.059$), but the Gen-Loop did

score 9.58 points higher (0.39–18.78 95% difference CI) than the Gen-List interface with the replication ($p = 0.039$).

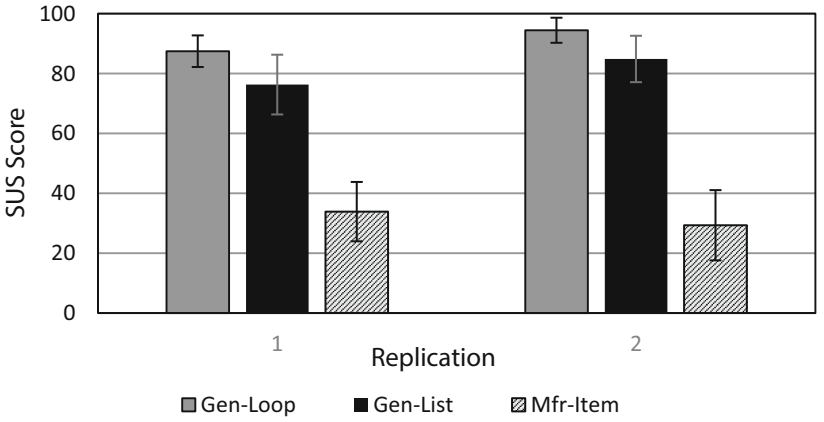

Fig. 5. The Interface × Replication interaction showing the marginal means of the SUS scores. The error bars show 95% normal-distribution confidence intervals for the marginal interaction means.

4 Discussion

When choosing their favorite interactors in Phase 1, participants had strong preferences for the Gen-List interactors that were designed for this experiment over "pre-existing" interactors. Particularly notable was the dislike of the Picker interactor (which is what is presented to people using the iOS VoiceOver screen reader whenever a person encounters a dropdown list (e.g., a `<select>` element in HTML). One blind participant, who had used pickers before on her own device, said bluntly that, "Pickers could go to hell." Nobody in the experiment chose the Keyboard interactor, which was closely modeled after the iOS keyboard mode that was presented for numeric entry (e.g., entry into a `<input type = "number">` element in HTML). Both the Picker and Keyboard interactors were examples of layered interfaces—where people use a screen reader interface layer that is reading (and interacting with) a graphical user interface that was specifically designed for mainstream users with vision. Having automatically generated interfaces can eliminate this layering of interfaces, because the interactors can be designed with users' primary modalities in mind. For this experiment, the "new" interactors were designed specifically with speech output and touchscreen gesture manipulation in mind; any visual representations were secondary and mostly included to make it easier for the sighted researchers to observe.

While participants liked the "new" interactors, several of the blind participants wished that additional gestures were supported by the two auto-generated interfaces.

For example, two blind participants use the VoiceOver drag-and-tap-with-a-second-finger gesture on their own iPhone devices to make selections rather than the double-tap that was required by the research system. Others had difficulty with the gestures required by both VoiceOver and the research systems and would rather have substituted their own gestures. The blind participant with tremor suggested having a dedicated area on the touchscreen or button on the device for activation rather than double-tapping anywhere on the screen. Such user preferences could be supported by personalized, automatically-generated interfaces.

In Phase 1, the interactors that participants chose and the reasons that they reported for making the choices did not always seem optimal to the researchers. For example, one blindfolded participant chose a double-tap interactor for the Select-1-of-7 task (which requires a person to double-tap to change values and thus double-tap multiple times to cycle through all the values). This participant reported some frustration in Phase 2 when he used Gen-List interfaces, because he would sometimes overshoot the desired selection and have to double-tap through the entire list all over again. Other participants occasionally wished that the Gen-List or Gen-Loop interfaces were a little different when they were trying the study tasks of Phase 2 rather than the preference elicitation tasks. For example, a blind participant who chose radio button interactors in both the Gen-List and Gen-Loop interface reported later in Phase 2 that it was awkward to have to switch pages so much when completing the copier or thermostat tasks (which is the tradeoff with choosing radio buttons over menu-based or other interactors). This suggests that a different approach to preference elicitation may be helpful. It is also possible that users who knew ahead of time that they are choosing interactors for personalized interfaces might choose differently than participants who are just told to pick favorite interactors without context as in the study. Users may also pick better interactors if they were to do a first pass to screen out the interactors that they strongly dislike, and then try more comparisons with more realistic interfaces and the finalist interactors. Creating a better preference elicitation process would be a good avenue for future work.

Even with some participants' interactor choices seeming to be suboptimal, the two automatically-generated interfaces tested better than the manufacturer-created interfaces for all measures on both the copier and thermostat devices. The magnitude of the difference was greater than expected. Before conducting the experiment, it was expected that at least the copier's accessible web-based interface (i.e., the copier's Mfr-Item interface designed specifically for blind screen reader users) would have been a much closer match to the two automatically-generated interfaces, because it had been hand-crafted specifically for people who are blind and using screen readers. One blind participant said that copier's Mfr-Item interface was "100-percent do-able," and that she probably would have scored it more highly on the usability questionnaire (SUS) if she had not just tried a superior interface beforehand (in her case, the Gen-List interface). For participants in this experiment, the automatically-generated interfaces were consistently better than the manufacturer-created ones.

The results of the sub-study of the performance and preference differences between the two automatically-generated Gen-List and Gen-Loop interfaces are not as clear, however. The subjective measures (ranking, SEQ, and SUS) showed limited evidence that the Gen-Loop interface may be better than the Gen-List interface, but the success/

failure and performance data did not. The fact that the Gen-Loop interface was perceived as generally preferred to or better than the Gen-List interface might also be a bias related to the good-subject effect [18]. The Gen-Loop interface was obviously the most different interface to people who were blind, so "good" participants might have felt that the study's aim was to show that the Gen-Loop interface was better and "good" participants might respond in such a way to support that perceived hypothesis. It would be fair to say that participants had different and sometimes changing preferences between the Gen-List and Gen-Loop interfaces. One blind participant strongly disliked the Gen-Loop interface concept in general because he felt it was confusing and unintuitive. Other participants liked how the Gen-Loop interface felt logical and efficient. Some participants did not have much of an opinion either way because they used both the Gen-List and Gen-Loop interfaces with swiping navigation gestures. If participants chose particular interactors, then they could potentially have exactly the same user experience when swiping to navigate with both Gen-List and Gen-Loop interfaces. Many participants did not like the Gen-Loop interface at first with only limited exposure and use, but later after becoming accustomed to the interface style, many participants' preferences changed. One participant had his own hypothesis and said that the Gen-Loop interface violated the two ways that blind people have interacted with user interfaces so far: (1) navigation using arrow keys, tabbing, or swiping and (2) scanning through the interface as with a visual magnifier or dragging on a touchscreen. He said that the Gen-Loop interface was good, but that he did not like it at first because it required him to break the force of habit: "Forget all that. [The Loop interface] might be a better, more efficient way to do this."

This lack of a clear difference between Gen-List and Gen-Loop interfaces may lend more support to model-based generation of user interfaces. Model-based, automatic interface generation can support people's preferences, even for very different layouts and interaction styles. People who prefer and perhaps better comprehend a more typical, linear interface could have automatically-generated interfaces that look and behave like a Gen-List interface. Other people might prefer to use a Gen-Loop interface because it is intuitive and more efficient for them.

While group differences were not the focus of the study, the data supported the expectation that people who were blind perform better with blind-specific interfaces than people who were blindfolded and had no prior experience with those interfaces or techniques and strategies that are used by people who are blind. It was remarkable that the blindfolded participants did as well as they did. VoiceOver and screen readers in general have a steep learning curve that can be particularly difficult for elders and others who are not technically savvy (K. M. Fountaine, personal communication, August 31, 2015). The two-way Group × Interface interaction of the success/fail data (plotted in Fig. 3 above) could be interpreted along with the flexibility and experience of participants. Blind participants were more flexible and experienced and thus did better on the Mfr-Item interface than the blindfolded participants. However, the automatically-generated interfaces studied here were very consistent and had relatively few interactors and gestures, which seemed to make it easier for even novice blindfolded users to experience success.

5 Conclusion

The self-voicing interfaces that were automatically generated using each participant's preferred interactors were statistically better on all performance and usability measures than the manufacturer-created interfaces. While preferences varied in the beginning when participants were first learning about the different interface styles, participants also preferred the auto-generated interfaces once they had finished training and started performing the actual tasks. These results are notable because the model and interface generators were both relatively simple.

This study supports the ability of the FIN-USI modeling approach to automatically generate user interfaces that participants prefer and on which they perform better compared to the manufacturer-created interfaces, including one interface designed by the manufacturer specifically for people who are blind. It also supports the sufficiency of even simple interface generators created using the model to outperform manufacturer interfaces. In the future, model-based auto-generation of interfaces could support a wide range of user needs and preferences, where people could choose the type of interface they want and from what components interfaces are built.

Acknowledgements. The contents of this paper are based on work carried out with funding from the National Institute on Disability, Independent Living, and Rehabilitation Research, U.S. Department of Health and Human Services, grant number H133E080022 (RERC on Universal Interface and Information Technology Access). However, the contents do not necessarily represent the policy nor imply endorsement by the funding agencies.

References

1. Gajos, K.Z., Weld, D.S., Wobbrock, J.O.: Automatically generating personalized user interfaces with Supple. Artif. Intell. **174**, 910–950 (2010). doi:10.1016/j.artint.2010.05.005
2. Meixner, G., Paternò, F., Vanderdonckt, J.: Past, present, and future of model-based user interface development. i-Com **10**, 2–11 (2011). doi:10.1524/icom.2011.0026
3. Myers, B., Hudson, S.E., Pausch, R.: Past, present, and future of user interface software tools. ACM Trans. Comput. Hum. Interact. **7**, 3–28 (2000). doi:10.1145/344949.344959
4. Nichols, J., Myers, B.A., Higgins, M., Hughes, J., Harris, T.K., Rosenfeld, R., Pignol, M.: Generating remote control interfaces for complex appliances. In: Proceedings of the 15th Annual ACM Symposium on User Interface Software and Technology, pp. 161–170. ACM, New York (2002). doi:10.1145/571985.572008
5. Coelho, J., Duarte, C., Biswas, P., Langdon, P.: Developing accessible TV applications. In: The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 131–138. ACM, New York (2011). doi:10.1145/2049536.2049561
6. Peissner, M., Häbe, D., Janssen, D., Sellner, T.: MyUI: generating accessible user interfaces from multimodal design patterns. In: Proceedings of the 4th ACM SIGCHI Symposium on Engineering Interactive Computing Systems, pp. 81–90. ACM, New York (2012). doi:10.1145/2305484.2305500
7. Nichols, J., Chau, D.H., Myers, B.A.: Demonstrating the viability of automatically generated user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1283–1292. ACM, New York (2007). doi:10.1145/1240624.1240819

8. Sears, A., Hanson, V.: Representing users in accessibility research. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2235–2238. ACM, New York (2011). doi:10.1145/1978942.1979268
9. Caldwell, B., Cooper, M., Reid, L.G., Vanderheiden, G., Chisholm, W., Slatin, J., White, J. (eds.): Web Content Accessibility Guidelines (WCAG) 2.0 (2008). http://www.w3.org/TR/WCAG20/
10. Jordan, J.B.: A circular direct-selection interface for non-visual use. IPcom Prior Art Database. Disclosure Number: IPCOM000241004D (2015)
11. Sauro, J., Dumas, J.S.: Comparison of three one-question, post-task usability questionnaires. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, pp. 1599–1608. ACM, New York (2009). doi:10.1145/1518701.1518946
12. Brooke, J.: SUS: A "quick and dirty" usability scale. In: Usability Evaluation in Industry. Taylor and Francis, London (1996)
13. Finstad, K.: Response interpolation and scale sensitivity: evidence against 5-point scales. J. Usability Stud. **5**, 104–110 (2010)
14. Lee, J.-H., Herzog, T.A., Meade, C.D., Webb, M.S., Brandon, T.H.: The use of GEE for analyzing longitudinal binomial data: a primer using data from a tobacco intervention. Addict. Behav. **32**, 187–193 (2007). doi:10.1016/j.addbeh.2006.03.030
15. Bangor, A., Kortum, P.T., Miller, J.T.: An empirical evaluation of the system usability scale. Int. J. Hum. Comput. Interact. **24**, 574 (2008). doi:10.1080/10447310802205776
16. Sauro, J., Lewis, J.R.: Correlations among prototypical usability metrics: evidence for the construct of usability. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, pp. 1609–1618. ACM, Boston (2009). doi:10.1145/1518701.1518947
17. Sauro, J., Lewis, J.R.: Quantifying the User Experience: Practical Statistics for User Research. Morgan Kaufmann, Waltham (2012)
18. Nichols, A.L., Maner, J.K.: The good-subject effect: investigating participant demand characteristics. J. Gen. Psychol. **135**, 151–166 (2008). doi:10.3200/GENP.135.2.151-166