GENERALIZABLE USER-INTERFACE RESEARCH

Edmond W. Israelski, Joel S. Angiolillo-Bent, Douglas J. Brems,
LaVerne L Hoag, Linda A. Roberts, and Robin S. Wells

The authors are with AT&T Bell Laboratories. **Joel S. Angiolillo-Bent,** **LaVerne L Hoag,** *and* **Robin S. Wells** *are members of technical staff in the Human Factors and International Planning Department in Lincroft, New Jersey, and* **Edmond W. Israelski** *is supervisor of the Human Factors Group in that department.* **Douglas J. Brems** *is a member of technical staff in the User Interface and Quality Planning Department in Holmdel, New Jersey.* **Linda A. Roberts** *is a distinguished member of technical staff in the Systems Engineering Department in Middletown, New Jersey. Mr. Angiolillo-Bent works on the usability of customer instructions for PBX products and designs user interfaces for system management products for AT&T PBXs. He joined the company in 1982 and has a Ph.D. in cognition and communications from the Uni-*

A high-quality, friendly user interface can and does affect how users perceive a product's quality. Human-factors specialists have been using a variety of methods to bring behavioral data into the process of specifying and evaluating user-interface design. Empirical studies are often used to evaluate design issues for a specific product, but we have found that more generic studies can influence a wider variety of products. In this paper, we describe three such generic user-interface studies whose results led to new product-specific user interfaces. Although each of these studies was done for a specific product, the findings have also proved useful for a wide variety of telecommunications products.

31

Introduction

High-quality, friendly user interfaces to AT&T products are more likely when requirements have a solid empirical base. Through a variety of methodologies, human-factors specialists at AT&T Bell Laboratories bring behavioral data into the process of specifying the user interface. These methodologies include: literature reviews of previous studies, evaluations of iterations of designs through rapid prototyping, and both controlled laboratory and field studies. Although empirical studies are often applied narrowly to a product-specific set of concerns, human-factors specialists do find opportunities to conduct studies that are more generic. Such studies produce results that not only can provide user-interface requirements for a specific product, but can also be generalized across a range of telecommunications products.

This paper presents three different sets of generic user-interface studies that illustrate the practical use of traditional experimental design in psychology. The areas covered are:

- *User perception of switch performance.* This study measured the noticeability and acceptability of performance parameters for telecommunications switches. The work was done in setting design criteria for AT&T's Definity™ PBX (private branch exchange) systems but has since been applied as design criteria for electronic key-

Panel 1. Acronyms and Terms	
ANOVA	analysis of variance
icon	a graphical image used on a computer screen to represent a real-world object
p	probability of error; level of statistical significance
PBX	private branch exchange
PC	personal computer
SD	standard deviation
UNMUIDS	Unified Network Management User Interface Design Standards
UNMA	Unified Network Management Architecture

telephone systems (such as AT&T's Merlin® communications system) and central-office switches (such as AT&T's 5ESS® electronic switch).

- *Distinctiveness and pleasantness of auditory alerting patterns.* This work has resulted in patents that cover new ways to indicate call-priority information, while at the same time providing personalized ringing signals. These features will be available in the next generation of AT&T business telephones for use with PBXs, key systems, and central-office switches.

- *Effectiveness of icons in human-computer interfaces.* (An *icon* is a graphical image used to represent a computer object or process, such as a word-processing file or the delete function.) This work has led to the specification of icons in direct-manipulation user interfaces for system administration products that are used to administer a large business customer's network. (The term *direct manipulation* means that a pointing device, such as a mouse, is used to select and move representations on the computer screen to perform various functions.)

The outcomes of these studies were used to specify new product-specific user interfaces and also provided useful findings that have been of value across a range of tele-

communications products. (Panel 1 defines terms and acronyms used in this paper.)

User Perceptions of Switch Performance

In the first generalizable study, the effect of telecommunications-switch performance parameters on users' perception was investigated. Specifically studied was how users judged the performance quality of a telecommunications switch (such as a PBX), based on the functioning of its user interface. As call volume and load on the switch's processor increase, it is important to know how changes in the switch's performance parameters affect its users. This investigation comprised a multimethod human-factors research program that included a literature review, a laboratory study, and two field studies.

From the literature review and knowledge of reported customer problems, we surmised that a particular switch performance parameter, *voice-path cut-through time*, had a strong influence on customer satisfaction. This parameter represents the period of time that starts when a called party lifts the telephone receiver off-hook and ends when the switch establishes a voice circuit for the conversation. When voice-path cut-through delay exceeds the voice-onset time, the calling party could hear a partial or clipped version of the called party's initial greeting. *Voice-onset time* is the interval between when the called party lifts the handset (going off-hook) and begins to speak. If this interval is shorter than the voice-path cut-through time, then the calling party will not hear the part of the initial greeting that is spoken before the voice-path is established (i.e., the greeting is *clipped*). As the amount of information clipped increases, users' perception of system quality would suffer accordingly.

The laboratory study systematically investigated perception of the voice-path cut-through delay parameter, while the field studies investigated users' perceptions of a series of switch performance parameters.

Laboratory Study. The 134 participants in the laboratory study ranged in age from 20 to 50; 85 percent were

32

female, 15 percent were male; and all lived in nearby communities. The subjects participated in pairs.

To provide different voice-path cut-through delay distributions, two AT&T PC 6300 computers (one for each subject's telephone) were programmed to add a delay to the normal distribution. Each subject's telephone was connected to a dual-trace oscilloscope. When a subject's telephone went off-hook, the ensuing speech and switch activity appeared as signals on the oscilloscope. A stereo tape recorder attached to each telephone recorded subject sessions.

Experimental sessions involved pairs of subjects playing a board game. Each player was in a separate room and relayed information about his or her board moves through telephone calls to the other player. During a session (which lasted about 2-1/2 hours), subjects usually participated in two blocks of trials. In each block, a subject alternately placed and received a total of 80 telephone calls.

One of nine voice-path cut-through delay distribution conditions (see the second column in Table I) was assigned to each block in an experimental session. However, we counterbalanced the order of presentation of delay distribution conditions across sessions for all subjects. (Half the subjects got one presentation order, while the other half received the reversed order.) At least 20 subjects evaluated each condition. Table I gives the mean and the 99th-percentile cut-off point for each distribution, as well as the label used for the condition.

At the end of each block of trials, subjects were given a questionnaire to complete. Questions referred to more than 20 call setup delays, in addition to voice-path cut-through delay. The experiment also measured voice-onset times and the frequency and duration of clipped speech.

Delay results. Subjects used a five-point scale to rate several aspects of call setup on two dimensions: *noticeability* (i.e., did they noticed the specified delay) and *acceptability* (i.e., was the delay acceptable to them). The results of these ratings provided answers to two

Table I. Voice-Path Cut-Through Delay Distributions

| Condition (labeled delay, ms)	Actual mean delay (ms)	99th percentile (ms)
1	400	413	540
2	850	844	1000
3	1000	990	1120
4	1150	1139	1240
5	1300	1301	1400
6	1600	1598	1680
7	1900	1901	2425
8	2100	2115	2600
9	2700	2703	3200

important questions:
1. Did the manipulation of voice-path cut-through delays generate ratings that were related to the delay conditions?
2. Is it possible to determine exactly when subjects significantly shift their evaluations of voice-path cut-through delays?

As Figure 1 shows, changes in these delay conditions were directly related to subjects' ratings of the noticeability and acceptability of voice-path cut-through delay, as well as to ratings of the overall quality of service. [For all analyses of variance (ANOVAs), $p < 0.0001$. When the level of statistical significance is expressed in terms of the probability of error (p), this states that the result is statistically significant.] However, changes in the delay conditions did not relate to any ratings for dial-tone delay or post-dialing delay. Thus, in this experiment, subjects perceived the attributes of voice-path cut-through delay to be different from those of dial-tone and post-dialing delay, but distinctly related to overall quality of service.

After tests of multiple paired comparisons, three findings emerged that provided a positive answer to the second question posed (i.e., when did ratings shift markedly). As Figure 1a shows, subjects significantly shifted

33

Figure 1. Laboratory study of voice-path cut-through delay conditions. (a) Mean subject sensitivity to voice clipping, and (b) mean rating of overall quality of service.
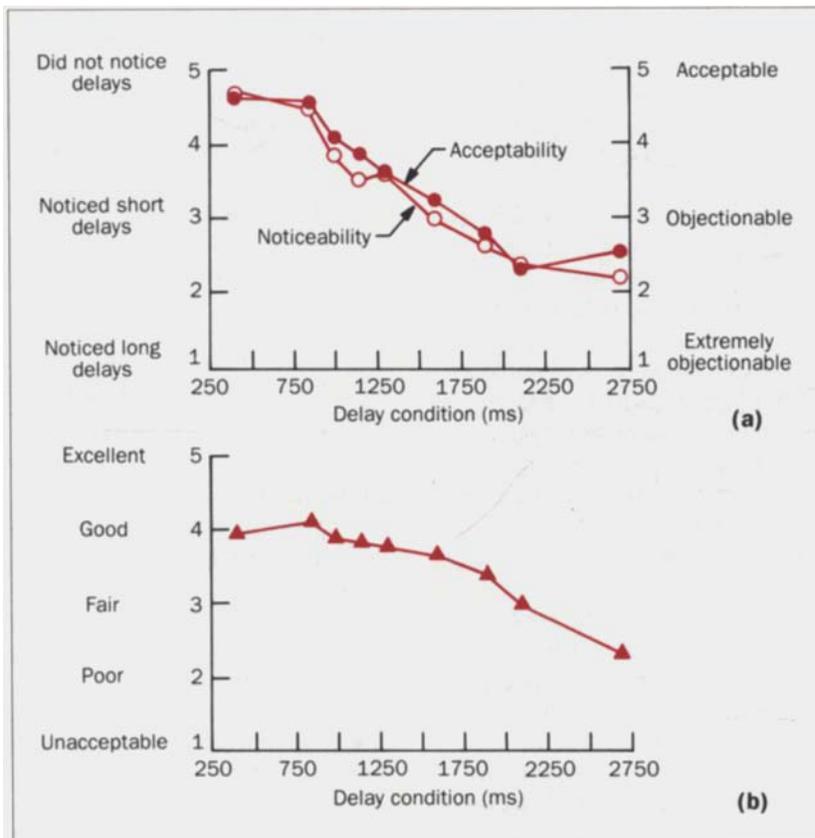
(p < 0.05) their ratings of the *noticeability* of voice-path cut-through delay after the 850-ms (milliseconds) mean-delay condition. A significant shift (p < 0.05) in the subjects' ratings of the *acceptability* of voice-path cut-through delay occurred after the 1000-ms mean-delay condition. And finally, the subjects' ratings of the *overall quality* of switch performance (Figure 1b) shifted significantly downward (p < 0.05) after the 1600-ms mean-delay condition.

Onset-time and clipping results. These subjects used regular handsets. The mean of the distribution of voice-onset times was 1143 ms, the median was 1040 ms, and the standard deviation was 372 ms. Voice-onset times did not fall below 400 ms, and the lowest individual mean voice-onset time was 761 ms. Fewer than 10 percent of the subjects had mean voice-onset times under 1000 ms. As the average standard deviation of 289 ms suggests, individual voice-onset times also varied across trials.

These data support the view that, among handset users, a less stringent criterion for voice-path cut-through delay would not lead to noticeable instances of speech clipping. When the mean delays exceeded 1150 ms, the frequency of speech clipping increased significantly (p < 0.0001); e.g., the proportion of trials clipped went to 23 percent, then to 59 percent. However,

34

the duration of clipping began to show a significant increase starting at the 1600-ms delay condition. At the same mean delay (i.e., 1600 ms), the subjects' ratings of the overall quality of service showed a significant downward shift.

Thus, 99 percent of the time, users appear to tolerate the consequences of voice-path cut-through delays well beyond a criterion level of 900 ms, even in the context of highly concentrated trials.

Field Studies. Two field studies were conducted to clarify questions raised during the laboratory study. The first field study surveyed switch users for their acceptance of switch performance when the switch was controlled for specific load levels. The second study measured the difference in voice-onset time for users of regular telephones with handsets and of speakerphones, to determine if speakerphone users would tolerate longer voice-path cut-through times.

To evaluate user acceptance of switch performance, a survey asked users to rate the acceptability and noticeability of 22 aspects of their telephone service. The survey included:
- Overall switch performance
- Overall switch delays
- Call setup delays (including voice-path cut-through)
- Lamp activation delays on voice terminals
- Delays on alphanumeric displays
- Feature activation delays (e.g., call forwarding, send all calls).

Acceptability and noticeability were rated on the same five-point scale as in the laboratory study.

The users on the selected switch were divided into two groups. The first group received the survey form and rated switch performance while the switch was working under its normal load condition. Then, a high-load condition was induced using a load generator. The second group experienced the new load for one week before receiving the survey form. The 99th-percentile voice-path cut-through times for the two conditions (normal and high) were 500 ms and 750 ms, respectively.

In the second field study, we wanted to determine the difference in voice-onset times for speakerphone users and regular-handset users. We believed that speakerphone users had shorter voice-onset times and that they, therefore, would experience increased amounts of clipping.

We attached an oscilloscope to the tip and ring of a central-office line and then called 184 managers at AT&T Bell Laboratories who had both speakerphones and regular telephones in their offices. When the telephone was answered, the time to the appearance of voice energy could be measured from the oscilloscope's trace. We held a brief conversation with each person to determine the type of telephone being used.

Switch performance results. Under both normal (500 ms) and high (750 ms) load conditions, users rated the switch's performance as good. According to the users' ratings, neither noticeability nor objectionability was perceived as being much greater under the high-load condition than under the normal condition. This was true for all 22 of the switch performance parameters, including voice-path cut-through delay.

In addition, people with high telephone usage (more than 20 calls per day) did not give ratings significantly different from people with low telephone usage (no more than 10 calls per day).

Voice-onset difference. The "speakerphone" field study provided oscilloscope measures of 69 speakerphone users and 115 regular-handset users. In this study, the mean voice-onset time measured was 734 ms for regular-handset users (SD = 313 ms), and 352 ms for speakerphone users (SD = 254).

Because we could not determine which call setup activity triggered the oscilloscope, these times are lower than the true voice-onset times by a constant, δ. The constant appeared to be about 400 ms, but this paper's space restrictions prevent our discussing the evidence for this conclusion. In any case, the constant's value was the same for both telephone types. Therefore, the 418-ms difference between voice-onset times of the

35

two groups is the real difference; i.e., $(\mu_1 - \delta) - (\mu_2 - \delta) = 418$ ms, where μ_1 represents regular-telephone users, μ_2 represents the speakerphone group, and δ is a constant.

The probability that speech clipping will occur is the probability that the voice-path cut-through time is greater than the voice-onset time. To estimate these probabilities, we used the sample field data from the switch performance study to estimate voice-path cut-through times. For handset users, the probabilities of clipping are 0.0003 (for a normal load) and 0.002 (for a high load). For speakerphone users, the probabilities of clipping are 0.031 (for normal load) and 0.053 (for high load).

Summary and Conclusions. The results of laboratory and field studies provided useful insight on user perceptions of telecommunications-switch performance parameters. The laboratory data indicate that voice-path cut-through delays are clearly acceptable to users, until delay values exceed the mean delay distribution of 1000 ms. Users in the field study rated voice-path cut-through delay under two load conditions. Field-study data supported the laboratory data up through a 750-ms mean delay. The inclusion of data from speakerphone users (whose response is relatively faster) provided additional insight into the limits of user tolerances for switch performance delays.

Initially, this work was applied to design criteria for Definity PBXs. Since then, these results and recommendations have been used to set performance criteria for electronic key-telephone systems and central-office switches.

Auditory Alerting Studies

The second generalizable study investigated auditory alerting signals. Designers of computer and telephone systems frequently use *auditory signals* to inform, warn, or prompt a user. Error tones, dialing tones, and ringing signals are three examples of the hundreds of tones in use on AT&T products. Often the tones carry simple meaning, such as a programmable tele-

---

Panel 2. Dimensions Assessed in the Sound Study

The alerting signal study assessed nine dimensions and used the following criterion for the levels to be measured:

1. *Single frequency*—The four frequencies, which ranged from 300 to 1200 Hz, were equally distant from each other (in terms of musical notes on a staff), and were in the most pleasing pitch range (as determined in pilot work).
2. *Two-note melody*—Four different musical intervals (two-note "melodic" sequences) were chosen; two were "consonant" (e.g., the major third), and two were "dissonant" (e.g., the major second).
3. *Octave*—The four patterns consisted of melodic sequences, where the tones were in an octave relationship. Two of the patterns contained four-note sequences, and two patterns had five-note sequences.
4. *Three-note melody*—Each pattern consisted of a three-note melodic sequence, where each sequence differed in its pattern of "ups and downs."
5. *Four-note melody*—Each pattern consisted of a four-note melodic sequence. As for the three-note melodies, each sequence had a unique contour.
6. *Warble*—The same frequencies were used as in the "melody" condition. All patterns were frequency modulated at a 15-Hz rate.
7. *Timbre*—Four preprogrammed timbres of the Yamaha DX7 sound synthesizer were selected. Each timbre had the same pitch.
8. *Modulation rate*—Four different frequency modulation rates were selected that ranged from very slow (5 Hz) to very fast (17 Hz). The interval of a major third was used for each pattern.
9. *Rhythm*—A single frequency was used for all patterns. The four patterns were: one short tone, two short tones, three short tones, and one long tone.

phone's error tone to inform the user that he or she has entered an incorrect sequence. In a more complex scenario, the ringing signal on a business telephone may have to tell the user who is being called, where the call originated, and what is its priority.

Suppose we want to design a system of telephone-ringing signals that identify both the call's origin (e.g., inside the building, outside the building, a transferred call, or a priority call) and the call's destination (e.g., the user's phone or the office mate's phone). For all combinations of four call origins and two call destinations, we would need eight different ringing signals. But to differentiate among these eight states, we can manipulate only a few dimensions of sound in a simple electronic system; for example, frequency, amplitude, duration, patterns of pitch changes, or pitch differences between two notes. (See Roberts and Angiolillo-Bent[1] for a selective bibliography on these issues.)

It is desirable for a system of auditory signals to have several properties, principally pleasantness and distinctiveness. If users turn off the tones because they are irritating, then these signals cannot be informative. Moreover, if users confuse different signals, then the system has again lost its informational content.

Dimensions of the Sound Study. In the first experiment of this research program, we investigated sound qualities (*dimensions*) that contribute to an auditory signal's pleasantness and distinctiveness. There is rich psychological literature on the perception of tones.[2,3]

Panel 2 lists the nine different dimensions of sound quality that the experiment assessed. For each of these dimensions, we selected four different levels that we felt covered a range of reasonable values for modern electronic devices. For example, the dimension *single frequency* was represented by four frequencies that ranged from 300 to 1200 Hz (hertz). Another constraint was that the levels should be distinct from each other and pleasant.

Procedure. Forty-four subjects participated in the experiment, in groups of four. Each of the nine different dimensions formed a block, and each subject heard all nine blocks. In each block, the four levels of the dimension being tested were repeated four times, resulting in 16 trials per block. To balance sequence effects, we created three different orderings of the nine blocks. Each block was divided into three phases:

- *Familiarization*—In this phase, subjects heard each of the four levels of a particular dimension twice. The stimulus was on for 1.2 seconds, with a 3.6-second pause between the two repetitions. There were 6 seconds of silence between one auditory signal and the next.

- *Learning*—The learning phase followed the same format as familiarization. But in this phase, each of the four subjects was paired with one of the four stimulus levels and instructed to remember his or her assigned level.

- *Test*—In this phase, the subjects received 16 trials. Each of the four levels was presented four times in a block but in a random position in the block.

For each of the 16 trials, subjects made two separate judgments. First, they indicated if a pattern was theirs. Second, they indicated how much they liked this pattern, using a seven-point rating scale. In addition, at the end of a block and at the end of the experiment, subjects rated the stimuli's pleasantness and distinctiveness (i.e., how easy was it to discriminate among the stimuli in the block).

Results. We analyzed the data in terms of perceived pleasantness, perceived distinctiveness, and percent-correct identification. In each of the three analyses, the only reliable difference was the main effect of dimension (ANOVA, $p < 0.01$), which demonstrates that, in all three tasks, the nine dimensions received reliably different ratings.

In Figure 2, we show the results for each of the dimensions; that is, the percent-correct identification versus pleasantness. Clearly, the dimension judged to be most pleasing entailed warbles (frequency modulated
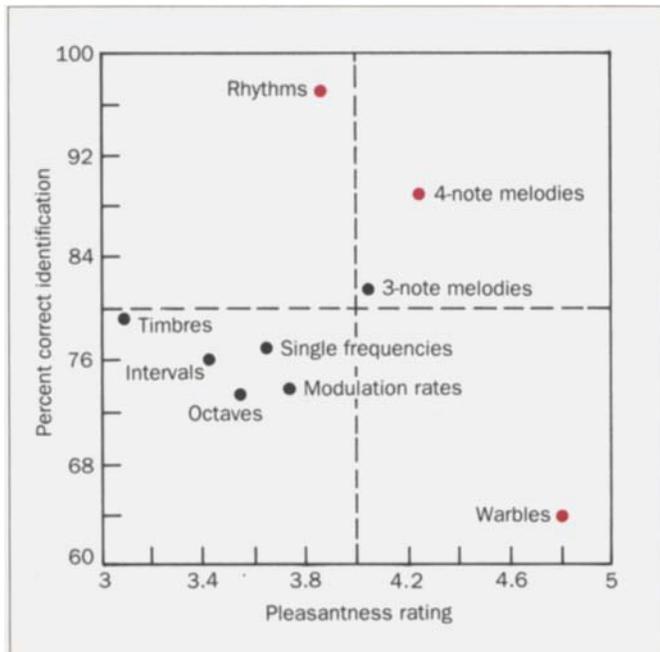
Figure 2. Sound study results. Shows the pleasantness and correct-identification ratings for nine dimensions of sound.

intervals); next were the four-note and three-note melodies. The least pleasant dimension entailed timbres, followed by two-note melodic intervals. For the correct-identification data, rhythms were most easily identified (nearly 97-percent correct), and warbles were least discriminable (65-percent correct). Results of the perceived distinctiveness data were similar to the correct-identification data.

These results permit us to create a two-dimensional (distinctiveness versus pleasantness) representation of each dimension.

Thus, if an application requires maximum discrimination of auditory patterns, then patterns that vary along the rhythmic domain are most appropriate. Such

signals would be most beneficial where immediate understanding of the signal is critical—for example, in an aircraft cockpit. These patterns would not be particularly pleasing unless other variations are involved (perhaps frequency).

If pleasantness must be maximized—for example, in a dentist's office—then, frequency-modulated tone pairs (warbles) appear to be the best candidate. However, it is difficult to achieve distinctiveness as well, because most listeners are not good at discerning interval differences. Training may be required if distinctiveness is also important.

When both distinctiveness and pleasantness are critical, the best dimension appears to be four-note melodic sequences. Although these patterns did not achieve the highest rankings for either pleasantness or distinctiveness, they reached the next-to-the-highest position in both dimensions. It may be possible to optimize both pleasantness and distinctiveness even more by creating better patterns within the dimensions or by creating patterns that combine some of these musical dimensions. Moreover, other aspects of sound (e.g., loudness) may also be critical to enhanced distinctiveness or pleasantness.

A final caveat is that these conclusions are limited to the levels of the dimensions chosen. For example, subjects could easily distinguish rhythmic patterns with two or three beats, but they probably would not do as well when distinguishing patterns that contain seven, eight, and nine beats.

Follow-Up Studies. The findings reported above were extended in four additional experiments:
- Different melodic patterns (experiment 2)
- More complex patterns involving both rhythm and melody (experiment 3)
- Different rhythmic patterns (experiment 4)
- Different frequencies (experiment 5).

In the first of these follow-up studies (experiment 2), 40 listeners rated the pleasantness of 96 differ-

ent three-note and four-note melodic patterns. The most pleasant were musically resolved four-note patterns drawn from a major chord with small interval sizes.

In experiment 3, both rhythmic and melodic information were combined to create a more complex signal. Each signal comprised two short segments; one contained a sequence of beats (rhythm), and the second a four-note melody (derived from the "best" of the patterns assessed in experiment 2). Subjects tried to categorize signals that varied in these two different dimensions. (For example, test signal 1 might contain melody A followed by two beats, while test signal 2 might contain melody B preceded by three beats.)

We found that subjects were most successful in picking out both the melody and the rhythmic pattern when the rhythmic information followed the melody and when the rhythmic portion was a different frequency than the melody.

Experiment 4 looked at different types of rhythmic patterns to see what types yielded the best discrimination performance. In addition, we tried to replicate the results of experiment 3 but under different conditions. In this replication, the rhythmic patterns were found to be more effective when they were syncopated with the four-note melodies (that is, when they did not begin on the beat set up by the melody). Varying the duration of the rhythmic beats did not affect performance.

Finally, in experiment 5, we assessed a variety of frequencies for the rhythmic portion of the signal. The goal was to find those frequencies that were the most pleasant and easiest to discriminate. The rhythmic frequency that worked best was a unique one; i.e., it was a frequency that was different from any of the notes in the melody.

Conclusions. This rich set of findings on the important dimensions of frequency, rhythmic pattern, timbre, and modulation has yielded generic results that can be applied when designing a variety of electronic devices.

These data allow signals to be selected that are distinctive and discriminable, as well as pleasant. The results have led to ringing-signal specifications for our next generation of business telephones.

Graphics-Based User Interfaces

Another active area of current human-factors work at AT&T is in design and development of human-computer interfaces.[4-8] Work on human-computer interfaces is becoming increasingly important for AT&T telecommunications products (e.g., network-management systems) and general data-processing systems, such as AT&T personal computers. With each new advance in computer capabilities, user-interface engineers must determine how to implement new functions and information types in ways that will be most beneficial to users.

The use of graphics in computer systems provides a good example of important human-factors work. As recently as 10 years ago, virtually all computer input and output was text. But current systems often rely on graphical information such as windows, icons, maps, graphs, and line drawings. (*Windows* are separate viewing areas that are provided on the screen by software.) Human-factors contributions to the use of graphics can be found in many forms including work on the grouping, consistency, layout, and color of computer graphics elements.[4]

An especially innovative and successful case is the development of icon-based systems; i.e., systems in which small pictures and symbols are a critical part of the user interface. The development of icon-based interfaces has included the creation of new metaphors and interaction styles for communicating with computers. Until recently, computer systems had been based on a conversational metaphor; i.e., the user communicated with the computer only through a computer language or code. But today, many systems are based on analogies to real-world objects and a user-interface style of "touching and moving" graphical images.[5,6] For example, in the popular desktop computing metaphor:

39

- Word-processing files are represented by icons that resemble sheets of paper.
- Directories are images of common office-file folders.
- Icons for electronic-mail functions are the familiar "in" and "out" boxes found in many offices.

This metaphor presents the user with an analogy to real-world objects and helps make the computer system easier to learn, especially for computer novices.

Many icon-based systems also feature a new type of interaction style, sometimes referred to as *direct manipulation*. With direct manipulation, a user performs functions by selecting and moving graphical images, often without using verbal commands. For example, to delete a word-processing file, the user might point at a *sheet of paper* icon, and then move the icon across the screen to a spot where the image of a *wastebasket* is located.

One project in the User Interface and Quality Planning Department (51134) in AT&T Bell Laboratories has been concerned with how to use icons in human-computer interfaces. The study's results serve as input to standards that guide the design of AT&T user interfaces,[7] including the Unified Network Management User Interface Design Standards (UNMUIDS) and functional specifications for the Open Look™ software for a graphical user interface. In this way, the AT&T standards for developing user interfaces are based on the needs and preferences of customers, and on data about the types of user interfaces that are successful. The issues that were studied include: how to design a "good" icon; the value of adding verbal labels to icons; and interaction styles, or ways of performing functions, that are appropriate with icons.

The rest of this section considers one of the more general experiments on types of interaction with icon-based user interfaces.

Representation and Interface Style. Several motives drove the work on this generic study:
- There seemed to be some confusion about what made recent icon-based systems successful. The computer industry tends to equate the use of icons with a *user interface that is good for novices*. But can the success of

any one system be properly attributed to the use of icons? Should the advantages of icons be limited to computer novices?
- There is the issue of interface style. The relationship of successful icon-based systems and successful methods of interaction is not clear.

This study evaluated two categories of interface styles: direct manipulation and keyboard input.

By direct manipulation, we mean an interface style in which the individual uses a pointing device (such as a mouse) to move representations on the screen to perform various functions. Direct manipulation has been used primarily with icon-based systems, but can be implemented with textual representations as well. For example, to delete a file, the individual might drag a filename to a delete function represented by the word *delete*.

By keyboard input, we mean an interface style in which the individual uses a keyboard to input verbal commands. For example, to delete a file, the individual might type: `rm file1`. Keyboard input has been used primarily with textual representations, but can be implemented with icons. For example, typing `rm file1` might cause the icon that represents `file1` to disappear.

The true relationship of icons and interface style has important implications for user-interface standards. For example, if icons are just as effective and helpful with keyboard input as with direct manipulation, then keyboard input could be used as the primary interface style for AT&T systems to promote consistency across graphics and nongraphics systems.

To consider these issues, we performed a two-variable factorial study, with object representation (icons or word representations) and interface style (keyboard input or direct manipulation) as the independent variables. Four prototype systems were evaluated: direct-manipulation-icon, direct-manipulation-word, keyboard-icon, and keyboard-word systems. This permitted us to analyze the advantage of icons for each interface style and analyze the two interface styles.

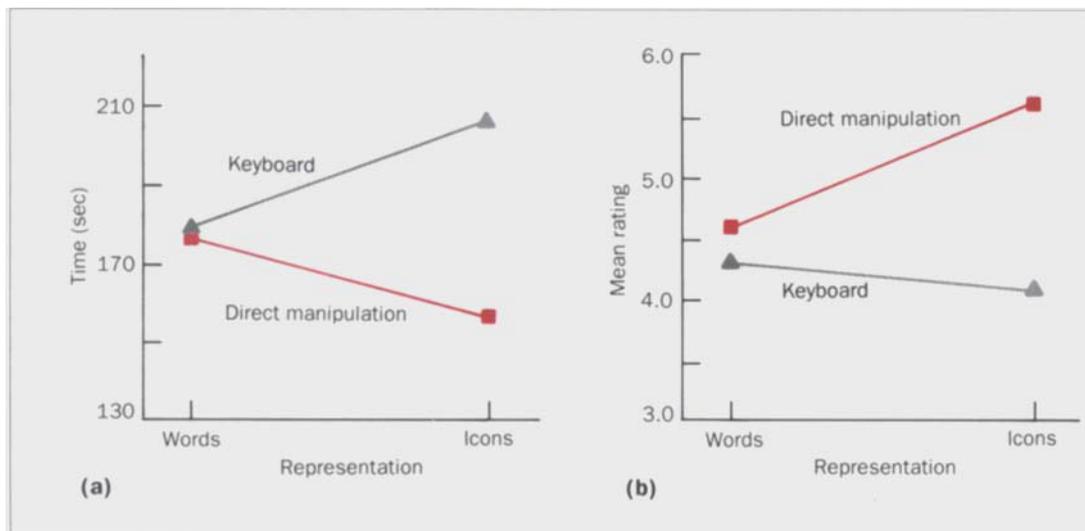Procedure. In the experiment, the subjects were

Figure 3. Evaluation
of four prototype
user-interface models:
direct-manipulation-icon,
direct-manipulation-word,
keyboard-icon, and
keyboard-word.
(a) Mean time to
complete file-
manipulation tasks,
and (b) mean prefer-
ence rating for user
interface.

32 technical employees at AT&T Bell Laboratories with no formal human-factors experience. Most were experts on keyboard-based user interfaces, especially with the UNIX® operating system, and most had little experience with direct-manipulation interfaces. Thus, this study provided a good test of whether the advantages of icons and direct manipulation can be generalized from computer novices to experienced users of keyboard user interfaces.

The main task involved a set of common file-manipulation procedures, a task common to virtually all computer systems. Successful direct-manipulation and keyboard-, icon-, and text-based systems for manipulating files are currently available. In this study, the direct-manipulation systems resembled the user interface provided by the Open Look software, and the keyboard systems resembled the successful UNIX operating system with the Bourne shell.[9] The tasks included moving, copying, and deleting files within a hierarchical directory.

In the icon systems, we used icons to represent files, directories, and the delete function. For a file, the icon also pictorially represented the file type (i.e., text, data, map, or graph) and was always accompanied by

textual presentation of the file's name. Directories were represented by a folder icon that was also labeled by name. In the text-based systems, files were represented by textual names and directories by textual names enclosed in square brackets.

Each participant was given about 30 minutes of personal instruction on how to use each system, as well as brief reference notes. Each participant was then asked to execute a series of standard procedures with each of the four prototype systems.

The main performance measure was the time to complete standard file-manipulation tasks. Preference measures included overall ratings of the four systems, and ratings on 12 bipolar-interface dimensions.

Results. Speed and accuracy are the two most common measures of performance for evaluating human-computer interfaces. In this study, all subjects were able to complete the tasks correctly, but differences in speed of performance were found.

Figure 3a presents the median times to complete each task in the four main conditions. The fastest completion times were found in the direct-manipulation-icon

41

condition. An analysis of variance on this data showed a statistically significant interaction ($p < 0.05$), and a Newman-Keuls analysis indicated that performance with the direct-manipulation-icon system was significantly faster than with the keyboard systems ($p < 0.05$).

After completing the tasks with all four systems, each subject gave his or her preference ratings. Each subject rated the four systems on a scale from 1 to 7 where a 1 equaled *least preferred* and a 7 equaled *most preferred*. Figure 3b presents the mean preference ratings. The direct-manipulation-icon condition received the highest preference ratings. As with the performance data (Figure 3a), the interaction was statistically significant ($p < 0.01$), and ratings of the direct-manipulation-icon system were significantly higher ($p < 0.01$) than for the other three systems.

These findings were somewhat surprising, because the subjects in this study were AT&T Bell Laboratories technical employees who had much experience on keyboard-based UNIX systems. Nevertheless, a direct-manipulation-icon system seemed to be the best system for these participants when performing these file-manipulation tasks. The results also indicate a cumulative, interactive effect of icons and the direct-manipulation interface style; that is, the advantages of icons and of direct manipulation were fully realized only when icons and direct manipulation were integrated.

In addition to rating simple preferences, subjects rated each of the four systems according to 12 bipolar dimensions, such as inefficient-efficient, dull-interesting, and complicated-simple. Based on these ratings, we performed multivariate discriminant analyses to find those attributes that distinguish icon systems from word systems, and those attributes that distinguish direct-manipulation systems from keyboard systems.

These analyses showed that the icon-based systems were "more visually attractive," had a "new technology appearance," and were "more interesting" than word-based systems. But the analyses also shows that the direct-manipulation interface was distinguished from the keyboard interface by a greater number and a greater breadth of attributes. More specifically, the direct-manipulation interface was rated higher on the variables: good for novices, simple, pleasant to use, fast, interesting, new technology appearance, comfortable, visually attractive, accurate, reliable, and efficient. This suggests that the direct-manipulation interface style is an important factor in the design of successful icon-based user interfaces.

Conclusions. With the introduction of any new user interface, tradeoffs will often exist. For direct manipulation, the benefits are a user interface form that is easier to learn and apparently better liked by users. On the other hand, inconsistency with the user interfaces of nongraphical systems and terminals is one disadvantage.

When evaluating such tradeoffs, generalizable human-factors research can provide valuable information for standards and user-interface design decisions. While this paper has provided only a glimpse of the data on icons and interface styles, clearly this study provided important evidence that the user-interface style largely affects the benefits of icon-based user interfaces. We also provided clear evidence that even experienced users of keyboard-based systems could quickly appreciate the direct-manipulation interface style.

This study suggests that the further development and improvement of direct-manipulation user interfaces as a part of AT&T standards should be well worth the effort. These standards have been used to help developers design high-quality and friendly user interfaces to telecommunications network management products.

Summary

The three sets of studies described here illustrate the value of designing product-specific user-interface studies on a broader scale and, as a result, providing results that are generalizable to related products. Often, little extra effort or time is required to apply some aspects of telecommunications product-specific research to a broader range of products.

42

For example, generalizability should ideally be done during the design phase of any study when the human-factors investigator usually will search the literature and, during the process, will often communicate with other developers and human-factors specialists. Together, they may discover that, with slight changes to the study's design, the findings might generalize to other related products. For example, the voice-path cut-through studies were initially done to satisfy design concerns for the Definity PBX product. But the findings have since been applied to electronic key-telephone systems (e.g., the Merlin communications system) and central-office switches, such as the 5ESS electronic switch, with virtually no extra effort required.

The work on auditory alerting signals has resulted in patents for designs that are being incorporated into the next generation of AT&T's digital business-telephone products. This product line includes terminals used on PBXs, key telephone systems, and Centrex service from central-office switches.

Finally, the results that demonstrate the human-performance advantages of coupling icon-based user interfaces with direct manipulation will broadly affect the quality of user-interface design for a wide range of telecommunications network management products. These products fall under the broad umbrella of Unified Network Management Architecture (UNMA) products.

References
1. L. A. Roberts and J. S. Angiolillo-Bent, "The Relative Pleasantness and Distinctiveness of a Variety of Auditory Patterns," *Proceedings of the Human Factors Society 32nd Annual Meeting*, Anaheim, California, October 24-28, 1988, The Human Factors Society, Santa Monica, California.
2. R. Erickson, *Sound Structure in Music*, University of California Press, Berkeley, California, 1975.
3. T. D. Rossing, *The Science of Sound*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1983.
4. B. Shneiderman, *Designing the User Interface*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1987.
5. D. Gittins, "Icon-Based Human Computer Interaction," *International Journal of Man-Machine Studies*, Vol. 24, 1986, pp. 519-543.
6. C. G. Wolf, "A Comparative Study of Gestural and Keyboard Interfaces," *Proceedings of the Human Factors Society 32nd Annual Meeting*, Anaheim, California, October 24-28, 1988, The Human Factors Society, Santa Monica, California, pp. 273-276.
7. J. M. Farber, "The AT&T User-Interface Architecture," *AT&T Technical Journal*, Vol. 68, No. 5, September/October 1989, pp. 9-16.
8. D. J. Brems and W. B. Whitten II, "Learning and Preference for Icon-Based Interfaces," *Proceedings of the Human Factors Society 31st Annual Meeting*, New York, New York, October 19-22, 1987, The Human Factors Society, Santa Monica, California, pp. 125-129.
9. S. R. Bourne, *The UNIX System*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1982.

Biographies (continued)

versity of Chicago. Mr. Brems designs user interfaces for network management products. He joined the company in 1986 and has a Ph.D. in psychology from Rice University. Mr. Hoag works on user interfaces for maintenance systems and data services on AT&T PBXs. He joined the company in 1980 and has a Ph.D. in industrial engineering (human factors) from the University of Michigan. Mr. Israelski's group supports the design of features and equipment for customer premises switching systems and their adjuncts. He joined the company in 1970 and has a Ph.D. in engineering and industrial psychology from Stevens Institute of Technology. Ms. Roberts works on style and physical design, audio, documentation, and user-interface issues for small-business products. She joined the company in 1982 and has a Ph.D. in experimental psychology from Rutgers University. Ms. Wells designs user interfaces for PBX voice and attendant features. She joined the company in 1984 and has a Ph.D. in experimental psychology from New York University.

43