

A Collection of Papers from FirstPerson, Inc.*

Jeff Johnson
Mark Keavney

SMLI TR-95-41

August 1995

Abstract:


The papers included in this collection are the results of work conducted at the prior FirstPerson, Inc., the group which later produced the interactive browser, HotJava.

The five papers report on experiments conducted:

- 1) to determine which of several candidate user interfaces for panning is most usable and intuitive;
- 2) to determine the smallest touchpad that subjects could accurately use to point to objects on a TV-sized video display;
- 3) to determine the effects of various factors on users' ability to hit an animated character (a sprite) moving across the TV screen, using a touchpad separate from the display;
- 4) to determine how accurately subjects could use a touch-sensitive remote control to hit targets on a TV screen; and
- 5) to determine which of several candidate pointing devices is most usable for controlling interactive TV applications.

Because FirstPerson, Inc. is no longer an active business unit, and because of its close association with the Labs, this collection was printed as a courtesy by Sun Microsystems Laboratories to document the results of these experiments.

* FirstPerson, Inc. existed from 1993 to 1994.

 *Sun Microsystems*
Laboratories, Inc.
A Sun Microsystems, Inc. Business
M/S 29-01
2550 Garcia Avenue
Mountain View, CA 94043

email addresses:
jeff.johnson@eng.sun.com

Table of Contents

FP-1994-1	A Comparison of User Interfaces for Panning on a Touch-Controlled Display	1
FP-1994-2	The Effect of Touch-Pad Size on Pointing Accuracy	21
FP-1994-3	Factors Affecting Users' Ability to Hit Animated Characters	33
FP-1994-4	An Evaluation of the Accuracy of a Touch-Sensitive Remote Control	45
FP-1994-5	A Comparison of Remote Pointing Devices for Interactive TV Applications	59
	About the authors	86

A Comparison of User Interfaces for Panning on a Touch-Controlled Display

Jeff Johnson

FirstPerson Technical Report FP-1994-1

May 1994

Abstract:

An experiment was conducted to determine which of several candidate user interfaces for panning is most usable and intuitive: panning by pushing the background, panning by pushing the view/window, and panning by touching the side of the display screen. Twelve subjects participated in the experiment, which consisted of three parts: 1) subjects were asked to suggest panning user interfaces that seemed natural to them, 2) subjects each used three different panning user interfaces to perform a structured panning task, with experimenters recording their performance, and 3) subjects were asked which of the three panning methods they preferred. One panning method, panning by pushing the background, emerged as superior in performance and user preference, and slightly better in intuitiveness than panning by touching the side of the screen. Panning by pushing the view/window fared poorly relative to the others in all three parts of the experiment.

A Comparison of User Interfaces for Panning on a Touch-Controlled Display

Jeff Johnson

1 Introduction

Often, computer-based displays provide views of a scene or information that is too expansive (wide or tall) to be shown in its entirety. A common solution is to provide a panning or scrolling function, allowing users to control which portion of the subject is visible in the display. If such a function is provided, a user interface—a way for users to invoke and control panning—must also be provided.

Panning user interfaces for touch-displays has not been studied much, if at all. In fact, very little has been published on panning and scrolling user interfaces of any kind. A few panning and scrolling user interfaces for specific task-domains have been described (e.g., Glassner, 1990; Beard et al., 1991), and researchers have written about the merits of alternative user interfaces for scrolling and panning (Beck and Elkerton, 1988) and about how to compare alternative designs rationally (MacLean et al., 1989). However, this work has focused on user interfaces controlled via keyboard or various pointing devices, not touch-displays. Though empirical studies of touch-display interfaces have been performed (Plaisant and Sears, 1993; Potter, Weldon, and Shneiderman, 1993; Sears and Shneiderman, 1993), these have not included studies of panning and scrolling. Several ways of using touch-displays have been proposed (e.g., Shneiderman, 1993), but panning through a flat information space is not among them. Recently, a product that incorporates a user interface for controlling panning has appeared in the marketplace (Sullivan, 1993), but nothing describing how that design was validated has yet been published.

The goal of the current study was to determine the best—meaning most easily used—user interface for panning and scrolling in a simulated physical world displayed on a touch-controlled display.

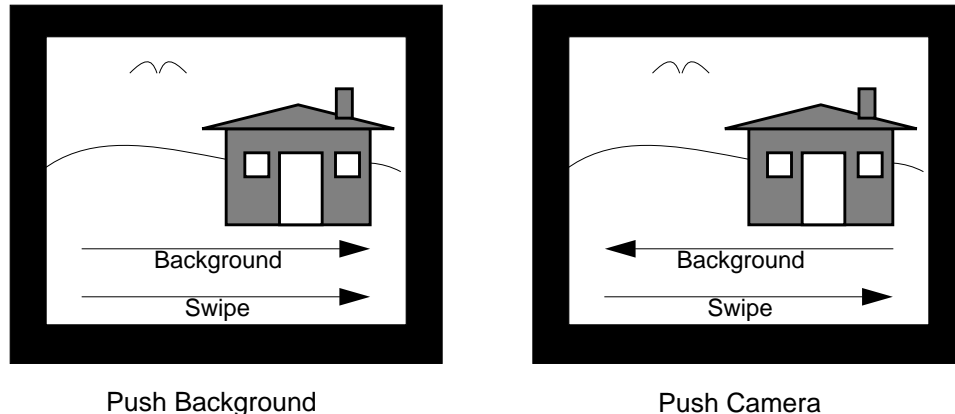


Figure 1. Two Contrasting Touch-Controlled Panning Methods

Initial experience with a prototype information space controlled by a touch-display raised an old scrolling user-interface issue in a new form. The prototype had been designed such that swiping one’s finger across the display caused the scene to shift. The question was: What should the relationship be between the direction of the swipe and how the scene shifts? Is the user’s finger swipe best regarded as pushing the scene or pushing the “camera” or “window” that is viewing the scene? These two different approaches have very different effects. With a Push Background user interface, users swipe in the direction they want the scene to go. This seems completely natural until one considers that to bring information onto the screen that is off-screen left, users would swipe towards the right. With a Push Camera user interface, users swipe in the direction of the off-screen information. This, too, seems natural until one considers that the physical display doesn’t actually move, so it is the scene that shifts—in the opposite direction as the swipe. Of course, both interfaces can’t be “natural,” since they have opposite effects.

The correct interface depends on how users think about it: do they view themselves as shifting the background or panning the camera? Therefore, the right way to resolve the issue is via user testing. The designers of the initial prototype had used the Push Camera approach, based on an informal paper-and-pencil test they had conducted using a few co-workers as subjects. However, experience with the prototype suggested that a significant number of users considered its panning user interface to work “backwards” from what they expected.

In an attempt to resolve this issue, a more formal study was devised, using subjects from outside the company who had no experience with the existing prototype and so were not already familiar with its user interface for panning. The idea of the study was that subjects would first be asked—before they had tried or seen any panning user interface—to describe one that would seem natural to them. Subjects would then perform a panning accuracy test with different panning user interfaces. Finally, they would be asked to say which panning user interface they preferred of those they had tried.

As the study was being designed, it was decided that panning user-interfaces other than Push Background and Push Camera should be considered. One plausible alternative to panning by swiping a finger is panning by touching the sides of the touch-display. In fact, this user-interface for panning was already used in the prototype in certain situations. When a user is dragging an object to a new location that is off-screen, it is impractical to have to drop it, pan the scene, pick the object up, and resume dragging. Therefore, in the prototype, when a user drags an object to a side of the display and holds it there, the “camera” pans towards that side. It is reasonable to consider having panning work that way *all* the time rather than only when the user is dragging something. This interface is best regarded as Touch Edge Camera because the user feels as if hitting an edge of the display pushes the “camera” or “viewing window” in that direction. This interface was added to the set to be compared.

Another possible panning user interface is the opposite of Touch Edge Camera: users press on the side of the display they want the background to move towards. However, because this interface—referred to herein as Touch Edge Background—is in the author’s experience very counter-intuitive, it was not included in the comparison; it also seems incompatible with the goal of having panning work well with dragging.

Regarding the first part of the study, it is clear that the appearance of the touch-display can influence what people suggest when subjects would be asked to suggest a panning method. After all, that is what Norman (1988) means by an “affordance”: an aspect of an artifact’s design that suggests how it should be operated. We thought that adding a brightly colored border around the displayed image might suggest “touch here” to users, and might therefore suggest Touch Edge panning (camera or background). We therefore designed the study so that half of the subjects would see such a border, and half would see the “normal” display (with the displayed image taking up the entire screen). We also had an idea for how to design the touch-display casing so it would suggest Touch Edge Camera to users. A casing was created that could be slipped on over the normal touch-display and

tried out on all subjects after they had suggested a panning method based upon the “plain” prototype touch-display cover.

2 Method

2.1 Subjects

The subjects were six males and six females, ranging in age from 18 to 65 years old. The distribution of subject ages was similar across gender. Subjects were recruited by company employees, and were paid \$15 for the one-hour test session. The subjects’ occupations included student, homemaker, retiree, clerical, salesperson, and business planner. None were computer engineers, but some do use computers.

2.2 Design

The experiment consisted of three tests: Panning Elicitation, Panning Accuracy, and Panning Preference. Since gender differences on performance measures like these have been observed in other studies, all three tests were designed to allow gender effects to be distinguished from individual differences.

2.2.1 Panning elicitation test

The six subjects of each gender were randomly assigned to one of two display types—plain or bordered—in such a way that three subjects of each gender had each display type (see Table 1). The dependent measure was the type of panning user interface the subject suggested (see Procedure and Materials), i.e., a categorical variable. However, as is described in the Analysis and Results section, the response categories were not predetermined, but rather emerged as the study progressed.

Display Type	Male	Female
Plain	3	3
Bordered	3	3

Table 1. Panning Elicitation Test Design

2.2.2 Panning accuracy test

Because high intersubject variability was expected in the panning accuracy task, a within-subject design was used for that part of the experiment. Each subject was tested with all three panning user interfaces (Push Background, Push Camera, and Touch Edge Camera), with the order of the interfaces counterbalanced across subjects both within gender and overall. For each panning user interface, there were twenty-three trials (see Procedure). The following dependent measures were taken on each trial: time, number of moves, and whether the subject started shifting in the wrong direction (direction errors). Thus, the design of the panning accuracy test was a three (panning user interface) by two (gender) design. Panning user interface was varied within subject and the order in which the three panning user interfaces were presented to each subject was counterbalanced across subjects (see Table 2).

Interface Order	Male	Female
Push Camera, Touch Edge Camera, Push Background	1	1
Push Camera, Push Background, Touch Edge Camera	1	1
Push Background, Push Camera, Touch Edge Camera	1	1
Push Background, Touch Edge Camera, Push Camera	1	1
Touch Edge Camera, Push Background, Push Camera	1	1
Touch Edge Camera, Push Camera, Push Background	1	1

Table 2. Panning Accuracy Test Design

2.2.3 Panning preference test

The panning preference test was a very simple design: Ask all 12 subjects (six of each gender) which panning user interface they preferred of the three they had tried, to determine whether any interface is systematically favored or disfavored.

2.3 Procedure and materials

The display was a small (15 cm diagonal) color LCD display mounted in a flat case that added about 2 cm of border to each edge of the display. A transparent touch-pad was affixed to the front of the display. The display housing was mounted on a pedestal, which held the display about four inches above the surface of a desk (see Figure 2). The display and touch-pad were connected to a SunTM SPARCstationTM 2 workstation.

Subjects participated in the study individually. A video camera was focused on the display for the entire session, recording what happened on the display as well as

the subject's hand when near or on it. Each session was conducted by an experimenter and an assistant (to operate the camera and a timer).

Subjects were told at the start of the session that we (the experimenters) “are testing various design ideas for our product, to make the product easy to use.” The experimenter explained that some aspects of the prototype might be hard to use, but if so, that indicated a bad design rather than anything wrong with the subject. Subjects were seated in front of a prototype touch-sensitive display, and their attention was directed to it. The session consisted of three phases: Panning Elicitation, Panning Accuracy, and Panning Preference.

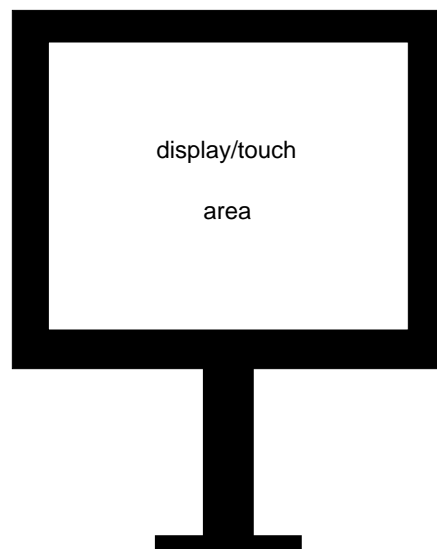


Figure 2. Touch-Display Apparatus

2.3.1 Panning elicitation

A cartoon scene of a living room was visible on the display. For half of the subjects, the living room display occupied the full area of the display; for the other half, the display had a quarter-inch, bright blue border that decreased the area available for the living room scene.

Subjects were shown that the living room scene was wider than the display (by panning the scene via keyboard commands to the right and then back to the initial position). In other words, they were shown the panning function without being shown a panning user interface. They were then asked to show and tell how they would expect to operate the touch-display to “bring that other part of the scene

back into view.” Words such as “pan” that might suggest a particular mental model or user interface were intentionally not used in the instructions. Subjects’ responses were categorized (see Analysis and Results).

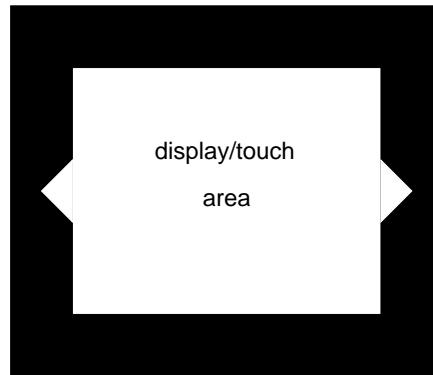


Figure 3. Alternate Touch-Display Cover

Next, an extra cover was placed over the display that decreased the screen size by approximately 1/4” on each side, but left two fingertip-sized sideways V-shaped notches at the midpoint of the left and right sides of the screen. The notches exposed the edge of the screen in those two spots (see Figure 3). Subjects were asked whether the cover suggested anything different to them about how to shift the scene. Again, their responses were categorized (see Analysis and Results). The extra cover was then removed.

2.3.2 Panning accuracy

The experimenter placed a sticker on the display’s bottom bezel that marked three horizontal positions: “A,” “B,” and “C.” Then the experimenter started a program that changed the displayed image from the living room scene to a set of evenly spaced vertical lines, each labeled by a number (0-10 from left to right) at its top end (see Figure 4). The experimenter directed the subject’s attention to the numbered lines and said that in this part of the test, he would ask the subject to shift the scene so that certain numbered lines were over specified letters. The experimenter explained that the subject would be timed while shifting the scene to the target letter. The experimenter continued: “We’ll do that several times, then I’ll change the way the scene is shifted and ask you to shift it some more; then we’ll change it again and I’ll ask you to shift it some more.” The subject was reminded that we were testing our designs, not the subject.

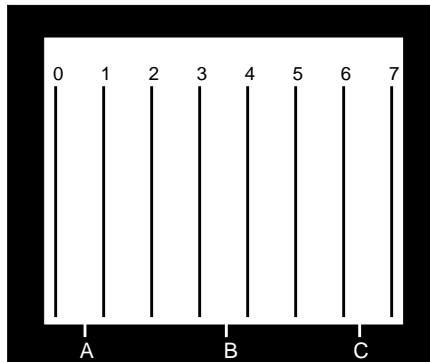


Figure 4. Panning Accuracy Test Display

The experimenter then demonstrated to the subject how to operate the first of the three panning user interfaces. Subjects were not allowed to practice panning before starting the timed trials. Subjects were instructed to move the line given on each trial to the indicated letter position, and once the line was positioned to his or her satisfaction, to say “OK” so that the assistant would know when to stop the timer. Twenty three panning trials followed (i.e., all lines moved to all reachable targets) in which the experimenter stated the line number and target letter, the assistant timed the trial, and the experimenter recorded the time on a data sheet. All trials were videotaped to allow rechecking of the times and collection of other data.

After the first set of panning accuracy trials, the experimenter changed the panning method, demonstrated the new panning method, and commenced a second set of trials with a different panning user interface. These were followed by a set of trials with a third panning user interface.

2.3.3 Panning preference

After completing all three sets of panning accuracy trials, subjects were asked “Which of the three shifting methods do you like best?” Their answers were recorded, both on the data sheet and on videotape.

3 Analysis and Results

3.1 Panning elicitation

When subjects were asked to indicate how they would operate the touch-display to bring into view the part of the living room scene that was off-screen-right, they gave a variety of responses. Note that at this point, subjects had not yet seen any

actual panning method for the prototype display (though some had used scrolling mechanisms on computers). All they had seen was that the living room scene was wider than could be seen in the display at once, and that the portion in view was the left side of the scene. Thus, their responses may be considered to be based upon their prior experience and whatever the prototype display suggested to them.

Subject responses were categorized initially into five categories corresponding to the four panning methods under consideration, plus an "Other category." The panning categories were:

- Push Background: finger motion on screen, shifting the background
- Push Camera: finger motion on screen, panning the viewer/camera
- Touch Edge Camera: press edge of screen that camera is to pan towards
- Touch Edge Background: press edge of screen that background is to shift towards
- Other: anything else, e.g., scrollbar, case buttons (i.e., physical scrolling buttons on the case), and miscellaneous

Six (i.e., half) of the twelve subjects indicated by their actions and words that Push Background was the panning method they first thought of when faced with the prototype touch-display. Three subjects suggested Touch Edge Camera. The remaining three subjects suggested methods that were classified as Other. A statistical test of the skewedness of the observed distribution requires a null hypothesis specifying what distribution would be expected by chance. Because a null hypothesis of equal category probabilities seems naive, such a statistical test wasn't feasible. Nonetheless, it is notable (if not "significant" in the formal sense) that half of the subjects suggested Push Background, and that no subject initially suggested Push Camera or Touch Edge Background (see Figure 5).

After subjects gave their initial suggestions for how panning should work, the notched cover was placed over the display, changing its appearance. Subjects were then asked whether the new appearance of the touch-display suggested anything different to them regarding how they might expect to shift the scene. In response to this question, seven of the twelve subjects suggested one of the two Touch Edge methods as the panning method suggested by the display. In other words, the notched display cover apparently provided an "affordance" that suggested Touch Edge panning to subjects. Six of these subjects said that one would bring the right side of the scene on-screen by touching the indented area on the right side of the display (i.e., Touch Edge Camera). These six included all three subjects who had originally suggested Touch Edge Camera (i.e., the notched display cover didn't affect their suggestion of panning method) and one each of subjects switching

from Push Background, Scrollbar, and Other. The seventh subject who suggested a Touch Edge panning method said that one would bring the right side of the scene on-screen by touching the indented area on the left side of the display (i.e., Touch Edge Background). The remaining five subjects stuck with their original suggestions. Four of these felt that Push Background was the most natural panning method, and one fell into the Other category.

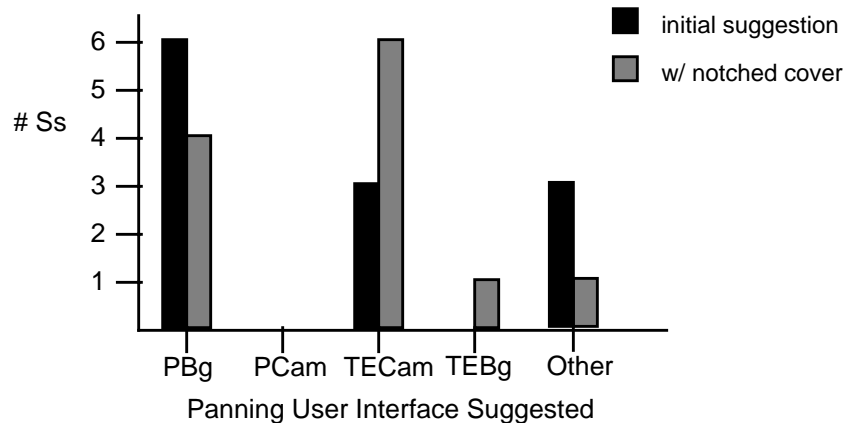


Figure 5. Panning Elicitation Results

The preceding analysis treated all 12 subjects as one group. However, the scene shown to half of the subjects had a bright blue border, which might suggest a Touch Edge panning user interface. With the notched cover on, the blue border was visible only in the notches. Orthogonally to this grouping, half the subjects were of each gender. Neither the presence/absence of a blue border, nor gender, appeared related to subjects' suggested panning user interfaces.

3.2 Panning accuracy

Each subject's time scores were averaged over each trial block, yielding, for that subject, an average time for each of the three panning user interfaces. Similarly, each subject's moves scores were averaged for each of the panning user interfaces. Direction errors were summed for each panning user interface. For each dependent measure, a Friedman ranks test for matched scores (Hays, 1963) was used to test for an effect of panning user interface type. To perform this test, the three scores for each subject were ranked; then the rank scores were submitted to a formula that yields a chi-squared score. The chi-square statistic indicates whether any interface had more low or high ranked scores than would be expected by chance.

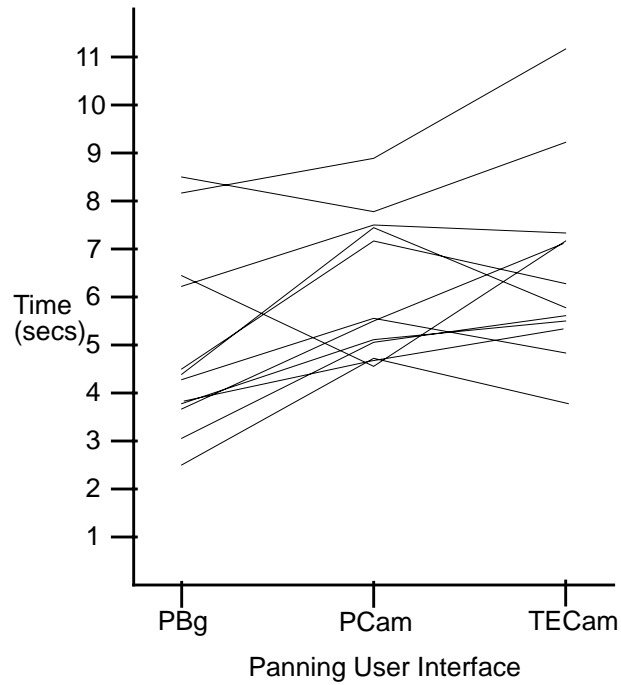


Figure 6. Panning Accuracy: time data (average over 23 trials)

Subject	Gender	PBg	PCam	TECam
1	M	3.06	5.03	5.66
3	“	6.48	4.66	7.18
7	“	4.58	7.16	6.33
8	“	6.29	7.48	7.35
9	“	8.14	8.9	11.22
12	“	4.44	7.45	5.83
2	F	4.27	5.69	4.91
4	“	8.57	7.88	9.20
5	“	3.88	4.07	4.57
6	“	3.71	5.49	7.08
10	“	3.81	4.65	5.29
11	“	2.53	4.79	3.88

Table 3. Panning Accuracy: time data (average over 23 trials)

For all three measures (i.e., time, number of moves, and direction errors), a significant effect of interface type was observed ($p < .01$; see Tables 3-5, and Figures 6-8). With the Push Background panning user interface, subjects took less time, required fewer moves, and made fewer direction errors than with the other two panning user interfaces. The effect was particularly strong for the direction errors measurement: when using the Push Background, subjects made almost no direction errors, in sharp contrast to the other two panning user interfaces (see Table 5 and Figure 8). A *post-hoc* t-test of difference scores between the Push Camera and Touch Edge Camera interfaces showed no significant difference between those two interfaces for any of the dependent measures.

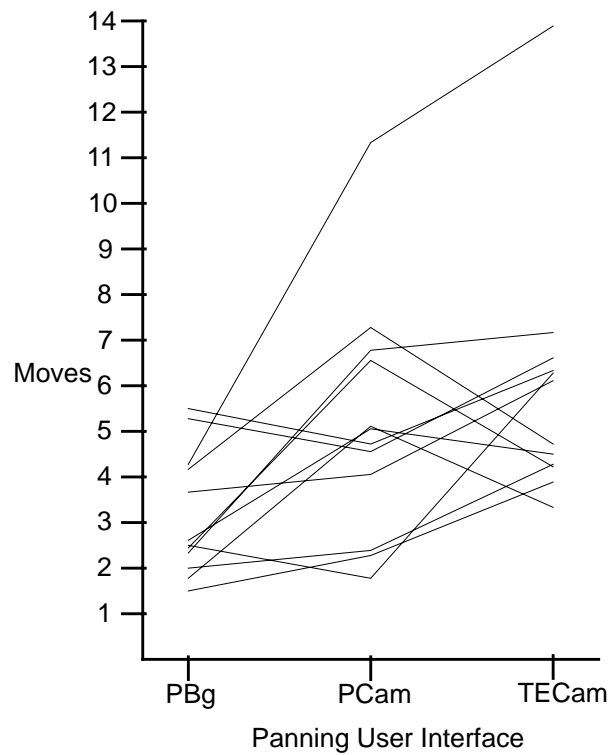


Figure 7. Panning Accuracy: moves data (average over 23 trials)

Subject	Gender	PBg	PCam	TECam
1	M	1.48	2.26	3.96
3	“	2.57	1.74	6.26
7	“	2.39	6.87	7.13
8	“	4.26	11.35	13.96
9	“	5.57	4.70	6.30
12	“	2.43	6.61	4.22
2	F	4.13	7.30	4.70
4	“	5.22	4.61	6.65
5	“	3.74	4.04	6.13
6	“	2.57	5.09	4.57
10	“	2.09	2.39	4.30
11	“	1.87	5.09	3.30

Table 4. Panning Accuracy: moves data (average over 23 trials)

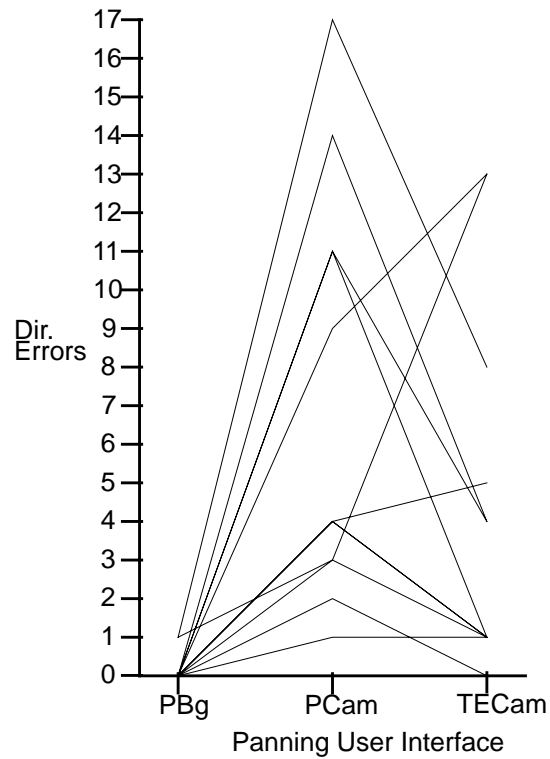


Figure 8. Panning Accuracy: direction-error data (sum over 23 trials)

Subject	Gender	PBg	PCam	TECam
1	M	0	4	5
3	“	0	3	1
7	“	0	11	1
8	“	1	17	8
9	“	0	4	1
12	“	0	2	0
2	F	0	11	4
4	“	0	1	1
5	“	0	4	1
6	“	1	3	13
10	“	0	9	13
11	“	0	14	4

Table 5. Panning Accuracy: direction-error data (sum over 23 trials)

No gender differences appeared on any measure, either for overall performance or effect of panning user interface type.

One possibility worth checking is that the three panning user interfaces might differ in how performance improves with practice. For example, performance with one user interface might start out worse than with others, but improve faster. Learning effects were examined by comparing, for each subject, performance in the first half of the trials with that in the second half. Since each user-interface-block of trials consisted of an odd number (23) of trials, the “first half” was defined as trials 1-11, and the “second half” was defined as trials 13-23. Trial 12 was ignored.

Overall, there was clear evidence of learning over the trial blocks. Superimposing the trial blocks for the three panning user interfaces, performance in the second half of the trials in a block almost always exceeded that in the first half. Simple sign tests were significant for all three performance measures: time ($p < .01$), number of moves ($p < .02$), and direction errors ($p < .01$).

To determine whether the learning rate depended on the panning user interface, subject’s difference scores (first-half-second-half) were submitted to a Friedman ranks test. For all three performance measures (time, moves, and direction

errors), the test showed no significant difference in learning between the three interfaces.¹

It would be interesting to know whether panning time and number of moves depends upon the distance panned, and whether the dependency is affected by the panning user interface. It might be, for example, that panning time is directly proportional to distance for one panning user interface but not for another. The exact panning distance required for each panning trial was known, so, for each subject, regressions were computed for distance vs. time and distance vs. moves for each of the three panning interfaces.

Overall, both time and number of moves were positively related to the panning distance. This was determined by a simple sign test on the regression scores: many more of them were positive than would be expected by chance ($p < .01$). On the other hand, Friedman ranks tests applied to the regression scores showed no effect of panning interface type on the strength of the relationship between distance and either time or number of moves.

3.3 Panning preference

Two of the twelve subjects, one male and one female, were accidentally not asked which of the panning user interfaces they preferred of those they had tried. Of the ten remaining subjects, eight preferred the Push Background interface, two preferred the Touch Edge Camera interface, and none preferred the Push Camera interface. A multinomial calculation showed that this distribution is significantly skewed ($p < .01$).

Of the ten subjects who were asked to state a preference, five were male and five were female. The distribution of preference scores was exactly the same for males as for females: four out of five preferred Push Background, one out of five preferred Touch Edge Camera. Thus, no gender difference in preference was observed.

4 Discussion and Conclusions

4.1 Panning elicitation

Push Background seems to be the single most common way people expect to pan a touch display if there is nothing about the design that suggests anything else.

1. No visible decrease in direction errors was possible with the Push Background user interface, because direction errors were almost nonexistent to begin with.

However, the notched casing shifted subjects' intuitions to Touch Edge panning. A visible border on the display, however, did not seem to be a sufficient affordance for Touch Edge panning.

4.2 Panning accuracy

Push Background panning was the clear winner of the accuracy test: it resulted in faster times, less effort, and fewer errors than the other two panning methods.

It is possible that Touch Edge Camera would not do so badly against Push Background if the task were grosser, e.g., just moving something offscreen on, or just moving something to the left half of the screen. Touch Edge (Camera or Background) is just a difficult interface for panning to an exact position. A less exact panning task is thus a good candidate for a follow-up study.

4.3 Panning preference

Push Background was also clearly the preferred panning method for most subjects. However, Touch Edge Camera is preferred by some. One possible reason for this is that Touch Edge panning requires less physical movement on the users' part than Push panning does.

4.4 Overall

Overall, Push Background seems to be the best—including the most intuitive—of the three panning user interfaces examined in this study. Subjects' preference generally matches their initial impressions and performance scores. The results strongly suggest that the decision to use Push Camera panning in the prototype based on the informal paper-and-pencil study was wrong.

Touch Edge Background (i.e., touch the side that you want the background to move towards) was not included in this study because informal experience with scrollbars indicates (at least to the author) that it is counter-intuitive. However, the number of direction errors made with Touch Edge Camera and some subjects' comments that it is "backwards" suggest that not everyone would find Touch Edge Background to be a counter-intuitive panning method.

As a follow-up to the present study, we will consider a test in which subjects see a scene pan (without seeing a panning *user interface*, as in the present study) and then are asked: "How do you prefer to think of what just happened: Did the room shift to the left or did the camera view shift to the right?" In other words, subjects would be given a forced choice between two ways of thinking of panning instead of an open-ended question.

5 References

- Beard, D.V., Brown, P., Hemminger, B.M., Misra, R. (1991) "Scrolling Radiologic Images: Requirements, Designs, and Analysis," *Proc. Intl. Symposium on Computer Assisted Radiology (CAR'91)*, West Germany: Springer-Verlag, pages 636-641.
- Beck, D., Elkerton, J. (1988) "Development and Evaluation of Direct Manipulation Lists," *Proc. ACM Conference on Computer-Human Interaction (CHI'88)*, Washington, DC, pages 72-78.
- Glassner, A.S. (1990) "A Two-Dimensional View Controller," *ACM Transactions on Graphics*, January, 9 (1), pages 138-141.
- Hayes, W. (1963) *Statistics*, New York: Holt, Rinehart and Winston.
- MacLean, A., Young, R.M., Moran, T.P. (1989) "Design Rationale: The Argument Behind the Artifact," *Proc. ACM Conference on Computer-Human Interaction (CHI'89)*, Austin, TX, pages 247-252.
- Norman, D.A. (1988) *The Psychology of Everyday Things*, New York: Basic Books, Inc.
- Plaisant, C. and Sears, A. (1993) "Touchscreen interfaces for alphanumeric text display," in B. Shneiderman (Ed.), *Sparks of Innovation in Human-Computer Interaction*, Norwood, NJ: Ablex Publ.
- Potter, R.L., Weldon, L.J., Shneiderman, B. (1993) "Improving the accuracy of touchscreens: an experimental evaluation of three strategies," in B. Shneiderman (Ed.), *Sparks of Innovation in Human-Computer Interaction*, Norwood, NJ: Ablex Publ.
- Sears, A., and Shneiderman, B. (1993) "High precision touchscreens: design strategies and comparisons with a mouse," in B. Shneiderman (Ed.), *Sparks of Innovation in Human-Computer Interaction*, Norwood, NJ: Ablex Publ.
- Shneiderman, B. (1993) "Touchscreens now offer compelling uses," in B. Shneiderman (Ed.), *Sparks of Innovation in Human-Computer Interaction*, Norwood, NJ: Ablex Publ.
- Sullivan, J. (1993) "Magic Cap," informal demonstration at Sun Microsystems, Inc., March.

The Effect of Touch-Pad Size on Pointing Accuracy

Jeff Johnson and Mark Keavney

FirstPerson Technical Report FP-1994-2

April 1994

Abstract:

Two experiments were conducted to determine the smallest touchpad that subjects could accurately use to point to objects on a TV-sized video display. The touchpads were on a remote control, separate from the TV display. Three touchpad sizes were compared. In Experiment 1, fifteen subjects were divided into three experimental groups: one for each pad-size. The task consisted of attempting to hit specified lettered "buttons" on the television display. A "land-on" criterion was used to determine hit/miss. Subjects poked at each button repeatedly until the button registered a "hit"; the number of pokes required to hit each button was the dependent measure. There was a significant effect of pad size on subject accuracy, but not in the expected direction: accuracy was worst with the mid-sized pad, and the smallest and largest pads were not significantly different. Experiment 2 was conducted to check these results using a within-subject design and to correct some apparatus problems that may have affected the results of Experiment 1. Twelve subjects used both the mid-sized and the small pad-size, in counterbalanced order. No significant difference in accuracy was seen between the mid-size and small pads. Also, there was no noticeable learning observed between the two blocks of trials, corroborating a lack of learning over trials seen in Experiment 1.

The Effect of Touch-Pad Size on Pointing Accuracy

Jeff Johnson and Mark Keavney

1 Background and Rationale for Study

A new television remote control was being designed that used a touchpad as the primary interactive component, with a separate display on the TV set. The size of the touchpad had to be decided. From the perspective of design aesthetics and cost, the smaller the touchpad, the better. However, it seemed likely that users would have more trouble mapping positions on a smaller pad to positions on the TV screen. The smaller the pad, the more precise users would have to be in order to hit a target of a certain size. Also, since the contact area between users' fingers and the touchpad is not a point, the smaller the pad, the greater the portion of the display screen represented by the contact area, further degrading accuracy at smaller sizes. We had to find out how small the pad could be and still be usable.

From informal pilot testing, we knew that a pad 2.25" wide and 1.75" tall was usable, and that a pad less than 1.5" wide and 1.15" tall was not. We tested three pad sizes within these limits, using a "land-on" criterion (Potter, Weldon, and Shneiderman, 1993) for determining whether a target had been hit. We were interested in how well users could map from the touchpad to the display screen with these different sizes (i.e., how easily they could determine the position on the pad to press to hit targets on the display). We expected that the larger pads would be easier to use, but we wanted to determine the overall level of performance and how much of a difference there was between the different sizes.

2 Experiment 1

2.1 Subjects

The subjects were fifteen FirstPerson employees. They all had some experience using computers, but little or no experience using a separate touchpad to control a display screen. Four subjects were female; 11 were male. Subjects were mostly between 35 and 45 years old, with a few in their early thirties or mid-fifties.

2.2 Materials

The touchpad used was about five inches high and seven inches wide. To simulate three touchpad sizes, software was written that mapped one of three different-sized rectangles in the bottom-right corner of the touchpad to the entire display. Cardboard templates were created to cover the unused portion of the touchpad. The largest pad-size tested was 2.1" x 1.52"; the medium and small pad-sizes had the same proportional shape, but were 1.9" and 1.73" wide, respectively. The touchpad was mounted on a remote-control-like wooden handle so that while holding the handle, a user could reach the exposed portion of the touchpad with his or her thumb.

The display was a large (25" measured diagonally) ordinary TV screen connected to a Sun workstation with some special-purpose hardware.

Software was written to present an array of rectangular "buttons," each labeled by a letter of the alphabet (A-Y). The 25 buttons were arranged on the display in a 5 x 5 matrix. The software reacted to a touch on the touchpad by placing a pointer (a circled +) at the same relative position on the screen until the finger was removed from the pad. The software reacted to the pointer landing on a lettered button by sounding a tone and causing the button to expand briefly. A "land-on" criterion (Potter, Weldon, and Shneiderman, 1993) was used to determine whether a button had been hit, i.e., a touch was regarded as a hit only if its point of initial registration was on a button. Thus, touching just off of a button and then moving the touch-point onto the button did not count as a hit.

2.3 Design

The 15 subjects were assigned randomly to one of the three pad-size conditions, with five subjects in each condition. Since there was neither interest in, nor expectation of, gender differences, no attempt was made to balance gender across conditions.

Each subject repeated the basic task (trying to hit a specified target) 50 times (twice for each of the 25 letter targets). Each repetition of this task is one trial. The order of the trials was random, but was the same for each subject.

The subject's score for each trial was the number of attempts the subject needed to hit the specified target. When a subject had made 20 unsuccessful attempts to hit a target, the experiment proceeded to the next trial. Perfect performance would yield a score of 1 (i.e., the subject hit the target in one attempt), and the worst possible score on any trial was 20.

2.4 Procedure

Subjects were tested individually. They were seated before the TV display, which showed the 5 x 5 array of the letters A-Y. A video camera was set up and aimed at the TV screen to record display events.

Subjects were handed the touchpad-handle assembly configured as either a large, medium, or small touchpad. They were instructed to hold the handle so that they could touch the exposed area of the touchpad with their thumb. They were shown the effect of touching the display and of “hitting” one of the lettered “buttons.”

Subjects were told that they would be asked to try to hit specified lettered “buttons.” They were asked to do this by poking with their thumbs at the pad, and if they missed, lifting and poking again until they hit the target. They were asked not to press and drag their thumb to the target, but rather to poke. They were told that the experimenter would count the number of attempts required to hit the target.

On each trial, the experimenter announced a letter; the subjects’ task was to touch the touchpad so that the pointer “hit” the button labeled with that letter. Subjects were allowed to look at the touchpad or at the TV screen. If necessary, they were reminded not to drag their thumb on the pad, but only to poke at the pad, at what they thought was the appropriate pad location. The experimenter recorded the number of times a subject “poked” before hitting the correct target. Each subject was given 50 trials, two for each letter on the display.

When the subjects had completed the 50 trials, they were thanked for their participation and given chocolate as a reward.

2.5 Results and discussion

During the experiment, the experimenter noticed that the subjects seemed to have great difficulty hitting targets on the upper row and left column of the display screen. Subjects did have significantly more trouble hitting these targets than the others ($p < .01$). Also, pad size affected this difference: the difference was greatest for the medium pad ($p < .05$). The upper and left edges of the usable surface of the touchpad bordered on the cardboard template that covered the rest of the pad. It is possible that there was a problem calibrating the software with the template, so that the part of the pad that activated the topmost and leftmost targets was partly covered by the template. This would explain the subjects’ poor performance on these targets. In any case, all analyses were performed both with and without the topmost and leftmost targets to control for any possible artifacts resulting from this problem.

The results of the pad size comparison were surprising. There was a significant effect of pad size on performance ($p < .05$), but it was not in the predicted direc-

tion (see Figure 1). We had expected that subjects would perform best with the large pad and worst with the small pad. In fact, subjects who used the large pad hit their targets in the fewest number of attempts (3.55); those who used the medium pad required the most (6.98); and the small pad was inbetween (4.42). Contrary to expectations and common sense, the subjects found it easier to use the small pad than the medium pad. These results held even when the topmost and leftmost targets were excluded from the analysis: there was still a significant effect of pad size on performance, with the medium pad still the hardest to use (5.27), followed by the small (3.96) and large (2.79) pads.

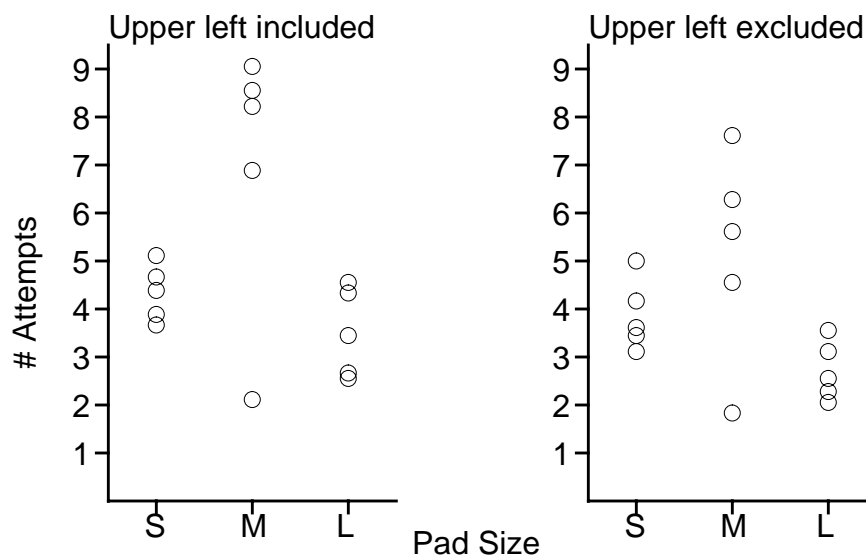


Figure 1. Number of attempts required to hit target vs. pad size

Another surprising result is that there were no detectable effects of practice. Although the subjects performed slightly better on the last 25 trials of the task than on the first 25, this difference was small and did not approach significance. These results also held when the topmost and leftmost targets were removed from the analysis. Also, there was no practice by pad-size interaction, i.e., the subjects didn't learn any better or worse with any particular pad size.

There were large individual differences in how well the subjects performed the task: several subjects averaged fewer than 3 attempts per trial, but most averaged more, and some averaged as high as 9 attempts per trial.

3 Experiment 2

3.1 Rationale

The result of the pad size comparison in Experiment 1 was that the medium pad was the hardest to use, followed by the small pad and the large pad. This result was unexpected and difficult to understand. The pad sizes were not greatly different, and it is hard to believe that there should be such a dramatic increase, and then decrease, in difficulty of use within this small window of size. We suspected that this result was a Type I statistical inference error, i.e., just by chance the subjects who had the most trouble with this task were assigned to the medium pad. Therefore, we conducted a follow-up study to see if this finding would replicate.

In order to increase our power and prevent individual differences from giving us misleading results, we decided to run Experiment 2 using a within-subject, rather than a between-subject, design. Rather than, as in Experiment 1, having each subject perform the task with a single pad size and then comparing the scores between the different groups of subjects, each subject in Experiment 2 performed the task with several different pad sizes, and we compared the scores within each subject. This design has the advantage of eliminating the error variance caused by individual differences in task performance. It is ideal for situations (like this one) where there are large individual differences in performance.

The reason that we didn't design Experiment 1 in this way was that we expected practice effects, and a within-subjects design is somewhat problematic in that case. However, despite our expectations, there were no significant practice effects in Experiment 1, so we felt comfortable running Experiment 2 using this design.

Because the most surprising result of Experiment 1 was that the medium pad was harder to use than the small pad, only those two pad sizes were included in Experiment 2. Each subject performed the same task as in Experiment 1 twice, once with the small pad size, and once with the medium pad size. If the medium pad was again significantly more difficult to use than the small pad, then the results of Experiment 1 would be replicated and we would have confidence in them. If, on the other hand, the medium pad was not significantly more difficult to use than the small pad, then we would believe that the results of Experiment 1 were suspect.

In addition to testing this general hypothesis about the usability of the pad sizes, we also wanted to see if there were reliable subject differences which might be related to the subjects' performance with the different pad sizes. For example, subjects who have trouble hitting targets overall might perform best with one pad size, while subjects who are more accurate might do best with a different pad or might show no difference between pad-sizes. In order to test hypotheses about

subject differences, we re-tested two groups of Experiment 1 subjects in Experiment 2: six who had found the task relatively easy in Experiment 1, and six who had found it more difficult.

3.2 Subjects

The subjects were 8 men and 4 women, all employees of FirstPerson. All had previously participated in Experiment 1. Six of the subjects had had difficulty hitting the targets in Experiment 1, and six had hit targets relatively easily.

3.3 Design

Each of the 12 subjects performed the Experiment 1 task twice, once with the small pad size, and once with the medium pad size. The order of pad size was counterbalanced, so that half of the subjects used the small pad and then the medium pad, and half used the medium and then the small. The order was also counterbalanced within the high-performance and low-performance groups, so that half of the subjects in each group received the medium pad, then the small; the other half received the small pad, then the medium. Within these limits, the subjects were randomly assigned to the pad ordering.

As before, the subject's score for each trial of the task was the number of attempts the subject needed to hit each target. Performance on a given trial ranged from 1 (the best) to 20 (the worst). The order of the trials was the same as in Experiment 1.

3.4 Materials

The materials used were the same in Experiment 2 as they had been in Experiment 1, except that the large pad template was not used. Only the medium and small pad sizes were tested. In addition, the software that mapped the touchpad to the display screen was revised so that the uppermost and leftmost targets were further from the edge of the template and thus easier to hit.

3.5 Procedure

The procedure used in Experiment 2 was identical to that of Experiment 1, except that the entire set of trials was presented to each subject twice—once with each pad size—instead of only once with a single pad size.

3.6 Results and discussion

In Experiment 1, accuracy on the targets in the upper row and left column was significantly worse than accuracy on the other targets. For Experiment 2, we recalibrated the software, moving the uppermost and leftmost targets in an attempt to

solve this problem. The results of Experiment 2 showed no significant difference in accuracy for the upper and left columns vs. the rest of the targets. Therefore, our attempt to recalibrate these targets was apparently successful. All the analyses for Experiment 2 include these targets.

The results of the pad size comparison were somewhat complicated, but it is clear that the main result of Experiment 1 failed to replicate. Initial analyses suggested that subjects performed significantly better with the medium pad than with the small pad (3.03 vs. 3.28, $p < .05$), contradicting Experiment 1 but more in line with common sense. However, closer examination revealed that most of the difference was attributable to a single target: using the small pad-template, most subjects had difficulty hitting the letter M (in the middle of the screen). No such difficulty was seen with the medium pad-template. When we noticed, part way through the experiment, that subjects seemed to be having a hard time hitting the M, we tried it ourselves and convinced ourselves that there was truly some sort of problem with that target. We concluded that the touchpad used for most of the experiment had a flaw that corresponded to the M-target with the small pad-template, and fell between targets with the medium pad-template. The difficulty in hitting the M-target with the small template disappeared when, for other reasons, we switched to a different touchpad late in the study, supporting our belief that there had been a flaw in the first pad. Based on this belief, we removed the M target from the analysis, and found that most of the difference between the medium and small pad disappeared. There was still a small advantage for the medium pad (2.97 vs. 3.00), but the difference wasn't significant. In sum, there was no evidence that the subjects performed better with the small pad, unlike the result of Experiment 1.

Experiment 2 did replicate the lack of learning that was observed in Experiment 1: on average, subjects performed no better on their second block of trials than on their first (3.10 vs. 3.21, N.S.). It is possible that given enough practice, subjects' performance would begin to improve, but that doesn't seem to occur within the time it takes to complete this study (about 20 minutes).

Though the subjects had been sorted into two groups based upon their accuracy in Experiment 1, there was no difference in overall performance between the two groups in this study. There are three possible explanations for this:

1. Performance on this task may not be stable. Performance may depend more on factors that vary within individuals (such as mood or level of alertness) than on factors that are stable within individuals (such as hand-eye coordination).

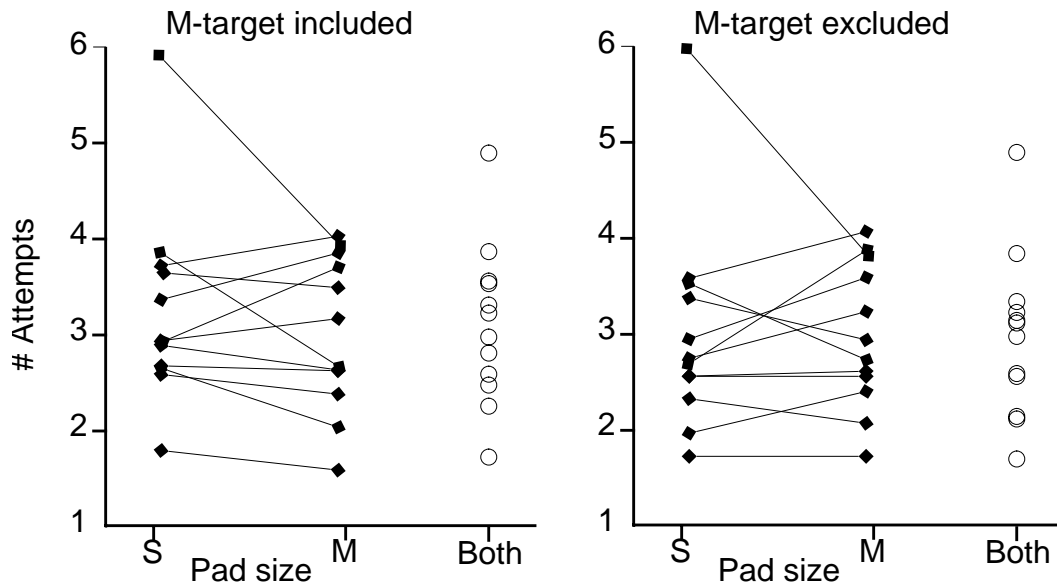


Figure 2. Number of attempts required to hit target (connected for each S) vs. pad size

2. Despite our having observed no learning within either study, perhaps the “low-performing” subjects improved between studies, catching up with the others.
3. The variation between subjects in Experiment 1 may have been caused by differences in how well they could hit the upper and leftmost targets (which were the hardest targets to hit in that experiment). Because the second study fixed the design problem that made the top and left targets hard to hit, that difference between subjects may have gone away, making people more equal.

We prefer the third of these explanations.

Although there was no difference in overall accuracy between the “low” and “high” performing subjects from the first study, there was a significant difference between the two groups as to which pad-size they did better with. The previous high-performers did better with the medium pad (2.70 attempts for medium vs. 3.45 for small), and the previous low-performers did better with the small pad (3.36 attempts for medium vs. 3.12 for small). We have no explanation for this.

4 Conclusions

Experiment 1 was designed to look, in the most straightforward way possible, for accuracy differences between the three pad-sizes. The results, however, were inexplicable. Either a positive relationship between pad-size and performance or no relationship at all would have seemed more sensible than the observed ordering of accuracy scores. The perplexing outcome of Experiment 1, and the problem subjects had hitting the topmost and leftmost targets, combined to suggest replicating the experiment (after correcting the target positioning). The large individual differences and lack of learning over trials seen in Experiment 1 suggested that the follow-up study should have a within-subject design. Thus, Experiment 2 was conducted.

The data-analysis for Experiment 2 was complicated by an apparatus problem, namely, a flaw in the touch pad that differentially affected performance in different conditions. However, once this problem was identified and the affected trials removed from the analysis, the result—no effect of pad-size upon accuracy in hitting targets—seems reasonable, even though a positive relationship had been expected. This result suggests that, within the range of pad-sizes tested, it doesn't matter much which size is chosen for inclusion in the first prototype remote-controls. The medium-sized pad fit the industrial design of the remote control casing best, and the results of Experiment 2 indicated that choosing that size would at least not be a serious mistake, so the medium-sized pad was chosen.

5 References

Potter, R.L., Weldon, L.J. and Shneiderman, B. (1993) "Improving the accuracy of touchscreens: an experimental evaluation of three strategies," in B. Shneiderman (Ed.), *Sparks of Innovation in Human-Computer Interaction*, Norwood, NJ: Ablex Publ.

Factors Affecting Users' Ability to Hit Animated Characters

Mark Keavney

FirstPerson Technical Report FP-1994-3

January 1994

Abstract:

FirstPerson conducted an experiment to determine the effects of various factors on users' ability to hit an animated character (a sprite) moving across the TV screen, using a touchpad separate from the display. Subjects were repeatedly shown sprites moving across the screen and told to touch a touchpad at the location that corresponded to the sprite. Three factors were systematically varied within each subject: the strategy used to hit the sprite (land-on vs. track), the type of sprite (which included size, shape, and regularity of movement), and the speed at which the sprite moved. The distance that the sprite moved across the screen before it was hit, the number of attempts subjects required to hit it, and the location of the subjects' missed attempts were all measured. Results suggested that land-on is a better strategy than track, that speed makes more of a difference than size in subject performance, and that there is a tendency for subjects to hit slightly behind the moving sprite rather than directly on it.

Factors Affecting Users' Ability to Hit Animated Characters

Mark Keavney

1 Introduction

FirstPerson, Inc. was creating a demonstration of an animated interactive environment in which users could choose movies to watch in their homes. This was called the Video on Demand (VOD) demonstration. Part of this demonstration included animated characters called “extras,” which were walking advertisements for specific movies. While the user was exploring the environment, an extra would sometimes move onto the screen. When the user touched it, the extra would activate, giving a brief preview of its movie and offering the user the option to watch it.

The user interface for this demonstration was a touchpad that was separate from the display screen. In order to “hit” an extra, the user would have to touch the appropriate location on the pad. We needed to know how easily people could do this and how we could design the extras to make this easier. Therefore, we conducted a study in which the subjects used a touchpad to hit animated characters moving across the screen. We varied many of the important aspects of the animated characters (such as speed, size, and regularity of movement) to determine how these affected subject accuracy, and we analyzed the subjects' missed attempts to hit the extras in the hope that might show us how to make the extras easier to hit.

2 Method

2.1 Subjects

The subjects were nine employees of FirstPerson, seven men and two women. All had experience using computers, but none had experience hitting animated characters using a touchpad.

2.2 Materials

The touchpad used for the study was a 1.25" high by 1.85" wide area on the bottom right corner of a 4.5" high by 5.5" wide touchpad. The rest of the pad was covered with a white cardboard mask. The software mapped the unmasked area to the entire display screen (a 22" TV set). The touchpad was mounted on a wooden handle so that while holding the handle, the user could touch the pad with his or her thumb. The design of the touchpad is illustrated in Figure 1.

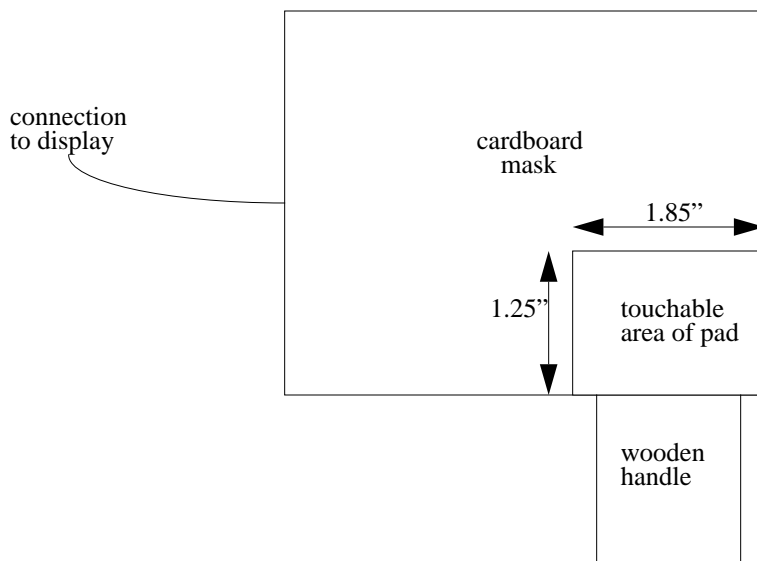


Figure 1. Diagram of touchpad. Only the area in the lower right corner was used in the study. The software registered touches in that area and displayed them on the screen.

The software used for the study displayed an animated character (a sprite) moving across the screen, registered touches on the touchpad, and displayed a cursor on the screen at the location corresponding to each touch. The program stopped the animation and gave auditory feedback when the user hit the sprite.

Each time the user touched the pad, the software recorded the location of the touch and the current location of the sprite. The sprite's location was recorded by measuring the center of its "bounding box," an invisible box which contained the sprite and moved along with it. Six out of the seven sprites were centered within their bounding boxes and did not move relative to the box, so the center of the bounding box was a good indication of the center of the sprite. As noted in the Results section, however, one of the animations did move within its bounding box, and so the measure of sprite location for that sprite may not be meaningful.

Although the record of the sprite's location was based only on the sprite's bounding box, the determination of a hit was not. In order for the program to register a hit, the user had to touch part of the sprite's image (i.e., a pixel with a non-zero alpha), not just part of the bounding box.

In writing the program, there was concern that a cursor that was not repainted quickly enough would be "sluggish" (i.e., it would lag behind the touch when the user moved his finger across the pad). Therefore, the program was designed to maximize the speed of the cursor. Because the environment was very simple (the background was small, there was no panning, and there was only a single sprite on-screen at any one time), it was possible to write a cursor that moved quickly enough that there was no noticeable sluggishness.

2.3 Design

Three factors were systematically varied within each subject. The first factor was the strategy used to hit the sprite. There were two possible strategies: land-on and track. The land-on strategy was to keep one's finger off the pad until the sprite appeared, and then to poke at the pad at the approximate location of the sprite. The track strategy was to track the sprite by moving one's finger across the surface of the pad, and then to quickly lift the finger and poke at the pad where the sprite was. Subjects were instructed to use the land-on strategy for half of the experiment, and the track strategy for the other half. The order of these two strategies was counterbalanced.

A previous study (Potter, Weldon, & Schneiderman, 1993) on the usability of touchable display screens (as opposed to touchpads) tested performance in hitting stationary targets using two strategies similar to those tested here. One of these strategies, land-on, was the same as the land-on strategy in this study. The other strategy, take-off, involved tracking just as the track strategy in this experiment, but the "sprite" was activated on an up motion rather than an up-down motion. (Take-off was not tested in the present study because it had already been decided that the extras in the VOD demo would be activated by a down motion.) Potter et al., found that subjects made significantly fewer errors using take-off than land-on, with no significant difference in time. However, it's not clear if those results will generalize to this experiment, which involves a strategy slightly different from take-off, a task with moving rather than stationary objects, and a touchpad rather than a touchable display screen.

The second factor in the study was sprite type. There were seven different animated characters in the study: three different-sized musical notes, a ball, a stick figure, a rectangular log, and Duke, the company mascot. These animations had

four different types of movement: the ball and three notes bounced diagonally from the top of the screen to the bottom; the log tumbled, turning around its center as it moved across the screen; the stick figure walked, somewhat irregularly; and Duke cartwheeled. Although the variation in sprite shape was somewhat unsystematic, we hoped to gain some information from a comparison of accuracy across the different sprites. For example, because the musical notes differed only in size and not in shape or movement, we hoped to see how sprite size affected accuracy.

The third factor was sprite speed. Each of the sprites was presented at three different speeds: fast, medium, and slow. Although the speeds varied somewhat from sprite to sprite, for all the sprites the slow speed was about half the fast speed.

These factors were fully crossed. Thus, there were a total of 42 trials for each subject: 2 (strategies) x 7 (sprites) x 3 (speeds). The trials were divided into two blocks of 21 trials, one for each strategy. In each block of 21 trials, the subject saw the different sprites at different speeds in an order that was random, but was the same for each block and each subject.

2.4 Procedure

The subjects were seated in front of the TV display and handed the touchpad. They were given one set of instructions (either the “land-on” or “track” instructions) for how to hit the sprite, and then they went through 21 trials using this strategy. On each trial, an animated character would appear somewhere on screen and move across the screen to the left. The subjects were instructed not to hit the sprite at this point. Eventually, the sprite exited the left side of the screen and, after a few seconds, reappeared on the right, still moving to the left. The subjects then tried to hit the sprite. Each time a subject touched the pad, the program recorded the position of the touch, the position of the sprite at the time of the touch, and whether or not the touch hit the sprite. If the touch hit the sprite, the sprite stopped. If the subject failed to hit the sprite before it moved all the way across the screen, after a brief delay, the sprite again appeared on the right and moved across the screen to the left. This continued until the subject hit the sprite or until the sprite moved across the screen 10 times.

After the subjects had completed 21 trials, they were given another set of instructions on how to hit the sprite, and went through another block of 21 trials. Once they had completed these trials, they were asked which of the two strategies for hitting the sprite they preferred. They were then thanked, debriefed, and given chocolate as compensation for their participation.

3 Results

3.1 General performance

In general, subjects had trouble hitting the sprites with either strategy. On average, the subjects needed 4.4 attempts (“pokes”) to hit the sprites, and the sprites traveled about 93% of the screen before they were hit. The sprites were hit on the first attempt only 20% of the time, and made it all the way across the screen without being hit about 15% of the time.

There were, however, large individual differences in performance: one subject was able to hit the sprites with only 2.2 attempts on average and when they had traveled across only 45% of the screen; whereas another subject hit them with 5.7 attempts and when they had traveled 160% of the screen (i.e., 60% into the next screen, on average).

There also seemed to be substantial individual differences in how the subjects reacted to the task. Some enjoyed the task and saw it as a game while others found it frustrating.

Although the overall level of performance in this study was below what is acceptable for users’ attempts to hit the VOD extras, there are at least two reasons why users should find it easier to hit the VOD extras than the sprites in this study. First, most of the sprites in this study were smaller and faster than the VOD extras. Second and more importantly, the remote control was a jury-rigged and improvised device that we expect to be much less usable than the VOD remote.

Because of these differences, the overall level of performance may not be informative. However, we can still look at variations in performance due to the different factors we manipulated. The variations that we find in this study may still be present in easier tasks (such as hitting the VOD extras).

3.2 Factors affecting performance

3.2.1 Strategy used

The first factor varied was the way in which subjects were instructed to hit the sprites. There were two different strategies: track and land-on. The results suggested that land-on was a more effective strategy. Seven out of the nine subjects hit the sprite earlier with land-on than with track, a marginally significant result ($t=1.83$, $p<.10$). Furthermore, out of the seven subjects who expressed a preference for a strategy, six preferred land-on.

However, this result may be due to the particular gesture model used in the experiment. According to this model, the sprites were only “hit” when the subjects pressed down. Therefore, when the subjects tracked the sprite, they had to pull their finger up and then press down, whereas when they were trying to “land on” the sprite, they only had to press down. If the sprite were hit by an up rather than a down motion (or even by a “press hard” motion in the case of gesture models that distinguish different finger pressures), a tracking strategy might be more popular and effective.

3.2.2 Sprite type

Another varied factor was the type of sprite. There were seven different sprites, of different sizes, shapes, and patterns of movement. There was a ball that bounced across the screen, three musical notes (small, medium, and large) that also bounced, a log that tumbled, a stick figure who walked (somewhat irregularly), and Duke, who cartwheeled. The two most irregular-moving and irregular-shaped sprites, the stick figure and Duke, were the hardest to hit: subjects required an average of 6.3 attempts to hit them, but only 3.8 attempts to hit the other sprites.

A comparison of the musical notes allows us to determine the effect of sprite size, since the notes all moved in the same way and with the same speeds. Not surprisingly, the larger the note, the easier it was to hit. However, this was a small effect: the largest note was hit with 3.5 attempts when it had traveled 77% of the screens, while the smallest note (which was less than half the size of the largest note) was hit with 4.1 attempts at 96% of the screen. This difference is not significant ($t=1.01$ for distance moved, $t=1.02$ for number of attempts). Given the small n of the sample, the fact that this difference isn’t significant doesn’t mean that it doesn’t exist. However, it does suggest that any accuracy differences due to size are likely to be small.

3.2.3 Sprite speed

Another factor was speed. Each sprite was presented at three different speeds. The slowest speed was about half of the fastest speed for most of the sprites. Speed made a large difference in how far the sprites traveled before they were hit. On average, subjects hit the fast sprites when they had traversed 130% of the screen, the medium sprites when they had traversed 95%, and the slow sprites when they had traversed 60%.

Speed also made a large difference in the number of attempts it took the subjects to hit the sprites: there were 5.4 attempts for the fast sprites, 4.8 for the medium sprites, and 3.1 for the slow sprites. The difference was in the same direction for

all of the subjects: every subject had a greater distance and more errors for the fast sprites than the slow sprites, making these differences statistically significant ($p < .01$ by sign test).

The speed of the sprite made little difference in the time it took subjects to hit it: the faster sprites covered more distance across the screen only because they were moving faster, not because they traveled for a longer time.

It is somewhat odd that the speed of the sprites affected the number of errors, but not the time required to hit them. Apparently the subjects “sped up” their attempts to hit the faster sprites, making more attempts in the same amount of time.

3.2.4 Order

The last factor was order. There was some indication of learning: 7 out of the 9 subjects performed better in the second block of trials than in the first block. However, this effect was weak and non-significant ($t = 1.14$, $p > .2$), suggesting that any learning effects are small compared to the between-subject differences.

3.3 Analysis of errors

3.3.1 Position of errors

Because the program recorded the position of each touch and the position of the sprite at the moment of each touch, we can look for a pattern in the subjects’ missed attempts. Since all of the sprites moved from right to left, the difference between the x-coordinate of the center of the sprite and the x-coordinate of the touch gives us an indication of whether the subjects are hitting at the center of the sprite, or whether they’re hitting behind or ahead of it. Some of these “errors” are misleading, since the touchpad was sensitive and occasionally recorded a touch when the subjects were clearly not trying to hit the sprite (for example, when the sprite was offscreen). However, we can eliminate most of these cases by only looking at errors within 120 pixels of the sprite (about half a screen). This distribution is shown below (see Figure 2):

As the graph shows, on average the subjects hit slightly behind the center of the sprite (positive numbers indicate touches behind the sprite’s center). Overall, the subjects were 13 pixels off center. Each individual subject’s average was positive (i.e., each subject, on average, hit behind the center), making this tendency statistically significant ($p < .01$ by sign test).

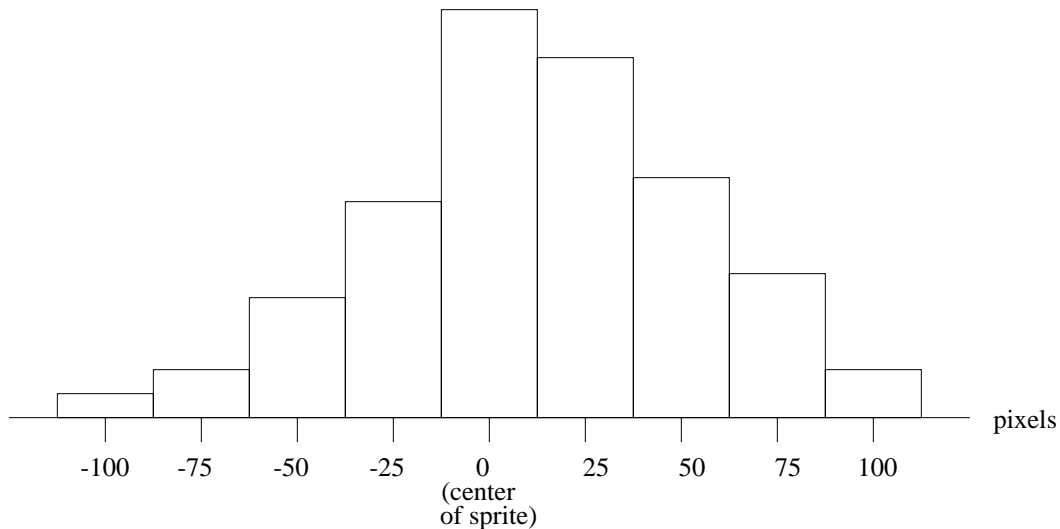


Figure 2. Distribution of subjects' unsuccessful attempts to hit the sprites. Positive numbers indicate misses to the right of (behind) the sprite's center; negative numbers indicate misses to the left of (ahead of) the sprite's center.

It is important to note that, on average, the touches were not far from the center of the sprite. In fact, a touch 13 pixels right of center is within most of the sprites, and would be counted as a hit unless the y-coordinate was off (i.e., unless the touch was too high or too low). However, as the graph shows, there were a number of errors that were much farther than 13 pixels from the center and would be outside the sprite no matter what the y-coordinate. At every level of distance from the center, there were more errors behind the sprite than ahead of the sprite.

As mentioned in the Materials section, one of the animations (Duke, the company mascot) moved differently from the others. Unlike the other animations, Duke moved relative to his "bounding box" (an invisible box that moved along with the sprite). So although the center of the box (which is what the program recorded) was an approximation of Duke's center, for any given touch we can't say for sure where the touch was relative to Duke. Thus, these numbers may not be completely accurate for that animation. However, 6 out of the 7 animations (all except one of the musical notes) show the same tendency to be hit to the right of center, and the analysis holds whether or not Duke is included in the data.

3.3.2 Factors affecting position of errors

Just as we examined the effect of various factors on performance, we can also look at how these same factors affect the position of the subjects' missed attempts.

Some patterns can be discerned. First of all, missed attempts lagged more for the sprites that move irregularly. On average, people were 25 pixels behind the log, Duke, and the stick figure (combined), but only 1 pixel behind the four “bouncing” sprites (combined). Second, this tendency was slightly stronger for faster sprites: the subjects were 15 pixels behind the fastest sprites and nine pixels behind the slowest sprites (true for 8/9 subjects, $p < .05$ by sign test). There was also a marginally significant learning effect: people were further behind the sprites in the first block of trials than in the second block (15 vs. 10 pixels, $t = 2.21$, $p < .06$). There was no difference in the position of subjects’ missed attempts based on which strategy the subjects were using.

The fact that subjects hit slightly behind the sprite centers suggests one possible way of improving performance: instead of or in addition to counting a touch at the location of the sprite as a hit, we could count a touch slightly behind the sprite as a hit. However, the fact that the tendency to lag varies depending on the sprite suggests that we should be careful in generalizing these findings to other sprites. In particular, none of the extras in the VOD demo were used in this study, so if the objective is to try to get subjects to hit those animations, those sprites should be tested to see if people are consistently off in trying to hit those as well (and if so, how far).

4 Conclusions

Although there are large individual differences in performance, people have difficulty hitting moving targets using the touchpad and gesture model tested in this study. As expected, the larger and slower the target, the more easily subjects are able to hit it. Speed seems to make more of a difference than size, at least within the ranges used in this study.

An analysis of the subjects’ missed attempts shows that, on average, subjects hit slightly behind the sprite. This suggests that one way of making the sprites easier to hit might be to count the area just behind the sprite as a hit. Further testing is required to see if this tendency to hit behind the sprite holds for animations other than those used in this experiment.

5 References

Potter, R.L., Weldon, L.J., and Schneiderman, B. (1993) “Improving the accuracy of touchscreens: an experimental evaluation of three strategies,” In *Sparks of Innovation in Human-Computer Interaction*, (B. Schneiderman, Ed.), pages 161-169. Norwood, NJ: Ablex Publishers.

An Evaluation of the Accuracy of a Touch-sensitive Remote Control

Mark Keavney and Jeff Johnson

FirstPerson Technical Report FP-1994-4

July 1994

Abstract:

An experiment was conducted to determine how accurately subjects could use a touch-sensitive remote control to hit targets on a TV screen. Two other input devices (a larger, table-mounted touchpad, and a makeshift handheld touchpad used in previous user-testing) were used as baseline measures. The study also compared accuracy between different hit criteria (hitting targets by touching vs. lifting off the pad), and between two different methods of using the touchpad (one of which provided constant feedback regarding the position of the finger and one of which did not). The remote control fared poorly relative to both other input devices, but the results suggested several ways in which it could be improved.

An Evaluation of the Accuracy of a Touch-sensitive Remote Control

Mark Keavney and Jeff Johnson

1 Introduction

FirstPerson, Inc. designed a remote control for use in interactive television applications. People use it by touching a touch-sensitive pad on the remote control in locations that correspond to selectable areas on the TV screen. When a prototype of the remote control became available, user tests were conducted to allow the designers to evaluate how easily people could use it to hit buttons on the screen. To provide an appropriate baseline for this evaluation (Burkhart et al., 1994), the study was designed to compare subjects' accuracy with the remote control to that with two other touchpad devices: a larger, table-mounted touchpad, and a make-shift prototype handheld touchpad used in our previous studies.

We also designed the study to compare accuracy between different hit criteria (activating buttons by touching vs. lifting off the pad; see Potter, Weldon, and Shneiderman, 1993), and two different methods of using the touchpad (which differed in the amount of feedback they provide regarding the position of the finger). Finally, we designed the study to begin testing the ergonomics of certain physical parameters of the remote control, e.g., the finger pressure required to click a switch.

2 Method

2.1 Subjects

The subjects were 18 FirstPerson employees and three of their teenage offspring, totalling 21 subjects. Subjects were pre-selected to have little or no experience with either of the input devices used in the study.

Seven of the FirstPerson employees had previously participated in a similar study (Johnson and Keavney, 1994) which used a makeshift remote control (the bottom

right corner of a 5" x 7" touchpad mounted on a wooden handle). We used these subjects to compare the accuracy of that device to the accuracy of the present remote control.

2.2 Materials

Two input devices were used in this study: a FirstPerson remote control and a larger, table-mounted touchpad. The remote control is a 4" long by 2.6" wide rounded black device designed to fit easily into a user's hand. Mounted on the device is a 1.4" long by 1.9" wide touchpad, reachable by the thumb of the hand that holds the device. The remote control transmits the coordinates of any touch on the pad that exceeds a certain threshold pressure. The threshold pressure of the pad is called its *actuation pressure*. The touchpad transmits no information about the pressure of a touch other than that it is greater than the actuation pressure. The entire touchpad is mounted on a switch such that when the pad is pressed hard enough, the entire touchpad is pressed down, depressing the switch, which then signals to the computer that the remote control's "button" has been pressed¹. The pressure required to depress this switch is called the *click pressure* of the remote control. The click pressure must of course be greater than the actuation pressure, but otherwise is independent of actuation pressure.

The other input device was a large touchpad. The large touchpad is a 5" high by 7" wide touchpad mounted on a piece of cardboard and laid flat on a table. Like the touchpad mounted on the remote control, the large touchpad has a threshold of actuation called its actuation pressure. Unlike the remote control, the large touchpad does not "click" and therefore has no click pressure.

In this study, we used only one remote control and one large touchpad, so the remote control's click and actuation pressures and the large touchpad's actuation pressure were constant for all subjects and conditions. Actuation pressure is difficult to measure reliably, but the actuation pressure of the remote control's touchpad was low relative to that of other touchpads we have tested, probably between 0.5 and 1.5 ounce. The actuation pressure of the large table-mounted touchpad was higher than that of the remote control.

1. Patent applied for.



Figure 1. First Person remote control

2.3 Design

The primary purpose of the study was to evaluate the accuracy of the remote control. All 21 subjects used the remote control to hit targets on the screen. To provide a baseline against which to compare subjects' performance (see Burkhart et al., 1994), all 21 subjects also used the large touchpad. Another baseline was provided by subjects' performance in a previous study, using a makeshift handheld remote control prototype that differed from both the remote control used in the present study and the large touchpad. Only seven of the 21 subjects in the present study participated in the previous study, so direct comparisons can only be made for those seven.

Besides the comparison between devices, two other variables were examined in the study: *tracking method* (finger-tracking vs. poke-to-hit), and *hit criterion* (land-on vs. lift-off). Tracking method describes whether the subjects were told to keep their finger on the pad as they moved from target to target (finger-tracking), or to lift their finger between trials and “poke” at each new target (poke-to-hit). In the finger-tracking trials, users hit targets using the remote control by positioning the pointer over the target and pressing the touchpad until it clicked. Users hit tar-

gets using the large touchpad by briefly lifting and then retouching the pad at the position corresponding to the target.

Device (within S)	Finger-tracking method (within subject)	
	Finger-tracking	Poke-to-hit
Large Touchpad	all 21 Ss: 7 from previous experiment + 14 new)	Hit criterion (between S):
		Land-on (14 Ss: 7 from previous study + 7 new Ss) Lift-off (7 new Ss)
Remote Control	all 21 Ss	Hit criterion (between S):
		Land-on: (14 Ss) Lift-off (7 new Ss)

Table 1. Design of study

The hit-criterion variable was varied only in the poke-to-hit trials. The seven subjects using the *land-on* criterion hit buttons by touching the corresponding location on the touchpad. To count as a hit, the first point of contact had to be on the button displayed on the screen. The seven subjects using the *lift-off* criterion hit buttons by positioning their finger on the touchpad’s corresponding location and then lifting it off. To count as a hit, the last point of contact had to be on the button. Neither the land-on nor the lift-off criterion required the use of the remote control’s clicking mechanism. Each depended only on the remote control’s (or large touchpad’s) registration of a touch.

Tracking method was varied for each subject. Thus, every subject completed four blocks of trials, one for each tracking method with each of the two devices. Hit criterion was varied between subjects, but only in poke-to-hit trials. For those trials, each subject used either the land-on hit criterion or the lift-off hit criterion. Hit criterion for the finger-tracking trials was not varied; it was always click-on for the remote control and land-on for the large touchpad.

The seven subjects who had participated in a previous study using a makeshift remote hand-held touchpad were not randomly assigned to a hit criterion. Because land-on was the hit criterion used in the previous study, all of these subjects were placed in the land-on condition in the present study in order to provide a matched comparison. Because these subjects were not randomly assigned to a hit criterion, they are not included in comparisons between the different hit criteria. The remaining 14 subjects were randomly assigned, seven to each hit criterion category.

2.4 Procedure

The subjects sat about eight feet from the TV, directly in front of a low table on which the two input devices lay. Subjects used the large touchpad without moving it from its position on the table. They used the remote control by picking it up and holding it with whichever hand felt more comfortable (for most subjects, this was the right hand). The experimenter instructed subjects to use their thumb to touch the remote control, but didn't tell them to use any particular part of the thumb. Some subjects used the pad of the thumb while others used their thumbnail. To touch the large touchpad, the experimenter told the subjects to use whatever finger was easiest. Most subjects used their right index finger.

Subjects were asked to hit lettered "buttons" on a 19" TV screen. The buttons were arranged in a 5 x 5 array and were labeled "A" through "Y." The software signalled that a letter was "hit" by a "click" sound; a different sound indicated a miss (i.e., a hit on the background). The software displayed a cursor at the location of each touch, which gave the subjects feedback about the location of their finger relative to the buttons.

In the poke-to-hit trials, the experimenter announced which letter the subject should hit and recorded the number of attempts the subject made in trying to hit that letter. In the finger-tracking trials, the computer indicated the next letter automatically by drawing a pink highlight around it and automatically recorded both the number of attempts and the time until the subject hit the letter. Thus, for the poke-to-hit trials, only the number of attempts was measured; for the finger-tracking trials, both the number of attempts and the time for each attempt were measured. In both sets of trials, the letters were presented in an order that had been randomly determined but was the same for each subject.

Each trial lasted until the subject successfully hit the target letter or made 20 unsuccessful attempts. Each subject completed four blocks trials, one for each combination of device (remote control vs. large touchpad) and tracking method (poke-to-hit vs. finger-tracking). Each block was to have consisted of 50 trials, but for some of the subjects, software problems caused the program controlling the finger-tracking trials to malfunction after 24 trials.

At the end of the experiment, the experimenter asked the subject to rank the four methods they had just used to hit targets. When necessary, the experimenter clarified the question by asking which method the subjects would prefer to use if they were doing this task over a long period of time. The experimenter also asked the subjects whether they found the remote control's click pressure too light, too heavy, or about right.

3 Results

3.1 Remote control vs. other input devices

Two comparisons can be used to evaluate the accuracy of the remote control. The first is the comparison between the remote control and the large touchpad. Every subject completed two blocks of trials with the remote control and two blocks of trials with the large touchpad (in each case, one block was with finger-tracking and the other was with poke-to-hit). These comparisons are illustrated in Figure 2 and Table 2 below.

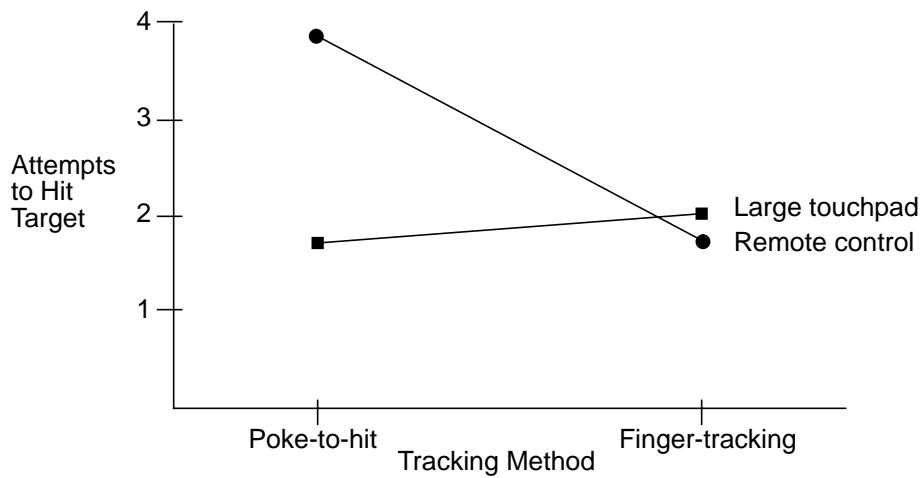


Figure 2. Performance by tracking method and input device

Device	Poke-to-hit attempts	Finger-tracking attempts	Finger-tracking time (secs)
Remote control	3.9	1.5	3.68
Large touchpad	1.7	2.0	2.37

Table 2. Performance by tracking method and input device

As Figure 2 and Table 2 show, the remote control fared poorly relative to the large touchpad on the poke-to-hit trials. Subjects made significantly more attempts when using the remote control than when using the large touchpad ($t = 5.13$; $p < .01$). The remote control also fared poorly on the time measure for the finger-tracking trials. Subjects took significantly longer to hit targets with the remote control than with the large touchpad ($t = 3.18$; $p < .01$). The one measure on which the remote control seemed to outperform the large touchpad was the number of attempts for the finger-tracking trials. Subjects made significantly fewer attempts with the remote control than with the large touchpad when using this tracking method ($t = 3.01$; $p < .01$). However, even here the remote control may not have actually outperformed the large touchpad, since many of the events recorded as errors for the large touchpad were not attempts to hit a target, but rather cases where the subjects had trouble keeping their fingers on the pad as they moved to the next target. Because the hit criterion for the large touchpad was land-on, the program interpreted accidental releases followed by a touch as an unsuccessful attempt to hit the target. Overall, subject performance with the remote control was poor relative to that with the large touchpad.

A similar pattern of results occurs in the analysis of subjects' preferences. At the end of the experiment, subjects were asked to rank the four combinations of tracking method and device that they had just used. The average preferences are shown in Table 3. The methods differed reliably in how much they were preferred: by a Friedman test, these four groups of scores differed significantly from a chance distribution ($\chi^2 = 16.83$, $df = 3$; $p < .01$).

Device	Finger-tracking	Poke-to-hit	Overall
Remote control	2.4	3.5	2.9
Large touchpad	2.1	2.0	2.1
Overall	2.2	2.8	2.5

Table 3. Mean preference ranking by tracking method and input device (1 = most; 4 = least)

Specific hypotheses about the preferences were tested through the use of linear contrasts. The effect of input device was significant: subjects preferred the large touchpad to the remote control ($t = 3.87$; $p < .01$). The effect of tracking method was not significant: although there was a tendency for subjects to prefer finger-tracking to poke-to-hit, this difference was not significant ($t = 1.63$; N.S.). The contrast for the

interaction between tracking method and input device was significant: subjects especially disliked the remote control when using poke-to-hit ($t = 3.68$; $p < .01$).

It is clear that the remote control did not perform well relative to the baseline of the large touchpad. The second baseline that can be used in evaluating the remote control's performance is the performance of subjects using a makeshift hand-held touchpad in a previous study. Only the seven subjects who participated in the present experiment as well as in the previous study are included in this comparison. In the present study, using the land-on hit criterion and the poke-to-hit tracking method (the same hit criterion and tracking method used in the previous study), those seven subjects averaged 4.2 attempts. In the previous study with the other device, the same seven subjects averaged 2.8 attempts. The difference is not significant ($t = 1.08$; N.S.). It was the experimenters' informal impression that although the remote control was much easier to use than the makeshift device, the makeshift device had a higher actuation pressure than the remote control, and this may have led to fewer errors (see the Discussion section for an explanation of how this might happen).

In sum, subjects performed significantly worse with the remote control than with the large table-mounted touchpad, and not significantly differently than with the makeshift handheld touchpad. Thus, the remote control produced performance that was at best no better than either of the devices used as baselines.

In addition to evaluating the performance of the remote control relative to other input devices, this study was designed to investigate the remote control's performance under different hit criteria and tracking methods. Every subject used both tracking methods and one of the two hit criteria.

3.2 Tracking method

For several reasons, it is problematic to compare performance between the two tracking methods. First, the programs and procedures used with finger-tracking were slightly different than those used with poke-to-hit. Second and more importantly, the program used for finger-tracking malfunctioned for several subjects after 24 trials, so the finger-tracking data are all based on only the first 24 trials, while the poke-to-hit data are based on all 50 trials. This makes statistical comparisons between the two methods inappropriate.

However, we can compare the means of the different tracking methods, to give us a tentative suggestion of how tracking method might affect performance. Subjects made slightly fewer errors with finger-tracking than with poke-to-hit (1.8 vs. 2.8), and there was a tendency to prefer finger-tracking to poke-to-hit (2.2 vs. 2.8).

3.3 Hit criterion

To compare performance between the two different hit criteria, we excluded from the analysis the subjects who had been in the previous study, since they were not randomly assigned to conditions. Among the remaining 14 subjects, those who used the lift-off criterion performed better than those who used land-on. Averaging across the two input devices, the lift-off subjects had significantly fewer attempts than the land-on subjects (1.8 vs. 3.6, $t = 3.64$, $df = 8$; $p < .01$). However, we didn't measure time for these trials. It is possible that the lift-off subjects took longer than the land-on subjects even though they made fewer errors. This would be consistent with Potter, Weldon, and Shneiderman (1993), who found in a study of touchscreens (as opposed to remote touchpads) that subjects using lift-off took longer but made fewer errors than subjects using land-on.

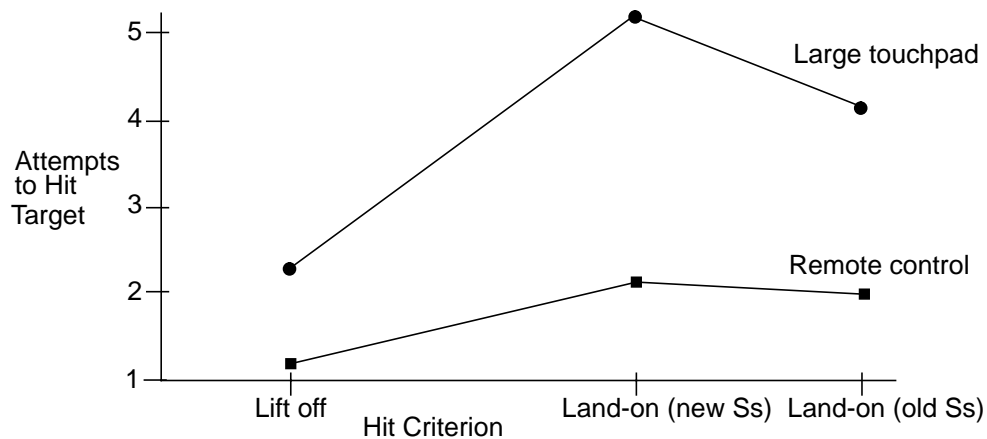


Figure 3. Attempts to hit target by hit criterion and input device

Device	Lift-off	Land-on New Ss	Land-on Old Ss	Overall
Remote control	2.3	5.2	4.2	3.9
Large touchpad	1.2	2.1	2.0	1.7
Overall	1.8	3.6	3.1	2.8

Table 4. Attempts to hit target by hit criterion and input device

In addition to affecting overall performance, the hit criterion also affected the difference in performance between the input devices. Although all subjects made fewer attempts with the large touchpad than with the remote control, this difference was much more pronounced among the subjects who used land-on than among those who used lift-off ($t(\text{diff}) = 2.71, df = 9; p < .05$). Thus, the subjects performed better using lift-off than land-on and especially poorly using land-on with the remote control. These results are shown in Figure 3 and Table 4.

3.4 Learning

The order of the trial blocks was counterbalanced. There was no evidence of learning across trial blocks. Within the poke-to-hit trials, for example, subjects performed slightly better on the second block of trials than on the first (2.5 vs. 3.1 attempts), but this difference was not significant ($t = 1.02; \text{N.S.}$). Within the finger-tracking trials, on average, subjects made slightly fewer attempts but took more time on their first block of trials than on their second (1.7 vs. 1.8 attempts, 3.34 seconds vs. 2.76 seconds). Again, however, neither effect was significant ($t = 0.34, \text{N.S.}$ for attempts; $t = 1.13, \text{N.S.}$ for time).

The hit-criterion variable (land-on vs. lift-off) did not affect the amount of learning from the first block to the second. There was a slight tendency for subjects to improve more with the land-on hit criterion (3.9 attempts on the first block vs. 2.8 on the second block) than with the lift-off model (1.6 attempts on the first block vs. 1.9 on the second block) but this difference was not significant ($t = 1.46; \text{N.S.}$). This is consistent with previous research, which has found weak and non-significant learning effects with similar tasks (Johnson and Keavney, 1994).

3.5 Click pressure

At the end of the experiment, subjects were asked whether they thought that the click pressure of the remote control was too heavy, too light, or about right. Fourteen out of the 21 subjects found it too heavy, six said that it was about right, and only one subject said that it was too light. By a sign test, this distribution is significantly skewed ($p < .01$) from what one would expect if subjects showed no bias and were therefore responding randomly (which would yield about one-third in each response category), so it is safe to say that the click pressure on the prototype remote control used in the study was too heavy.

4 Discussion

The results indicate that the remote control still has a lot of problems. People often made errors with it, they took longer to use it than to use the large touchpad, they

preferred the large touchpad to it, and they complained that its click pressure was too heavy. Clearly, the remote control is not yet as usable as it needs to be. The performance problems with the remote control may even have been underestimated in this study, since many of the subjects used their thumbnails to hit the targets, whereas for reasons of comfort, the remote is designed to be used with the pad of the thumb.

However, some aspects of the data suggest that the remote control does have promise. Most people were impressed with its physical design: all of the comments regarding how it was shaped and how it fit into the hand were positive. This is in contrast to previous studies with the makeshift handheld touchpad in which there were a fair number of complaints about the unwieldiness of that device. It also appears that finger-tracking improves both performance with and preference for the remote control considerably.

The main reason for the poor performance of the remote control was that the cursor wasn't stable. The area of contact between a user's thumb and the touchpad is a substantial portion of the total area of the touchpad. The coordinates transmitted (30 times a second) by the pad are each the centroid of the area of contact. Small changes in how the thumb contacts the pad can cause relatively large shifts in the position of the centroid and, hence, in the cursor position. In this study, the cursor tended to move whenever subjects changed the pressure of their finger on the touchpad, e.g., as they tried to click the touchpad or lift their finger off it. Also, involuntary small movements caused shifts in the centroid position, making the cursor move even when subjects were trying to keep their thumb stationary.

This problem is not unsolvable. There are at least three ways that the cursor could be made more stable. First, the software in the computer could average over time the points it receives from the remote control rather than repositioning the cursor at each individual point received. The firmware within the remote control already averages points to a certain extent. It samples 100 points per second and then (by discarding outliers and averaging) uses those to construct 30 points per second to transmit. Modifying the computer software to average the points received from the remote control in some way might make the cursor less jittery.

Second, the cursor could be stabilized by decreasing click pressure. It is clear from subject responses (see Click Pressure, above) that the click pressure of the remote control used in this study is uncomfortably heavy, and that alone is reason to change it; but in addition, decreasing click pressure could also make the cursor more stable during clicks. The reason that the cursor sometimes moves during a click is that the position of the thumb changes as the user applies pressure to the pad. The lighter the pressure required to click the pad, the less the position of the thumb should change, and the less the cursor should move.

For similar reasons, increasing the remote control's actuation pressure could stabilize the cursor during touches and releases. Just as a user's thumb position changes during a click, it also changes during a touch or a release. If the touchpad first registers a touch early (when the thumb is just barely touching the pad), the cursor will move more than if the pad registers the touch later, when the position of the thumb is close to its final position. Informal testing done after the completion of this study suggested that using a touchpad with a higher actuation pressure than that used in this study, e.g., closer to 2.5 oz., would reduce jitter and improve accuracy.

Of course, there is a risk in simultaneously increasing the remote control's actuation pressure and decreasing its click pressure: users may confuse them, and click the remote control when they had intended only to touch it. However, accidental clicks were not a problem in this latest study, so it seems that there is at least some room for the two pressures to converge before we need to worry about confusions between them.

5 Conclusions

The remote control has serious usability problems, but these problems can be reduced or eliminated by: 1) modifying the software in the computer to average base the cursor position on the average of several coordinate-points received from the remote control; 2) decreasing the remote control's click pressure; and 3) increasing its actuation pressure. Optimal values for the click pressure, the actuation pressure, and the number of points to average over in stabilizing the cursor should be determined through further testing.

6 References

Burkhart, B., Hemphill, D., and Jones, S. (1994) "The Value of a Baseline in Determining Design Success," *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'94)*, pages 386-391.

Johnson, J. and Keavney, M. (1994) "The Effect of Touch-Pad Size on Pointing Accuracy," *FirstPerson Technical Report FP-1994-2*. (In this collection)

Potter, R.L., Weldon, L.J., and Shneiderman, B. (1993) "Improving the accuracy of touchscreens: an experimental evaluation of three strategies," in B. Shneiderman (Ed.), *Sparks of Innovation in Human-Computer Interaction*, Norwood, NJ: Ablex Publishers.

A Comparison of Remote Pointing Devices for Interactive TV Applications

Jeff Johnson and Mark Keavney

FirstPerson Technical Report FP-1994-5

September 1994

Abstract:

A series of experiments was conducted to determine which of several candidate pointing devices is most usable for controlling interactive TV applications. Devices compared included a mouse, two types of trackballs, an isometric joystick, and a handheld remote control based on an absolute touchpad. The subjects' task was to hit targets on a TV display from a normal TV-viewing distance using one of the pointing devices. Experiment 1 compared the devices using device-control parameters that seemed reasonable to the authors, and also began exploring simple stabilization methods for the touchpad remote control. Experiments 2a and 2b were conducted to optimize device control-parameters to improve the validity of a device-comparison. Experiment 3 again compared devices, this time with optimized control-parameters. It also attempted to clarify learning effects and began exploring more complex stabilization methods for the touchpad remote control. In both Experiments 1 and 3, subjects were fastest and most accurate with the mouse. However, a mouse is unsuitable for interactive TV applications because it requires a surface upon which to operate. The order for the remaining devices, in both speed and accuracy, was: mini-trackball, trackball, isometric joystick, and touchpad remote control. Subject preferences showed a similar pattern, except that the touchpad remote control was preferred to the isometric joystick. Practice with the touchpad remote control quickly improved subject performance, approaching performance with the trackballs. Static stabilization methods (i.e., simple averaging of points over time) improved error performance with the touchpad remote control, but their effect on speed, if any, was very weak relative to the effect of practice. Preliminary results suggest that velocity-dependent stabilization may improve performance more than fixed-average stabilization.

A Comparison of Remote Pointing Devices for Interactive TV Applications

Jeff Johnson and Mark Keavney

1 Introduction

1.1 Background and motivation

As part of a prototype interactive TV system, FirstPerson developed a new type of remote control. The control was designed to be less complicated, easier to use, and less complicated than most present-day remote controls for electronic devices. Among the design goals for the FirstPerson remote control were:

- as few buttons as possible; perhaps only one,
- extremely cheap to manufacture, i.e., very simple electronics and casing,
- useful for controlling complex devices and for navigating large information spaces,
- highly ergonomic design, i.e., fits well into the user's hand,
- ruggedness: ability to survive rough treatment.

The remote control is 4" long by 2.6" wide (see Figure 1). Mounted in it is a 1.4" long by 1.9" wide touchpad, reachable by the thumb of the hand that holds the device. The touchpad is the primary control component of the control. It is an absolute positioning touchpad: each point on the pad corresponds to a point on the TV display. Users touch the pad with their thumb while holding the remote-control in their hand; the system responds by displaying a pointer at the corresponding screen location. Pressing directly on the touchpad depresses a button



Figure 1. First Person remote control

under the pad, providing a user-clickable button¹. With the interactive TV prototypes FirstPerson developed, users use the remote control, specifically the touchpad, to navigate in information spaces, make selections, and alter application settings.

An important issue for the designers was the accuracy with which users could, via various gestures on the touchpad, operate the prototype interactive TV system. What should the user interface for moving around in information spaces be? Can users reliably hit targets? Of what size? What if the targets move? How does the accuracy of the touchpad compare to other pointing devices that might have been used? Before the remote control was developed, some of these issues could be—and were—addressed through studies using touch-displays, table-mounted touchpads, and primitive, makeshift hand-held touchpads (Johnson, 1994; Johnson and Keavney, 1994; Keavney, 1994; Keavney and Johnson, 1994). However, studies addressing the last of the above issues—comparing the remote control with other pointing devices—had to wait until the first working models of the remote control were available.

1. Patent applied for.

One reason for comparing the FirstPerson remote control with other pointing devices was to provide a benchmark against which it could be measured; otherwise, there would have been no objective way to decide whether its accuracy was good or poor. For various reasons, the designers had concerns about the accuracy of the remote control's touchpad, and so felt the need to test it extensively. Another reason was that in case the FirstPerson remote control turned out to be unacceptable for one reason or another (e.g., accuracy, cost, fragility), the designers wanted to know what sort of pointing device to use instead.

1.2 Related work

The pointing device comparison experiment reported herein is certainly not the first such study to be performed. The results of several previous pointing device comparison studies provide a source of methodological ideas and expected results.

The seminal experimental comparison of pointing devices was conducted by Card and his colleagues (Card, English, and Burr, 1978). It was conducted when the mouse was just beginning to spread beyond the research labs where it was invented (English, Engelbart, and Berman, 1967). Card et al. compared users performing pointing and selection tasks with a mouse, an isometric joystick, arrow keys (up, down, left, right), and text keys (next character, next word, next sentence, next paragraph). They found that subjects were fastest and most accurate with the mouse. In addition to comparing devices, the study examined the effects of target size and distance on time and accuracy. The authors found that the mouse and joystick data were well accounted for by Fitts' law (Fitts, 1954), and that the arrow and text key data were not. They also found that the Fitts law parameters for the mouse suggested that mouse performance was close to optimal.

Subsequent experimental comparisons included pointing devices other than those Card et al. tested and tasks other than text selection. Epps and Snyder (1986) compared task-completion time in a target-pointing task for several devices: mouse, absolute touchpad, relative touchpad, trackball, displacement joystick, and isometric joystick. This comparison was performed using empirically-optimized gains² for the devices. Like Card et al., Epps and Snyder also systematically varied target size and distance. They found that subjects were faster at hitting targets with the mouse and the trackball than with the other devices, and this effect was greater the smaller or more distant the target. However, they found no differences between the mouse and the trackball or between the other group of devices. Epps

2. Gain is a parameter that determines how device movements are multiplied to obtain screen-pointer movements, i.e., $\text{PointerDisplacement} = \text{Gain} \times \text{DeviceDisplacement}$.

(1987) essentially replicated these findings, using graphics editing tasks instead of a selection task. He also asked subjects to indicate their relative preferences for the devices, and found that the mouse and trackball were the favorites.

Because previous studies had found the mouse and trackball to be close contenders for “best pointing device,” but had not indicated which might be better, Sperling and Tullis (1987) conducted experiments to compare the two. Their intent was to end, once and for all, frequent quasi-religious arguments between mouse and trackball advocates. First, they surveyed a population of Apple Macintosh users whose computers included both a mouse and a trackball to determine which device, if any, was more popular. Out of 39 respondents, 56% used the mouse exclusively, 15% used the trackball exclusively, and 28% use both depending on the task. Next, using subjects drawn from the surveyed group, Sperling and Tullis compared time and accuracy on three tasks: target selection, menu selection, and path tracing. They found no differences in *accuracy* (because the error rates were so low). However, they also found that subjects—even those who normally use the trackball—were *faster* with the mouse than with the trackball.

To resolve another “religious” issue, Jellinek and Card (1990) conducted a study to examine carefully the effect of gain—both fixed and variable—on user-performance with a mouse. Many people prefer so-called “power-mouse” (mouse acceleration) software to standard mouse software, claiming that a “faster” mouse lets them work faster. Noting that it has been well-established that mouse performance follows Fitts’ law (Fitts, 1954), in which gain plays no role, Jellinek and Card conducted three experiments using a standard Fitts’ Law task. The first experiment compared several fixed gains, and found that higher-than-usual gains did not improve performance, but rather degraded it, presumably by decreasing the effective screen-resolution. Lower-than-usual gains also degraded performance, mainly by requiring users to make multiple movements to reach targets. The second experiment examined “power-mouse” schemes, using a standard gain for slow mouse movements and a higher gain for faster movements, varying the slow/fast threshold and the higher gain. Neither manipulation affected performance. The third experiment compared a standard mouse (fixed gain of 2) with two popular power-mouse schemes, and found no significant differences. The authors concluded that gain and gain-acceleration *per se* do not affect performance with a mouse, as predicted by Fitts’ Law. However, they noted that performance may suffer if increased gain decreases screen-resolution, or if decreased gain makes the mouse-footprint too large.

As part of a program of research on touchscreen user interfaces, Sears and Shneiderman, (1993) conducted a study that compared performance (time and errors) on a selection task using a mouse and a touchscreen. Each subject used a mouse and

two versions of the touchscreen: one that used software techniques to stabilize the cursor, and one that didn't. In addition to comparing devices, Sears and Shneiderman varied the target-rectangle size from 1 pixel on a side to 32 pixels. They found that subjects were faster with the stabilized touchscreen than with the mouse for all but the smallest (1-pixel) targets. Also, though subjects made fewer selection errors with the mouse than with the stabilized touchscreen for the smaller targets, the difference disappeared for larger target sizes.

Based upon the results of the previous studies, it seemed reasonable to assume that the mouse and trackballs would emerge as the best of the alternative pointing devices. Furthermore, since the FirstPerson remote control is based upon an absolute touchpad that did not (at least initially) employ software stabilization, which in previous studies proved inferior to mice and trackballs in usability, it seemed unlikely that the FirstPerson remote control would emerge first in the current comparison. However, replication of the previous findings was not guaranteed because of differences between the conventional computer-use situation and the interactive TV situation, e.g., in the interactive TV situation, the display is much further from the user and the table on which the mouse rests (a coffee table) is much lower. Therefore, a series of experiments was conducted.

Before the experiments reported herein were conducted, two initial experiments were carried out. The first examined the effect of the size of the touchpad—and hence the ratio of pad-size to the TV display size—on pointing accuracy (Johnson and Keavney, 1994). The results indicated no clear relationship between pad-size and accuracy within the tested range of sizes. A second initial experiment compared the accuracy of the FirstPerson remote control against a larger, table-mounted touchpad and a makeshift handheld touchpad (Keavney and Johnson, 1994). The non-stabilized remote control fared poorly relative to the other devices, but the results suggested several ways to improve the usability of the remote control.

Having compared the FirstPerson remote control with other touchpad-based pointing devices, three experiments comparing it with non-touchpad-based pointing devices were conducted. The first experiment was designed and conducted quickly to provide input to a time-critical business decision; the others were conducted later to improve the methodology and thereby clarify the findings.

2 Experiment 1: Initial Comparison

2.1 Method

2.1.1 Subjects

Five subjects who were not FirstPerson (FP) employees participated in the experiment. Each was paid \$15 for the one-hour session. All subjects had previous experience with a mouse, but none had used the FP remote control before.

2.1.2 Materials

Five pointing devices were used in the experiment: a mouse, a mini-trackball, a large trackball, a device called a “pointing stick,” and an FP remote control. The mouse, trackballs, and pointing stick are commercially-available devices used for controlling an on-screen pointer. Each of these devices has a method of moving the screen-pointer and a separate button to select an object. The pointing stick has a thumb-pad that is pressed on its edges to control the screen-pointer’s direction of motion and velocity. It is similar in operation to an isometric joystick.

Prior to the experiment, the operating gain for these devices was set to values that seemed reasonable to the experimenters. The gains used were: 6.0 for the pointing stick and trackball, 8.0 for the larger trackball, and 4.0 for the mouse.

The FP remote control is a new pointing device developed by FirstPerson. It is 4” long by 2.6” wide, shaped so as to fit easily into a user’s hand. A 1.4” long by 1.9” wide touchpad is mounted on it and is reachable by the thumb of the hand that holds the device. The remote control transmits to the computer the coordinates of any touch on the pad that exceeds a certain threshold pressure. The threshold pressure of the pad is called its actuation pressure.

In contrast to the other devices tested, the FP remote control is an absolute positioning device. The screen-pointer’s position is determined by where a user is touching the pad, rather than by how the touch-point moves.

The touchpad transmits no information about the amount of pressure applied to it beyond the fact that its actuation pressure has been exceeded. However, the entire touchpad is mounted on a switch such that when the pad is pressed hard enough, the entire touchpad is pressed down, depressing the switch, which then signals to the computer that the remote control’s “button” has been pressed. The pressure required to depress this switch is called the click pressure of the remote control.

The click pressure of a remote control must be greater than its actuation pressure, but is otherwise independent of actuation pressure.

Both actuation pressure and click pressure vary among the different prototype FP remote controls. However, only one FP remote control was used in this experiment, so the click pressure and actuation pressure were the same for all subjects and conditions.

The remote control transmits the location of its touch-point 30 times per second. Normally, the software in the computer simply redisplay the screen-pointer at each point received. Because of the way users' thumbs contact the touchpad and the small size of the pad relative to the TV display, the screen-pointer tends to move around more than the user intends it to. When the device comparison experiment was being designed, we were investigating various methods of modifying the software to stabilize the screen-pointer. One such method was averaging the last two points received by the computer rather than simply positioning the screen-pointer at every received point. We used both the original software and this "average 2" stabilizing method in Experiment 1, as described below.

2.1.3 Design

In the first phase of the experiment, each subject completed five blocks of 50 trials, one block with each of the five pointing devices. To control for learning effects, the order of the devices was rotated from subject to subject (i.e., the device that was used first by the first subject was used second by the second subject, third by the third subject, and so on). In the first phase, the subjects used the non-stabilized remote control.

In the second phase of the experiment, each subject completed one more block of 50 trials, using the remote control with software that stabilizes the screen-pointer by averaging points over time.

2.1.4 Procedure

The subjects sat approximately eight feet from the TV, directly in front of a low table. The large trackball and the mouse required a flat surface, so they were used on the table. None of the other devices (the mini-trackball, the pointing stick, and the FP remote control) required a surface: the subjects held these devices in their hand as they used them.

The task was one used in a previous study of the FP remote control's accuracy (Keavney and Johnson, 1994): subjects saw a 5 x 5 array of lettered buttons on the screen, one of which was highlighted, and tried to "press" the highlighted button

as quickly as possible. When a button was hit, it unhighlighted and the next button in the randomized sequence highlighted. In addition, the auditory feedback for a button-hit differed from that for a button-miss.

With the off-the-shelf pointing devices, the pointer remained visible at all times; subjects simply moved it around and clicked the device’s physical button to hit the target button on the screen. With the FP remote control, the pointer was visible only when the subject was touching the pad, but subjects were instructed to keep their thumb on the pad so that the screen-pointer would remain visible. With the FP remote control, subjects clicked the pad to hit the target screen-button.

The software recorded the time required to hit each button (starting from when it was highlighted) and the number of missed attempts.

After trying all five pointing devices, subjects were asked to rank them in order of their preference.

After ranking the devices, subjects used the remote control with the “average 2” method of stabilizing the screen-pointer. For one subject, there was a technical problem with the remote control on this block of trials, so data were only collected for four of the five subjects.

2.2 Results

There was a striking difference between the pointing devices in how fast users hit the targets. The overall difference between the five devices (excluding the stabilized remote control since it was always presented last) is statistically significant ($p < .01$ by Friedman’s test; see Hayes, 1963). Table 1 lists the means by subject and device.

Subject	Mouse	Mini-trackball	Large trackball	Pointing stick	FP remote	FP rem. w/ avg. 2
S1	.92	1.42	1.49	2.98	3.21	no data
S2	.90	1.60	1.43	3.17	4.59	2.36
S3	1.38	2.15	2.13	4.47	5.25	2.73
S4	1.15	1.80	1.70	2.29	2.20	1.71
S5	3.44	2.06	2.63	3.97	5.68	4.66
Overall	1.56	1.81	1.88	3.38	4.18	2.87

Table 1. Average time to hit target (in seconds), by subject and input device

The differences between the devices (again, excluding the stabilized FP remote control) in how many errors (i.e., missed hit-attempts) users made were less striking, but significant nonetheless ($p < .01$ by Friedman’s test). See Table 2 for details.

Subject	Mouse	Mini-trackball	Large trackball	Pointing stick	FP remote	FP rem. w/ avg. 2
S1	.00	.04	.04	.06	.48	no data
S2	.02	.08	.06	.32	.58	.24
S3	.00	.00	.00	.06	1.00	.40
S4	.00	.00	.00	.00	.12	.00
S5	.04	.00	.06	.08	.40	.40
Overall	.01	.02	.03	.10	.52	.26

Table 2. Average number of errors per trial, by subject and input device

As the tables show, in both time and errors, the devices came out in the same order: mouse, mini-trackball, large trackball, pointing stick, with the non-stabilized FP remote control a distant last. However, when the FP remote control was stabilized by averaging over the last two points, its performance improved 35% on time and 50% on errors. The improvement in time is statistically significant (difference = 1.56 seconds, $t = 3.23$; $p < .05$) though the effect in errors is not (difference = .26, $t = 2.01$; N.S.). Thus, “average 2” stabilization—a rather minimal method of stabilization—improves its overall ranking somewhat: it is still last in terms of errors, but comes in ahead of the pointing stick on the time measure.

Subjects’ preference rankings were similar to their performance scores, though the FP remote control did score ahead of the pointing stick (see Table 3). However, the differences in these rankings are not significantly different from each other by Friedman’s test.

Mouse	Mini-trackball	Large trackball	Pointing stick	FP remote	FP rem. w/ avg. 2
1.5	2.9	3.1	4.0	3.5	no data

Table 3. Mean rankings (1 = most preferred, 5 = least preferred)

2.3 Discussion

Experiment 1 was conducted to compare the FP remote control to other, commercially-available pointing devices. Overall, the FP remote control fared poorly relative to these devices. It was a distant last in terms of time and errors, and second to last in subject preference.

However, the results of the experiment were not completely negative for the FP remote control. Subjects generally admired the look and feel of the remote control and liked the idea of having the selection-button be the touchpad itself. Also, subjects' performance with the remote control improved considerably in the second phase of the experiment, when the screen-pointer was stabilized slightly using the average-2 method. Of course, since subjects always used the stabilized remote control after the non-stabilized remote control, much of this improvement may be due to learning rather than to stabilization. To remove this confounding required further experiments. Informal pilot testing conducted after Experiment 1 suggested that average-2 stabilization does improve performance, and that averaging over more than two points improves performance even more.

We wanted to compare the alternate pointing devices with a more stabilized FP remote control. Before doing that, we wanted to find a better stabilization method than the “average 2” method. Also, for a fair comparison, we had to ensure that the alternative pointing devices were tested in their best configurations. This meant optimizing their gains empirically, as Epps and Snyder (1986) and Epps (1987) did, rather than simply setting the gains to reasonable values as we did in Experiment 1. Therefore, before conducting the improved comparison experiment (Experiment 3), we conducted two minor studies (Experiments 2a and b) to optimize the devices.

3 Experiment 2a: Optimization of Alternate Pointing Devices

Gain is a multiplier used in translating device movement into screen pointer movement. The larger the gain, the more the pointer moves for a given device movement. When the gain is too small, large movements (e.g., across the screen) are difficult; when the gain is too large, fine movements are difficult.

As described in the Introduction, Fitts' Law predicts—and Jellinek and Card (1990) showed empirically—that users' time to hit targets using a mouse is independent of gain. However, time-to-hit targets *is* degraded by low screen-resolution and by the need to execute multiple movements to reach the target, which can

result from gains that are too high or too low, respectively. Therefore, indirectly, the gain of a device *can* affect subjects' speed and accuracy.

Because the results of Experiment 1 were needed quickly for business reasons, we didn't spend time optimizing the gains for each device; we just set the gains to values that seemed appropriate. Experiment 2a was run to determine the optimal gains for the mouse and the trackballs in order to allow us to conduct a more meaningful comparison experiment. The pointing stick was not optimized in Experiment 2a because its comparison to the remote control was no longer of interest.

3.1 Method

To optimize the gains for the mouse, the mini-trackball, and the regular trackball, six employees of FirstPerson or Sun Interactive each participated on three successive days. Each of the three days was devoted to one of the three devices. On each day, each subject performed six blocks of 50 trials of the same task used in Experiment 1, using that day's device. Each block used a different gain. The order in which the gains were presented was permuted across subjects, such that each subject had a different order.

3.2 Results

The results are shown in Figure 2 and Tables 4-6 below. The time to hit targets and the error rates are listed in the tables, and the time to hit is graphed in Figure 1.

Friedman's test revealed no significant differences between the different gains on speed or accuracy measures for any of the three devices. However, while we would feel more secure in interpreting the observed trends if the effect of gain had been significant, our purpose was to optimize a parameter, not to look for significant effects. As the graphs of time-to-hit show, for all of the devices the subjects performed best in the middle of the range and worst at the two extremes, exactly what we would expect if the optimal gain was within the range tested. Therefore, given these data and the fact that many of the subjects said that they preferred values in the mid-range, we felt justified in believing that we had found the optimal gains for these devices. Therefore, for Experiment 3, a gain of 8 was used for the mouse, a gain of 12 was used for the mini-trackball, and a gain of 10 was used for the large trackball.

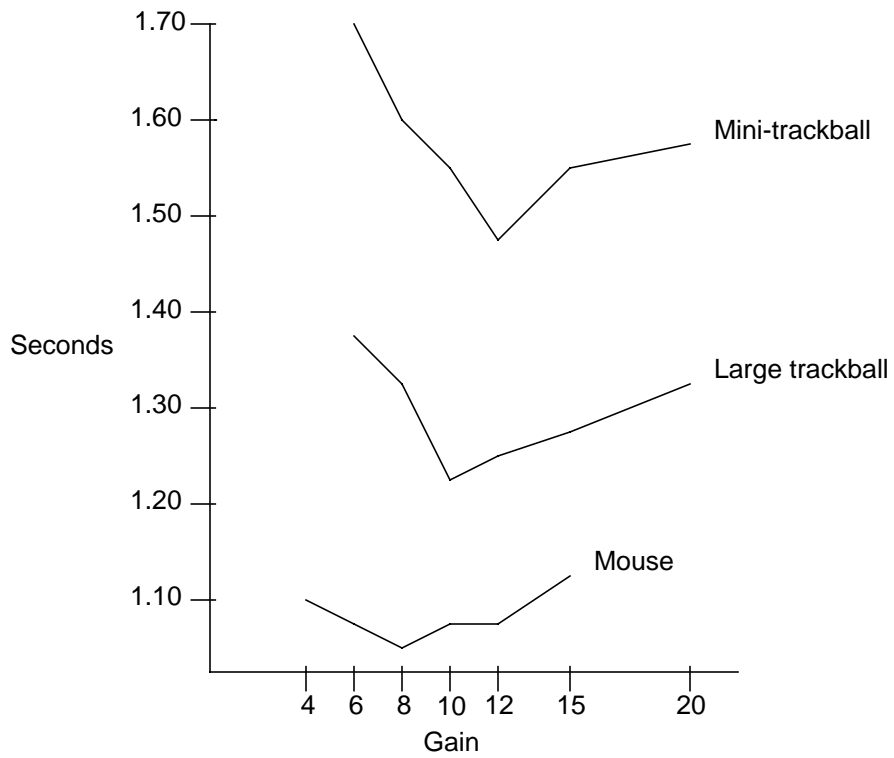


Figure 2. Optimization of mouse, mini-trackball, and large trackball: average time to hit target

Gain	Time to hit target (secs)	Errors
4	1.10	.01
6	1.08	.00
8	1.05	.03
10	1.06	.01
12	1.08	.04
15	1.11	.03

Table 4. Optimization of mouse: average time to hit target and number of errors

Gain	Time to hit target (secs)	Errors
6	1.71	.03
8	1.61	.07
10	1.55	.09
12	1.47	.06
15	1.54	.08
20	1.58	.10

Table 5. Optimization of mini-trackball: average time to hit target and number of errors

Gain	Time to hit target (secs)	Errors
6	1.38	.09
8	1.31	.12
10	1.23	.07
12	1.24	.08
15	1.28	.07
20	1.33	.09

Table 6. Optimization of large trackball: average time to hit target and number of errors

4 Experiment 2b: Optimization of FP Remote Control

In Experiment 2a, we found the optimal gains for the mouse and two trackballs, in order to better compare them with the FP remote control. The FP remote control does not have a gain parameter, since it is an absolute pointing device. However, the remote control does have a parameter that affects its speed and accuracy: the amount of averaging in the stabilization software. When the screen-pointer simply reflects every touch-point transmitted by the remote control each 1/30 second, it tends to jitter. This jitter can be reduced by having the screen-pointer's position reflect the running average of the last several points received from the remote control, rather than just the last point. Stated as a formula in which n is the number of

touch-coordinates that have been transmitted since the device was turned on, C_j is the j th touch-coordinate transmitted, P_n is the current (i.e., n th) position of the screen pointer P , and a is the number of touch-points to be averaged:

$$P_n = \frac{\sum_{j=n-(a-1)}^n C_j}{n}$$

In Experiment 1, we compared displaying the last point with displaying an average of the last two points (i.e., $a = 2$), and observed that averaging over two points seemed to be an improvement. However, that comparison was unfair in that the stabilized remote control was always tested after the non-stabilized one, making practice as likely an explanation for the observed improvement as stabilization. However, subsequent pilot testing suggested that the stabilization did help, and that averaging over more points would be even better. On the other hand, the pilot testing also showed that if the screen-pointer position is averaged over too many points, it becomes sluggish and lags behind the user’s movements. It seemed likely that there is some optimal amount of averaging in the stabilization software, just as there are optimal gains for the mouse and trackballs. Experiment 2b was designed to find the optimal number of points to average over in the stabilization software.

4.1 Method

Experiment 2b compared several different amounts of averaging; the range of averaging amounts tested was chosen based on pilot testing to be one that would be likely to contain the optimal value. We considered testing software with dynamic averaging, where the amount of averaging varies depending on the velocity of the user’s finger on the pad, but eventually we decided to test only different amounts of simple averaging. We used six FP employees as subjects. Each subject performed the same task as in studies 1 and 2a, with six different amounts of averaging, in permuted order.

4.2 Results

The results are shown in Figure 3 and Table 7 below. As in Experiment 2a, Friedman’s test revealed no significant differences, but as the graph and table show, the time to hit the target was least for averaging over six points (0.2 seconds). Therefore, for Experiment 3, the remote control stabilization software was set to average over six points.

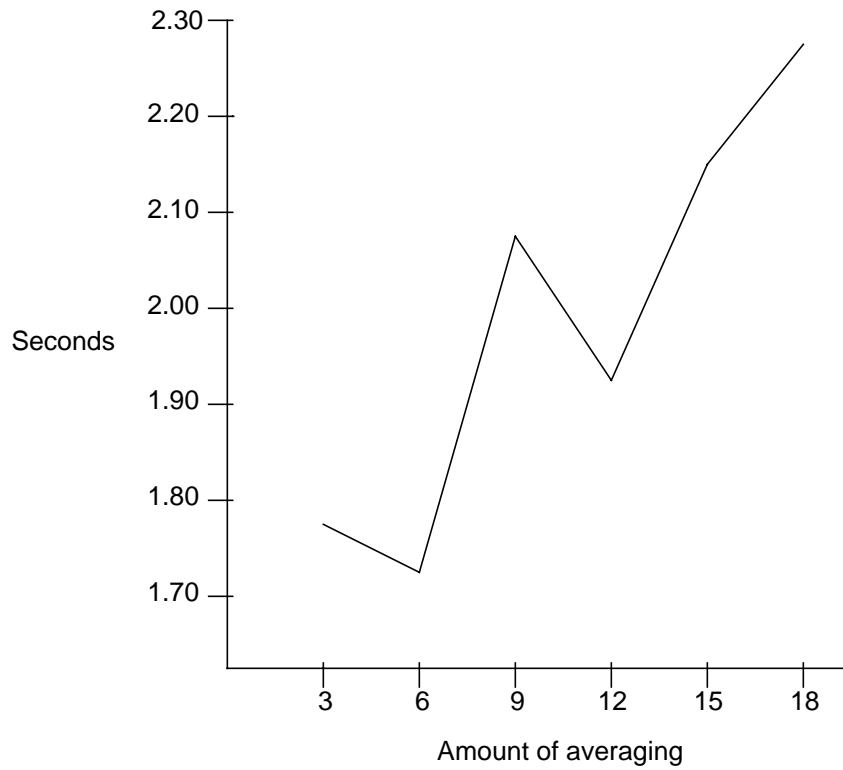


Figure 3. Optimization of remote control: average time to hit target

Amount of averaging	Time to hit target (secs)	Errors
3	1.78	.10
6	1.72	.10
9	2.08	.16
12	1.92	.05
15	2.14	.10
18	2.28	.12

Table 7. Optimization of remote control: average time to hit target and number of errors

5 Experiment 3: Optimized Comparison

Having found the optimal gains for the mouse and trackballs and the optimal amount of averaging for stabilizing the FP remote control, we could fairly compare these devices, with each device seen in its best light.

5.1 Method

5.1.1 Subjects

Eight subjects who were not employees of FirstPerson or Sun Interactive participated in the experiment. Each was paid \$15 for the one-hour session. Four of the subjects were preselected from our subject pool to have experience using a mouse, and the other four were preselected to have little or no mouse experience. Although there was some attempt to match the two groups with respect to age, on average the subjects without mouse experience were older than those having mouse experience.

5.1.2 Materials

Experiment 3 compared four devices: a mouse, two trackballs, and an FP remote control. The control software for each device was set to its optimal parameter as determined by Experiments 2a and 2b. The mouse was set to a gain of 8, the large trackball was set to 10, and the mini-trackball was set to 12. The remote control software was set to average the screen-pointer location across 6 points (i.e., the past 0.2 seconds; $a = 6$).

5.1.3 Design

The design of Experiment 3 was similar to that of Experiment 1. In the first phase, each subject completed four blocks of 50 trials, one block with each of the four pointing devices. The optimized values found in Experiments 2a and 2b were used for each of the devices. The order of the devices was permuted within each of the two groups of subjects (with and without mouse experience).

In the second phase, each subject completed three more blocks of trials using the FP remote control. The remote control stabilization software was varied on each of these trials. On the first and second blocks, the screen-pointer location was averaged dynamically, i.e., over different amounts depending on the velocity of the screen-pointer. On the last block, the screen-pointer location was not averaged at all, i.e., the software used on the final block of trials was the same as the non-stabilized software used in Experiment 1.

The second phase of the experiment had two purposes. First, we were experimenting with velocity-based averaging and wanted some exploratory data to help us set parameters for this more-complicated stabilization method. Since the first phase of the experiment took considerably less than the hour for which subjects were paid, it seemed reasonable to use the extra time to collect some data on this issue. Second, we wanted to see how practice would affect performance, to help us interpret the results of the second phase of Experiment 1. We ended Experiment 3 with what we knew was a relatively poor-performing pointer-positioning method—i.e., no stabilization at all—to see whether the extra practice would make up for the lack of stabilization.

5.1.4 Procedure

The procedure for each block of trials was nearly identical to that used in Experiment 1. The only difference was that each block of 50 trials was preceded by a short practice block of 10 trials, in order to allow the subjects to get some experience with the devices.

After performing the task with all four devices, subjects were asked to rank the four devices in order of preference.

5.2 Results

On average, across all four of the devices, the subjects without previous mouse experience were significantly slower than the subjects with mouse experience ($t = 3.43$; $p < .05$). However, the pattern of results across the different devices was similar for the subjects with and without mouse experience (see Tables 8 and 9). Therefore, the data were collapsed across this variable.

By Friedman's test, there was a significant difference between the pointing devices in how quickly users hit targets ($p < .01$). Subjects were fastest with the mouse, slightly slower with the mini-trackball and large trackball, and slowest with the remote control.

The error rates follow a similar pattern (see Table 9) and differ significantly between pointing devices. Subjects made fewest errors with the mouse, more with the mini-trackball and large trackball, and most with the remote control. The subjects' preference ranking of the devices was similar to the pattern of performance scores across devices but, as in Experiment 1, the differences among the rankings were not statistically significant (see Table 10).

Ss with mouse exp.	Mouse	Mini-trackball	Large trackball	FP rem. w/ avg. 6
S1	1.15	1.79	1.63	1.74
S2	1.23	1.37	1.44	1.54
S7	1.11	1.73	1.62	2.23
S8	1.52	2.64	1.78	5.35
<i>Average (S1,S2,S7,S8)</i>	<i>1.25</i>	<i>1.88</i>	<i>1.62</i>	<i>2.71</i>
Ss without mouse exp.				
S3	2.39	2.74	3.34	3.91
S4	2.54	2.33	4.63	3.01
S5	2.02	2.65	2.45	4.18
S6	2.20	2.85	3.32	3.63
<i>Average (S3,S4,S5,S6)</i>	<i>2.28</i>	<i>2.64</i>	<i>3.44</i>	<i>3.68</i>
Overall (all Ss)	1.77	2.26	2.53	3.20

Table 8. Average time to hit target (in seconds) by subject and input device

In the second phase of the experiment, subjects used the remote control for three more blocks of trials, two with velocity-based averaging and one with no averaging. The velocity-based averaging trials produced data that may be useful for finding the ideal parameters for this more complex method of stabilization, but those findings are exploratory and will not be reported here.

Experiment 3 was intended not only to provide a fair comparison of pointing devices, but also to clarify the effects, if any, of stabilization and practice on performance with the FP remote control. One way to test for an effect of stabilization is to determine whether the average-6 stabilization used in the device-comparison phase of Experiment 3 improved performance with the remote control relative to performance with the non-stabilized remote control in the corresponding phase of Experiment 1. To do this, t-tests were carried out on the time and error data for the remote control in the two experiments. Subjects who used average-6 stabilization (in Experiment 3) were slightly faster on the average at hitting targets than those who used no stabilization (in Experiment 1), but the difference was not statisti-

cally significant ($t = 1.27$; N.S.). On the other hand, the average number of errors was significantly lower for the average-6 group than for the no-stabilization group ($t = 3.04$; $p < .05$).

Ss with mouse exp.	Mouse	Mini-trackball	Large trackball	FP rem. w/ avg. 6
S1	.00	.02	.06	.02
S2	.14	.02	.28	.32
S7	.00	.02	.02	.04
S8	.04	.06	.10	.26
<i>Average (S1,S2,S7,S8)</i>	<i>.05</i>	<i>.03</i>	<i>.12</i>	<i>.16</i>
Ss without mouse exp.				
S3	.04	.10	.10	.22
S4	.04	.12	.24	.14
S5	.04	.06	.08	.12
S6	.02	.02	.10	.08
<i>Average (S3,S4,S5,S6)</i>	<i>.04</i>	<i>.08</i>	<i>.13</i>	<i>.14</i>
Overall (all Ss)	.04	.06	.13	.15

Table 9. Average number of errors per trial, by subject and input device

	Mouse	Mini-trackball	Large trackball	FP rem. w/ avg. 6
Ss with mouse exp.	1.5	2.1	2.5	3.9
Ss without mouse exp.	1.8	2.3	3.5	2.5
Overall (all Ss)	1.6	2.2	3.0	3.2

Table 10. Ranking of devices (1 = most preferred, 4 = least preferred)

Ss with mouse exp.	FP rem. w/ avg. 6	FP remote
S1	1.74	1.29
S2	1.54	1.42
S7	2.23	2.06
S8	5.35	3.89
<i>Average (S1,S2,S7,S8)</i>	<i>2.71</i>	<i>2.17</i>
Ss without mouse exp.		
S3	3.43	2.42
S4	3.91	3.41
S5	3.01	2.65
S6	4.18	2.92
<i>Average (S3,S4,S5,S6)</i>	<i>3.63</i>	<i>2.85</i>
Overall (all Ss)	3.17	2.51

Table 11. Average time to hit target (in seconds), by subject and stabilization method

A way to check the relative effect of stabilization and practice is to compare the difference in performance between the stabilized and non-stabilized remote control in Experiment 3 with the corresponding difference in Experiment 1. In Experiment 1, subjects hit targets faster using average-2 stabilization than they did with no stabilization, but because each subject used average-2 stabilization *after* no stabilization, practice was at least as likely an explanation for the improvement as was stabilization. In Experiment 3, subjects used the non-stabilized remote control in the last trial block of phase 2, after they had tried average-6 stabilization in the device-comparison phase of the experiment. In other words, the order of stabilized vs. non-stabilized trials was reversed between the two experiments. In Experiment 3, all eight subjects were faster in the final, non-stabilized trial block than in the average-6-stabilization block ($p < .01$ by sign test; see Table 11). This, combined with the aforementioned finding that stabilization did not significantly improve speed between the device-comparison trials of the two experiments, strongly suggests that a small amount of practice with the FP remote control improves target-hitting speed more than stabilization does.

Ss with mouse exp.	FP rem. w/ avg. 6	FP remote
S1	.02	.04
S2	.32	.22
S7	.04	.08
S8	.26	.36
<i>Average (S1,S2,S7,S8)</i>	<i>.16</i>	<i>.18</i>
Ss without mouse exp.		
S3	.22	.06
S4	.14	.30
S5	.12	.22
S6	.08	.22
<i>Average (S3,S4,S5,S6)</i>	<i>.14</i>	<i>.20</i>
Overall (all Ss)	.15	.19

Table 12. Average number of errors per trial, by subject and stabilization method

A similar analysis of the error data yielded a different result (see Table 12). In Experiment 1, subjects' error scores, like their time scores, were as good or better in the stabilized trial block presented at the end of the session than in the non-stabilized trial block presented earlier in the session, but it wasn't clear whether the improvement was due to stabilization or practice. In Experiment 3, subject accuracy was slightly worse in the final non-stabilized trial block than in the earlier stabilized trial block. Six of the eight subjects made *more* errors in the non-stabilized trial block than in the stabilized trial block, but this difference was not significant. Thus, practice did not overwhelm stabilization in the error data. This finding, combined with the aforementioned finding that stabilization significantly reduced target-hitting errors between the device-comparison trials of the two experiments, suggest that the primary benefit of stabilization may be to improve accuracy rather than speed.

5.3 Discussion

In Experiment 3, as in Experiment 1, the mouse was the clear winner in the device comparison. Even subjects who had little or no mouse experience performed best, in both speed and accuracy, with the mouse.

Also as in Experiment 1, subjects performed worst with the remote control. However, practice with the remote control improves performance considerably. With about 5-10 minutes of practice, subjects could use the remote control as easily as its most plausible competitor, the mini-trackball: the average time for the non-stabilized remote control on the last block of trials was almost as fast as the trackballs and faster than the optimally-stabilized remote control. This shows us that the best static stabilization is less effective than 180 trials of practice for improving target-hitting speed. Accuracy, however, is improved by static-average stabilization.

6 General Discussion and Conclusions

It seems clear that the mouse is the most accurate and quickest pointing device for the task used in these experiments. Note however, that for more realistic interactive TV situations, the mouse has a serious limitation: it requires a surface upon which to operate. For interactive TV, it is widely accepted that the controller must be hand-held. We included a mouse in these experiments not because it was in consideration for use in FirstPerson's interactive TV system, but rather because according to the research literature, it provides a benchmark that represents nearly ideal performance. And so it did.

Similarly, the regular trackball, which placed third in time, accuracy, and subject preferences in both experiments, requires a surface upon which to rest. Like the mouse, it was included because of its good showing in previous research rather than because it was under consideration for the interactive TV application.

A more serious contender for use in an interactive TV remote control is the mini-trackball, which placed second—just behind the mouse—in time, errors, and subject preference, in both Experiments 1 and 2. In fact, based on the results of Experiment 1, we recommended that the mini-trackball be considered the best alternative until the FP remote control was improved, or in case the company decided not to pursue development of the FP remote control. Experiment 3 strengthened that recommendation.

The so-called “pointing stick” was eliminated from further consideration after Experiment 1 because it was difficult to use, strongly disliked by subjects, and provided no clear opportunities for improvement.

The FP remote control, like the pointing stick, scored poorly in Experiment 1. However, unlike the pointing stick, which is a commercial product, the FP remote control was a prototype of a design still under development. Improvement was not only possible, but planned. Experiment 2b was an attempt to improve the usability of the remote control by optimizing the parameter of the average-N stabilization method. Experiment 3 then compared the more stable FP remote control with a mouse and trackballs, as well as with the non-stabilized remote control. (Experiment 3 also included some preliminary exploration of dynamic stabilization methods, which will not be discussed here.) The results indicate that static-averaging stabilization improves accuracy, but not speed. Speed, on the other hand, seems to improve greatly with relatively small amounts of practice.

The effect of various forms of stabilization on performance with the remote control needs to be studied further, both to look more rigorously at the effect of stabilization methods and to search for more effective stabilization methods. If further studies continue to show that simple stabilization has minimal impact on performance, more complicated stabilization methods, e.g., velocity-dependent averaging, may be necessary to further improve the performance of the remote control.

If it is decided that the FP remote control is unsatisfactory for interactive TV applications, the best alternative pointing device is the mini-trackball. If, on the other hand, the FP remote control is to be considered further, we have two alternative recommendations for improvements:

- If the touchpad on the remote control continues to be an absolute touchpad, develop and test stabilization methods that are more sophisticated than simple static averaging. For example, velocity-dependent averaging may provide stability while avoiding annoying lagging of the pointer. However, velocity-dependent averaging has at least three parameters. Finding the optimal combination of these parameters will be necessary, but it will not be easy.
- Consider replacing the absolute touchpad with a relative touchpad, similar to the one now used on Apple Computer's new PowerBook. This, combined with some sort of stabilization, would make operating the touchpad very similar to operating a handheld trackball, and so might improve performance to near the levels seen with the mini-trackball.

7 References

- Card, S.K., English, W.K., and Burr, B. (1978) "Evaluation of Mouse, Rate-Controlled Isometric Joystick, Step Keys, and Text Keys for Text Selection on a CRT," *Ergonomics*, 21(8), 1978, pages 601-613.
- English, W.K., Engelbart, D.C., and Berman, M.L. (1967) "Display-Selection Techniques for Text Manipulation," *IEEE Transactions in Human Factors in Electronics*, HFE-8, pages 21-31.
- Epps, B.W., and Snyder, H.L. (1986) "Comparison of Six Cursor Devices on a Target Acquisition Task," *SID Digest*, pages 302-305.
- Epps, B.W. (1987) "A Comparison of Cursor Control Devices on a Graphics Editing Task," *Proceedings of the Human Factors Society Annual Meeting*, pages 442-446.
- Fitts, P.M. (1954) "The Information Capacity of the Human Motor System in Controlling Amplitude of Movement," *Journal of Experimental Psychology*, vol. 47, pages 381-391.
- Hayes, W. (1963) *Statistics*, New York: Holt, Rinehart and Winston.
- Jellinek, H.D. and Card, S.K. (1990) "Powermice and User Performance," *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'90)*, pages 213-220.
- Johnson, J. (1994) "A Comparison of User Interfaces for Panning on a Touch-Controlled Display," *FirstPerson Technical Report FP-1994-1*. (In this collection)
- Johnson, J. and Keavney, M. (1994) "The Effect of Touch-Pad Size on Pointing Accuracy," *FirstPerson Technical Report FP-1994-2*. (In this collection)
- Keavney, M. (1994) "Factors Affecting Users' Ability to Hit Animated Characters," *FirstPerson Technical Report FP-1994-3*. (In this collection)
- Keavney, M. and Johnson, J. (1994) "An Evaluation of the Accuracy of a Touch-Sensitive Remote Control," *FirstPerson Technical Report FP-1994-4*. (In this collection)
- Plaisant, C. and Sears, A. (1993) "Touchscreen Interfaces for Alphanumeric Text Display," in B. Shneiderman (Ed.), *Sparks of Innovation in Human-Computer Interaction*, Norwood, NJ: Ablex Publ.
- Potter, R.L., Weldon, L.J., Shneiderman, B. (1993) "Improving the Accuracy of Touchscreens: An Experimental Evaluation of Three Strategies," in B. Shneiderman (Ed.), *Sparks of Innovation in Human-Computer Interaction*, Norwood, NJ: Ablex Publ.

Sears, A., and Shneiderman, B. (1993) “High Precision Touchscreens: Design Strategies and Comparisons with a Mouse,” in B. Shneiderman (Ed.), *Sparks of Innovation in Human-Computer Interaction*, Norwood, NJ: Ablex Publ.

Shneiderman, B. (1993) “Touchscreens Now Offer Compelling Uses,” in B. Shneiderman (Ed.), *Sparks of Innovation in Human-Computer Interaction*, Norwood, NJ: Ablex Publ.

Sperling, B.B., Tullis, T.S. (1987) “Are You a Better ‘Mouser’ or ‘Trackballer’? A Comparison of Cursor-Positioning Performances,” Interactive Poster at Graphics Interface and Human Factors in Computing Systems conference (CHI+GI’87).

About the authors

Jeff Johnson is a Staff Engineer in the Human Interface Engineering group of SunSoft Developer Products. He has worked in the computer industry for 18 years as a software developer, designer, manager, and researcher. Prior to working at SunSoft, he worked for Cromemco, Xerox, USWest, Hewlett-Packard, and Sun/FirstPerson. He has published numerous articles in the areas of human-computer interaction and technology policy. Johnson has a B.A. degree from Yale and a Ph.D. from Stanford. From 1991 to 1994, he was Chair of the Board of Directors of Computer Professionals for Social Responsibility, a public-interest organization that examines the impact of technology on society.

Mark Keavney received his Ph.D. in social psychology from Stanford University in 1993. His dissertation was on the psychology of humor. He worked as a user interface designer and usability tester for FirstPerson, Inc., from 1993 to 1994, and is presently working as a game designer for Sparkle, Inc., an educational software company.

© Copyright 1995 Sun Microsystems, Inc. The SML Technical Report Series is published by Sun Microsystems Laboratories, a division of Sun Microsystems, Inc. Printed in U.S.A.

Unlimited copying without fee is permitted provided that the copies are not made nor distributed for direct commercial advantage, and credit to the source is given. Otherwise, no part of this work covered by copyright hereon may be reproduced in any form or by any means graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an information retrieval system, without the prior written permission of the copyright owner.

TRADEMARKS

Sun, Sun Microsystems, and the Sun logo are trademarks or registered trademarks of Sun Microsystems, Inc. UNIX and OPEN LOOK are registered trademarks of UNIX System Laboratories, Inc. All SPARC trademarks, including the SCD Compliant Logo, are trademarks or registered trademarks of SPARC International, Inc. SPARCstation, SPARCserver, SPARCengine, SPARCworks, and SPARCcompiler are licensed exclusively to Sun Microsystems, Inc. NFS is a registered trademark of Sun Microsystems, Inc. All other product names mentioned herein are the trademarks of their respective owners.

For information regarding the SML Technical Report Series, contact Jeanie Treichel, Editor-in-Chief <jeanie.treichel@eng.sun.com>.

For distribution issues, contact Amy Tashbook Hall, Assistant Editor <amy.hall@eng.sun.com>.