# Chapter 8

# Longitudinal Tests

## Overview

This chapter describes the second set of tests that were undertaken by fewer subjects than the cross-sectional tests, but over a longer period of time. An explanation is given of how the procedures compared with the cross-sectional tests and how the reduced numbers of subjects and sounds were selected. The results are given in graphical form and some initial trends discussed.

## 8.1 Overview of Longitudinal tests

In section 7.8 the requirement for longitudinal tests was described; to study the development of a user's performance over an extended number of sessions. These tests were designed to plot the progress of three subjects over ten sessions. The test environment was exactly the same as for the cross-sectional tests and the same three interfaces were compared. There were no formal taped interviews with the test subjects but discussions were held at the end of each session.

The order that the subjects used the interfaces was kept constant throughout the ten sessions. In every case it was *mouse - sliders - multiparametric*. It is quite possible that the scores for the multiparametric interface were negatively affected by this, as the subjects may have been tired towards the end of a session, or they may have been 'induced' into an analytical learning mode by the mouse and sliders interfaces. All sessions had their sound examples arranged in increasing complexity.

To help prevent any tiredness in the subjects the overall number of sounds in each session was reduced. The selection criteria for these are outlined in the next section.

## 8.2 Choice of Musical Test Examples

The cross-sectional tests involved 24 sounds per session. Since this was very tiring for the participants it was decided to have just 9 sounds for the longitudinal tests. Nine of the sounds with high standard deviations were chosen. This was to ensure that the chosen sounds had been shown to elicit a range of responses from the users. In other words there would be no point in choosing a sound that was so hard to recreate that everyone got a low score, or conversely a sound so easy that every subject got a high score. Figure 8.1 shows the average standard deviation for every sound in the cross-sectional tests.

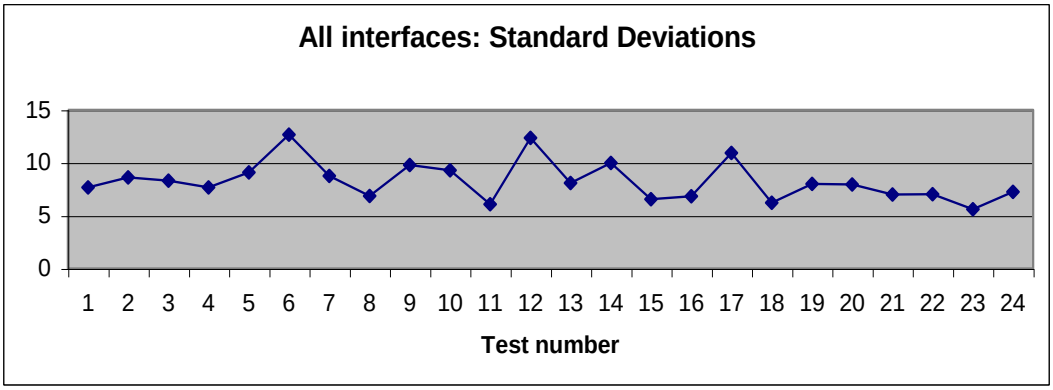


**Figure 8.1: Average Standard Deviations of cross-sectional sounds**

In addition it was required that 3 sounds were chosen from each of Group A, Group B and Group C (see section 7.2) so that the relative proportion of each group was the same as before.

## 8.3 Choice of Test Subjects

The three subjects highlighted in the previous chapter (section 7.7.4) were chosen as they demonstrated a suitably varied performance over the three sessions of the cross-sectional tests. Subject 5 had performed better with the multiparametric interface, subject 4 had scored well with the sliders, and subject 9's performance had been reasonably similar on all three interfaces.

Each of these subjects was also willing to embark upon another ten sessions. Each session took approximately 15 minutes to complete (5 minutes on each interface). The subjects completed on average between 2 and 3 sessions per week over a five-week period.

## 8.4 Results

These experiments have yielded ((9 sounds) x (3 interfaces) x (3 subjects) x (10 sessions)) = 810 individual results. They were analysed and plotted in the same manner as for the cross-sectional tests (see sections 7.5 & 7.7). The full results are shown in Appendix C. There are many conclusions that can be drawn from the trends shown in the results, but the most significant four conclusions are now presented.

### 8.4.1 Multiparametric is best for complex tests

Figure 8.2 shows how the three interfaces fared with different levels of test complexity (averaging across all the test subjects).

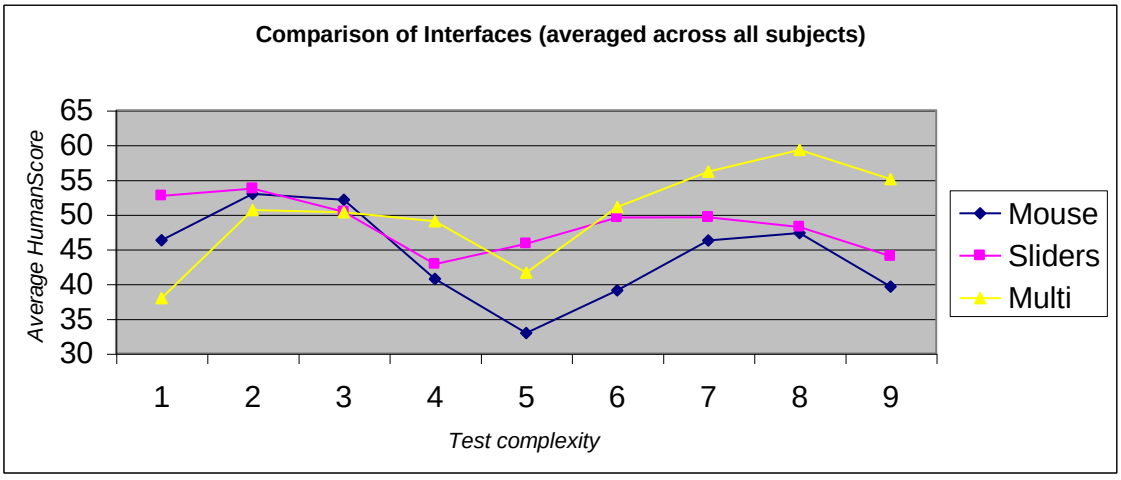


**Figure 8.2: The effect of Test Complexity**

The x-axis represents the nine sounds used in the longitudinal tests from sound 1 (simple) to sound 9 (most complex). The y-axis represents the average scores for each test. We can see that for tests 6 to 9 (where more than one parameter changes simultaneously) the multiparametric interface has the best results. This feature was also seen and discussed in section 7.7.3. Even with computer marking (which tends to give lower scores for the multiparametric tests) the results are best for sounds 8 and 9 (the most complex).

**For the more complex tests the multiparametric interface gives the best results.**

### 8.4.2 Multiparametric is *initially* worst for simple tests

Figure 8.2 (above) indicates that the multiparametric interface is indeed the worst overall for the simpler tests. However, this graph does not show the progression of scores over time.

Figure 8.3 shows the results for Sound 1 (the simplest test) over the ten session test period. Time is shown along the x-axis, and the average scores are shown, as before, on the y-axis.

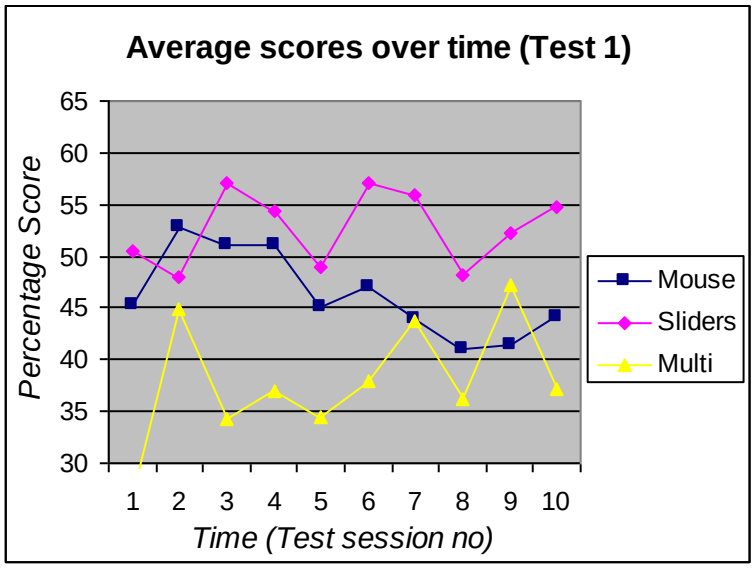


**Figure 8.3: Trends in the simplest test.**

Although the multiparametric interface has the lowest overall scores for the simpler tests, it does have a significant upward trend. Note how the trend for the 'mouse' interface is actually *downward* over time! The sliders interface is clearly best, with a slight upward trend.

**For the simpler (one parameter) tests the multiparametric interface comes out worst.**

### 8.4.3 Multiparametric nearly always has an upward trend

Contrast Figure 8.3 (above) with Figure 8.4 which shows the same type of plot over time but for the most complex test.

In Figure 8.4 the multiparametric interface is always the best interface and still with a strong upward trend. Also note how the trends for *both* the mouse and sliders interfaces are downward over time!

Taken together, these graphs imply that both of the sliders-based interfaces are clearly best for simple, single parameter changes. Their superiority is challenged by

the multiparametric interface where parameters change one after the other. However, they are truly beaten where several parameters change at once! This is an important result, as it demonstrates that it will probably take a complex interface to cope with a non-trivial control domain. We shall return to this point in the conclusions.

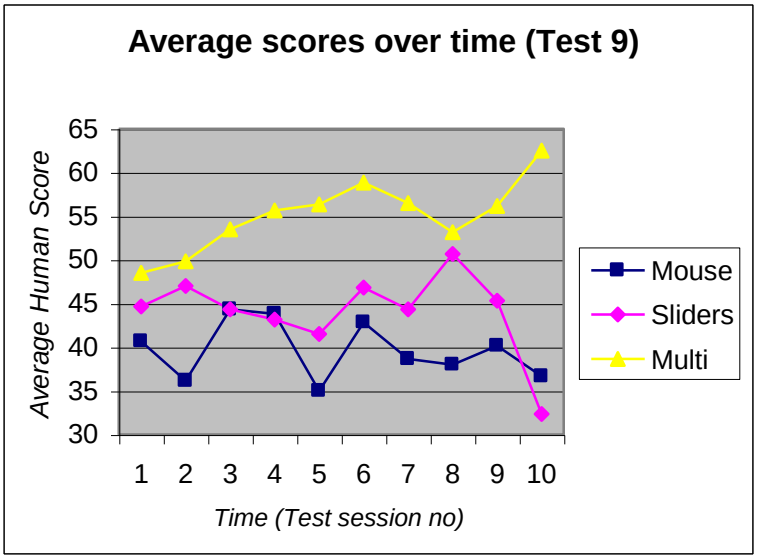


**Figure 8.4: Trends in the most complex test.**

As if to emphasise this point it is remarkable to notice that the multiparametric interface nearly always followed an upward trend - for *every subject* and for *nearly every sound* of whatever complexity. This is simply not true for the other two interfaces. Table 8.1 shows the number of overall upward performance trends (i.e. across all ten sessions) for each person on each interface.

| | **Mouse** | **Sliders** | **Multiparametric** |
|---|---|---|---|
| **Subject 5** | 2 | 8 | 8 |
| **Subject 9** | 6 | 3 | 8 |
| **Subject 4** | 5 | 7 | 8 |
| **Total / 27** | 13 | 18 | 24 |
| **Percentage** | 48% | 67% | 89% |

**Table 8.1: Relative number of upward performance trends**

For the mouse over half the trends are downward or flat. With the sliders interface the performance over time is improved to about two thirds upward trends. The multiparametric interface is best of all with nearly 90% upward trends.

**Across the whole range of tests the multiparametric interface has the most consistent upward trend.**

### 8.4.4 Mouse interface does not have long-term potential

Figure 8.5 shows the performance of each interface over time (averages for all subjects and all test complexities) along with a linear trend provided by *Excel*.

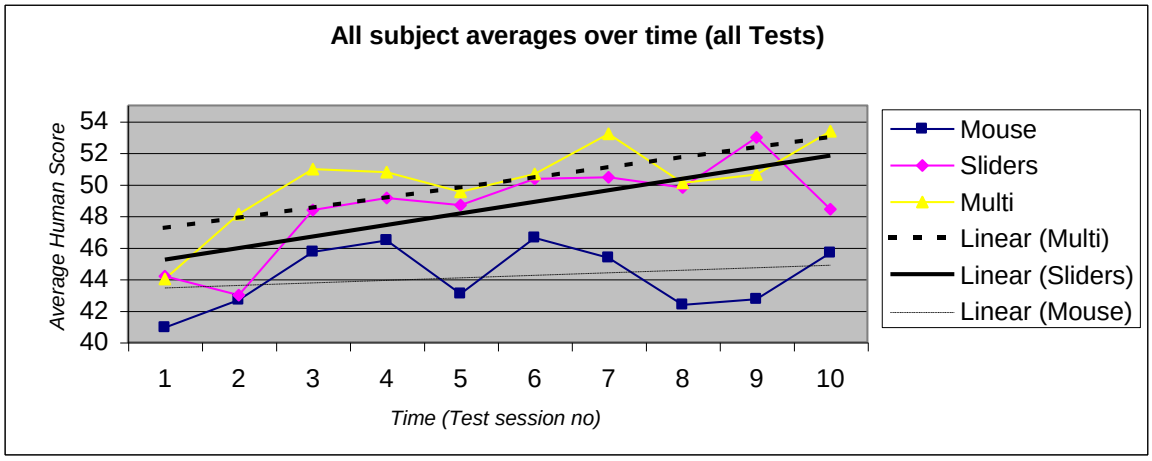


**Figure 8.5: Overall summary of scores**

The mouse interface can be seen to give the lowest average scores, with a very slight upward trend over time. This seems to concur with the users' perceptions that it was easy to learn, but was difficult to get any better on (see section 7.6.6). In contrast, both the sliders and multiparametric interfaces showed an average upward trend, the difference in the interfaces only really becoming clear when the complexity of the test is taken into account (section 8.4.1).

This seems to imply that the mouse (in conjunction with its on-screen sliders) is an interface that people can relate to and operate quickly. It may be this instant appeal of mouse interfaces that is the reason for the WIMP interface being popular. Alternatively, it may simply be that most of the test subjects are familiar with the mouse from other everyday computing operations and this gives them an initial 'head-start'. However as the task domain increases in complexity, and as people spend longer on the system, then other interfaces become more suitable.

**The mouse interface gives the most consistent (but low) scores across all tests and subjects.**

## 8.5 Summary

These tests have given an indication of how performance varies over an extended period of time. Each of the interfaces has been used by three test subjects over ten

sessions. The results clearly show that the multiparametric interface gives the best results on the complex tests and the worst results on the simplest tests. The mouse, however, gives the worst *overall* result and quite often has a downward trend over time! In contrast the multiparametric interface nearly always produces an upward trend over the ten sessions.

The results of both sets of tests (cross-sectional and longitudinal) as well as the recorded comments from the users are analysed in more detail and summarised in the next chapter.