TRIS
Design and evaluation in the real world: communicators and advisory systems

Introduction

Textbooks about design and usability testing often make the processes sound straightforward and able to be followed in a step-by-step manner. However, in the real world bringing together all the different aspects of a design is far from straightforward. It is only when you become involved in an actual design project that the challenges and multitude of difficult decisions to be made become apparent. Iterative design often involves carrying out different parts of a project in parallel and under tremendous pressure. The need to deal with different sets of demands and trade-offs (e.g. the need for rigorous testing versus the very limited availability of time and resources) is a major influence on the way a design project is carried out. The aim of this case study is therefore to convey what interaction design is like in the real world by describing how others have dealt with the challenges of an actual design project. More specifically we want you to see an example of how design and evaluation go hand-in-hand. So much so that whenever you develop a design you need to evaluate it.

This case study examines the redesign of a large interactive voice response system. In the original design, the focus was on developing a system where the programmers used themselves as models of the users. Furthermore, the programmers were more concerned with developing elegant programs than with users' needs for easy interaction. As you will see, this caused a mismatch between their design and how users tried to find information. This is a common predicament and interaction designers are often brought in to fix already badly designed systems.

The aim of this case study is to:

- Show how design and evaluation are brought together in the redesign of a large system.

Interaction Design: Beyond human-computer interaction Sharp, Rogers and Preece
© 2007 John Wiley & Sons, Ltd ISBN-13: 978-0-470-01866-8

- Show how a combination of evaluation methods are used in practice.
- Describe some of the design trade-offs and decisions made in the real world.

Key issues

User-centered approaches to interaction design involve iterative cycles of design-evaluate-redesign as development progresses from initial ideas through various prototypes to the final product. How many cycles need to take place depends on the constraints of the project (e.g. how many people are working on it, how much time is available, how secure the system has to be). To be good at working through these cycles requires a mix of skills involving multitasking, decision-making, team work and firefighting. Many practical issues and unexpected events also need to be dealt with (e.g. users not turning up at testing sessions, prototypes not working, budgets being cut, time to completion being reduced, designers leaving at crucial stages). A design team, therefore, must be creative, well organized, and knowledgeable about the range of techniques that can be brought into play when needed. Part of the challenge and excitement of interaction design is finding ways to cope with the diverse set of problems confronting a project.

A multitude of questions, concerns and decisions come up throughout a design project. No two projects are ever the same; each will face a different set of constraints, demands, and crises. Throughout the book we have raised what we consider to be general issues that are important in any project. These include how to involve users and take their needs into account, how to understand a problem space, how to design a conceptual model, and how to go about designing and evaluating interfaces. In the following case study, we focus on some of the more practical problems and dilemmas that can arise when working on an actual project.

We present the case study using a set of questions that draw out a number of key issues concerned with evaluating the current system when redesigning any large system.

Redesigning part of a large interactive phone-based response system

In this case study, we focus on the redesigned of a large system that provides the general public with advice about filling out a tax return—and those of you who have to do this know only too well how complex it is. The original product was developed not as a commercial product but as an advisory system to be interacted with via the phone. We report here on the work carried out by usability consultant Bill Killam and his colleagues, who worked with the US Internal Revenue Services (IRS) to evaluate and redesigned the telephone response information system (TRIS) (Killam & Autry, 2000).

Although this case study is situated in the US, such phone-based information systems are widespread across the world. Typically, they are very frustrating to use. Have you been annoyed by the long menus of options such systems provide when you are trying to buy a train ticket or when making an appointment for a technician to fix your phone line? What happens is that you work your way through several different menu systems, selecting an option from the first list of, say, seven choices, only to find that now you must choose from another list of five alternatives. Then, having spent several minutes doing this, you discover that you made the wrong choice back in the first menu, so you have to start again. Does this sound familiar? Other problems are that often there are too many options to remember, and that none of the options seems to be the right one for you. In such situations, most users long for human contact, for a real live operator, but of course there usually isn't one.

TRIS provided information via such a myriad of menus, so it was not surprising that users reported many of these problems. Consequently a thorough evaluation and redesign was planned. To do this, the usability specialists drew on many techniques to get different perspectives of the problems and to find potential solutions. Their choice of techniques was influenced by a combination of constraints: schedules, budgets, their level of expertise, and not least that they were working on redesigning part of an already existing system. Unlike new product development, the design space for making decisions was extremely limited by existing design decisions and the expectations of a large existing user population.

Background

Everyone over age 18 living in the US must submit a tax return each year either individually or included in a household. The age varies from country to country but the process is fairly similar in many countries. In the US this amounts to over 100 million tax returns each year. Completing the actual tax return is complex, so the IRS provides information in various forms to help people. One of the most used information services is TRIS, which provides voice-recorded information through an automated system. TRIS also allows simple automated transactions. Over 50 million calls are made to the IRS each year, but of these only 14% are handled by TRIS. This suggested to the designers that something was wrong.

The redesign

How do users interact with the current version of TRIS? The users of TRIS are the public, who get information by calling a toll-free telephone number. This takes them to the main IRS help desk, which is in fact the TRIS. The interface with TRIS is recorded voice information, so output is auditory. Users navigate through this system by selecting choices from the auditory menu that they enter by typing on the telephone keypad. First, the users have to interact with the Auto Attendant portion of the system—a sort of simulated operator that must figure out what the call is about and direct it to the proper part of the system. This sounds simple but there is a problem. Some paths have many subpaths and the way information is classified under the four main paths is often not intuitive to users. Furthermore, some of the functionality available through TRIS is provided by two other independent systems, so users can become confused about which system they are dealing with and may not even know they are dealing with a different system. Users get very few clues that these other systems exist or how they relate to each other, yet suddenly things may be quite different—even the voice they are listening to may change. Navigating through the system, with its lack of visual feedback and few auditory clues, is difficult. Imagine being in a maze with your eyes blindfolded and your hands tied so you can't feel anything, and where the only information you get is auditory. How can you possibly remember all the instructions and construct an accurate mental model in your head to help you?

Once in TRIS, users can take various paths that:

- Provide answers to questions about tax law (provided by one of the two other computer systems accessible through TRIS).
- Allow people to order all the forms and other materials they need to complete their tax return (provided by the two other systems accessible through TRIS).
- Perform simple transactions, such as changing a mailing address, ordering a copy of a tax return, or obtaining answers to specific questions about a person's taxation.
- Reach a live operator if none of the above options are applicable or the user cannot figure out how to use the system.

Activity

Why is developing an accurate mental model of TRIS difficult for users?

Comment

Much of TRIS is hidden to the users. Their interaction with it is indirect, through listening to responses from the system and pressing various keys (whose meaning is always context dependent). There is no visual interface and users have only speech output to support their mental model development. Because speech is transient, unlike visual feedback, users must work out the conceptual model without visual cues. The user interface to this system is a series of menus in a tree structure and, since human short-term memory is limited, the structure of

the system must also be limited to only a few branches at each point in the tree. Another problem is that TRIS accepts input only from the telephone number keypad, so it's not possible to associate unique or meaningful options with user choices. ■

What are the main problems identified with the existing version of TRIS? Because one of the main problems users have when using TRIS is developing a mental model of the system it is hard for users to find the information they need. In addition, TRIS was not designed to reveal the mapping of the underlying systems and often did things that made sense from a processing point of view but not from the user's. This is probably because the programmers took a data-oriented view of the system rather than a user-oriented one. For example, TRIS used the same software routine to gather both a social security number and an employee identification number for certain interactions. This may be efficient from a code-development standpoint, since only one code module needs to be designed and tested, but from the user's perspective it presented several problems. The system always had to ask the user which type of number was expected, even though only one of these numbers made sense for many questions being asked. Consequently, many users unfamiliar with employee identification numbers were not sure what to answer, those who knew the difference wondered why the system was even asking, and all users had yet another chance to make an entry error.

What methods did the usability experts use to identify the problems with the current version of TRIS? To begin with the usability specialists did a general review of the literature and industry standards and identified the latest design guidelines and current industry best practices for interactive voice response (IVR) systems. These guidelines formed the basis for a heuristic evaluation of the existing TRIS user interface and helped identify specific areas that needed improvement. They also used the GOMS keystroke-level modeling technique to predict how well the interface supported users' tasks. Menu selection from a hierarchy of options is quite well suited to a GOMS evaluation, although certain modifications were necessary to estimate values for average performance times.

What did they do with the findings of the evaluation? Once the analysis of the existing interface and user tasks was complete, the team then followed a set of design guidelines and standards, to develop three alternative interfaces for the Auto Attendant part of TRIS. An expert peer panel then reviewed the three alternatives and jointly selected the one that they considered to have the highest usability. The usability specialists also performed a further GOMS analysis for comparison with the existing system. The analysis predicted that it would only take 216.2 seconds to make a call with the new system, compared with 278.7 seconds with the original system. While this kind of prediction can highlight possible savings, it says little about which aspects of the redesign are more effective and why. The usability specialists, therefore, needed to carry out other kinds of user testing.

Activity

Why is it that the results from a GOMS analysis do not necessarily predict the best design?

Comment

The keystroke-level analysis predicts performance time for experts doing a task from beginning to end. Not all of the users of TRIS will be experts, so performance time is not the only predictor of good usability. ■

The usability specialists did *three iterations* of user testing in which they simulated how the new system would work. When they were confident the new Auto Attendant interface had sufficient usability, they redesigned a subset of the underlying functionality. A new simulation of the entire Auto Attendant portion of TRIS was then developed. It was designed to support two typical tasks that had been identified earlier as problematic, to:

- find out the status of a tax refund.
- order a transcript of a tax return for a particular year.

These tasks also provide examples of nearly all of the user–system interactions with TRIS (e.g. caller identification, numeric data entry, database lookup, data playback, verbal instructions, etc.). A separate simulation of the existing system was also developed so that the new and existing designs could be compared. The user interaction was automatically logged to make data collection easier and unobtrusive.

What conflicts can arise when suggesting changes for improvement? When carrying out an evaluation of an existing product, often ‘‘jewels in the mud’’ stick out—glaring usability problems with a system that, if changed, could result in significant improvements. However, conflicts can arise when suggesting such changes, especially if they may decrease the efficient running of the system. The usability specialists quickly became aware that the TRIS system was making too many cognitive demands on users. In particular, the system expected users to select from too many menu choices too quickly. They also realized that immediate usability improvements could be gained by just a few minor changes: breaking menu choices into groups of 3–5 items; making the choices easier to understand; and separating general navigation commands (e.g. repeat the menu or return to the top menu) from other choices with pauses. However, to make these changes would require adding additional menus and building in pauses in the software. This conflicts with the way engineers write their code: they are extremely reluctant to purposely add additional levels to a menu structure and resist purposely slowing down a system with pauses.

Activity

The gap between programmers' goals and usability goals is often seen in large systems like TRIS that have existed for some time. How might such problems be avoided when designing new systems?

Comment

It can be hard to get changes made when a system has been in operation for some time, but it is important for interaction designers to be persistent and convince the programmers of the benefits of doing so. Involving users early in design and frequent cycles of 'design-test-redesign' helps to avoid such problems in the design of new systems. □

How were the usability tests devised and carried out? In order to do usability tests, the usability specialists had to identify goals for testing, plan tasks that would satisfy those goals, recruit participants, schedule the tests, collect and analyze data, and report their findings. Their main goals were to:

- evaluate the navigation system of the redesigned TRIS Auto Attendant
- compare the usability of the redesign with the original TRIS for sample tasks.

Twenty-eight participants were recruited from a database of individuals who had expressed interest in participating in a usability test. There was an attempt to recruit an equal number of males and females and people from a mixture of education and income levels. The participants were screened by a telephone interview and were paid for their participation. The tests were conducted in a usability lab that provided access to the two simulated TRIS systems (the original design and the redesign). The lab had all the usual features (e.g. video cameras) and a telephone. Timestamps were included in the videotape and the participants' comments were recorded.

The order of the tasks and the order in which the systems were used was counter-balanced. This was done so that participants' experience on one system or task would not distort the results. So, half the participants first experienced the original TRIS design and the other half first experienced the redesigned TRIS system. That way, if a user learned something from one or other system the effects would be balanced. Similarly, the usability specialists wanted to avoid ordering effects from all the participants doing the same task first. Half the participants were therefore randomly allocated to do task A first and the other half to do task B. Taking both these ordering effects into account produced a 4 × 4 experimental design with eight participants for each condition.

TRIS is complex, particularly the mapping between TRIS and the underlying functionality, although the system's purpose is clearly defined. By the time the usability specialists started the tests, they believed that they had fixed the major usability problems because they had responded first to the expert reviewers' feedback and then to the GOMS analysis. They were therefore confident that the new design would be better than the original one, but they had to demonstrate this to the IRS. This style of testing was also possible because there were thousands of potential users and the cost savings over 50 million calls justified the cost of this elaborate testing procedure.

How did they ensure that the participants tested were a representative set of users? In order to get demographic information to make sure the participants were representative, a questionnaire was given to all of them. It revealed a broad range of ethnicity, educational accomplishment, and income among the 18 women and 14 men who took part in the tests. Most had submitted tax returns during the last five years and most were experienced with interactive voice response systems. Eight participants indicated strong negative feelings about IVR systems, saying they were frustrating, time-consuming, and user-unfriendly.

What data was collected during the user testing? A total of 185 subnavigation steps made up the two tasks for the current TRIS. Participants successfully completed 91 steps on their first attempt (49% of the total). This was compared with a similar number of steps for the redesigned system: 187 subnavigation steps made up the same tasks for the redesigned TRIS. Participants were able to complete 117 of the steps on the first attempt (62% of the total), indicating an improvement of over 10%.

The average time to perform tasks was also analyzed. The summary data for the two tasks is shown in Table 1. As you can see, performance time on the redesigned system was much better for both tasks.

Task	Original system (s)	Redesigned system (s)
A	264.3	186.9
B	348.7	218.1

Table 1 *Average total task completion time by systems in seconds (s)*

How was the user's satisfaction with the system assessed? At the end of each task, participants were asked to evaluate how well they thought the system enabled them to accomplish their tasks by completing a user satisfaction questionnaire. The responses again indicated that participants thought the redesign was easier to use and they preferred it. Regardless of the order in which participants used the two systems, the scores on the *redesigned* system were consistently much better than for the *original* system. The questionnaire provided statements that the participants had to rate on a seven-point scale. The difference between the two systems was highly significant, averaging over three rating-scale points higher on each statement.

Activity

User satisfaction questionnaires like the ones just described enable usability specialists to get answers to questions they regard as important. How can you make sure you collect opinions on all the topics that are most important to users?

Comment

Asking users' opinions informally after pilot testing the questionnaire helps to make sure that you cover everything, but it is not foolproof. Furthermore, you may not want to increase the length of the questionnaire. Two other approaches that could be used separately are to ask users to think aloud and to use open-ended interviews. However, the think aloud method can distort the performance measures, so that is not such a good idea. Open-ended interviews are better, and this was done by the usability specialists in this case. □

Participants were also invited to make any additional comments they wanted about the two systems. These were then categorized in terms of how easy the new system was considered to navigate, whether it was less confusing, faster, etc. Specific complaints included that some wording was still unclear and that not being able to return to previous menus easily was annoying. No matter how much usability testing and redesign you do, there is always room for improvement.

Would it have been better to redesign the entire system? It would have been far too expensive and time-consuming to redesign and test the whole system. A skill that usability specialists need when dealing with this much complexity is how to limit the scope of what they do and still produce useful results.

What other design features could be considered besides improving efficiency? Given that the system is aimed at a diverse set of users, many whose native language is not English, a system that uses different languages would be useful. A range of voices could also be tested to compare the acceptability of different kinds of voices.

Summary

This case study has illustrated how to use different techniques in the evaluation and redesign of a system. Expert critiques and GOMS analyses are both useful tools for analyzing current systems and for predicting improvements with a proposed new design. But until the systems are actually tested with users, there is no way of knowing whether the predictions are accurate. What if users can theoretically carry out their tasks faster but in practice the interface is so poor that they cannot use it? In many cases, testing with real users is needed to ensure that the new design really does offer an improvement in usability. In this case study, results from usability testing were able to indicate that not only was the new design faster but users also liked it much better.

Reference and Further Reading

KILLAM, H. W. and AUTRY, M. (2000) IVR interface design standards: A practical analysis. In Proceedings of HFES/IEA 44th Annual Meeting. This paper describes aspects of the TRIS study in more detail.