JUS
JOURNAL OF USABILITY STUDIES

Issue 3, Vol. 1, May 2006, pp. 121-135

Case Study: Conducting large-scale multi-user user tests on the United Kingdom Air Defence Command and Control system

Elliott Hey
IBM UK LTD
Birmingham Road, Warwick
elliott_hey@uk.ibm.com

Abstract

IBM was contracted to provide a new Air Defence Command and Control (ADCC) system for the Royal Air Force. The IBM Human Factors (HF) team was responsible for the design of the operations room, workstations and the graphical user interfaces. Because the project was safety-related, IBM had to produce a safety case. One aspect of the safety case was a demonstration of the operational effectiveness of the new system.

This paper is an in-depth case study of the user testing that was carried out to demonstrate the effectiveness of the system. Due to time constraints the HF team had to observe five participants working simultaneously. Further, to provide a realistic operational environment, up to twenty-eight operators were required for each test. The total effort for this activity was four person-years. The paper will detail the considerations, challenges and lessons learned in the creation and execution of these multi-user user tests.

Keywords

Multi-user, user test, safety, process, training, Air Defence, Command and Control.

Introduction

The UK ADCC system is used by a team of Royal Air Force (RAF) operators to police the UK's airspace. The team is roughly divided into two groups. One group is responsible for monitoring and correlating the received information (for example, from radar) and identifying (i.e. friend or foe) all the aircraft within the UK's airspace to produce a clear air picture. The second group is mainly responsible for the management and coordination of the RAF's aircraft to police the airspace. This responsibility includes guiding fighters to intercept unfriendly or suspicious aircraft and managing and coordinating exercise and practise flying.

An ADCC system has similarities to a commercial air traffic control system. A picture is presented to the operators showing aircraft movements on a geographical display. The operators can interact with these aircraft symbols by calling up textual information on the display or by speaking directly to the pilots on the radio. The primary goal of an air traffic control system is to ensure that aircraft do not collide. The primary goals of the ADCC system are to identify these aircraft as friend or foe, and to direct military aircraft to targets often under, over or through commercial air routes. Consequently there is a close working relationship between the operators of these two systems to ensure the UK's airspace is safe.

IBM was contracted to design and implement the new ADCC system for the Ministry of Defence (MoD) on behalf of the RAF. The MoD had rated this system as being safety-related; this means that a system or human error could contribute to the loss of human life. For this reason IBM had to produce an acceptable safety case before the customer would take possession of the new system. A safety case will contain evidence on aspects such as hardware and software reliability, working environment, and HF. For each aspect, the safety case will contain evidence gathered from areas such as functional test results, meeting minutes, processes, user tests and statistical test data. Using this evidence the safety team will form an argument to state that the system is safe enough to use. The customer will then assess the safety case and either accept it, or reject it and request further evidence or a stronger argument.

On most systems that require user interaction the majority of errors are made by humans; however, these are usually caused by poor HMI design or ineffective training. It is for this reason that HF figures so prominently in safety cases for safety-related systems. Starting with the premise that humans will make errors, the safety case needs to contain evidence that these have been adequately mitigated.

The section in the safety case that discusses the HF component is called the Human Factors Argument. The design of the Human Factors Argument is customised depending on the type of system and its safety categorisation. Our approach was to discuss five main aspects that would provide the evidence required to convince the customer that the system was safe enough to use. These five aspects were: effective procedures, effective training, rigorous development process, user assessment, and a user interface critique [4]. The user assessment aspect was a set of large-scale multi-user user tests.

This user assessment was originally planned as an MoD activity. The MoD requested (a contractual change) IBM to undertake this activity sometime after the contract was signed. Therefore, this activity was squeezed into the programme and replaced other simpler activities that were previously planned by IBM to collect the data.

This paper is a case study of the multi-user user test activity performed on this project. It will detail the considerations, challenges, process and lessons learnt in the creation and execution of these large-scale multi-user user tests. The following eight sections will provide more information on the purpose of the user tests, planning and logistics, training, conducting the user tests, results, lessons learned, conclusion and Practitioners' takeaways.

Purpose

The main purpose of the large-scale multi-user user tests was to gather data that could be used as evidence to construct arguments within the IBM safety case. The data would be used to show that IBM had done the following:

Met various user, system and safety requirements
There were many user, system and safety requirements that required evidence to be gathered from a user test activity. It was unlikely that strong evidence could have been gathered using any other mechanism for many of these requirements.
Met the derived safety requirements
During the development of a large system many derived safety requirements are produced in addition to the user, system and safety requirements defined in the contract. A derived safety requirement might be a clarification of a contractual requirement or may be due to the way the new system has been implemented.

Adequately mitigated previously identified hazards
During the programme hazards were identified using a Hazard and Operability Study (HAZOPS [1]). This HAZOP study is a recognised method for identifying potential hazards (for example, sharp corners on desks) and operability problems (for example, inappropriate visual display) caused by inadvertent or poor design.

Not introduced user problems by deviating from established HMI standards and design patterns or by implementing novel or contradictory design ideas
When designing a complex system there are occasions when it is necessary to deviate from a contracted design standard (such as Defence Standard 00-25 [2]). For example, many alert systems use the categories Warning, Caution and Information. The design deviated from this common standard and introduced four new categories: Critical Action, Must Action, Must Know and Should Know. The argument was that when these categories were targeted at the correct individual they were more useful, understandable and thus easier to learn. As this was a different categorisation scheme, evidence had to be collected to show that the introduction of these new categories would not cause the user confusion.

Provided an HMI that would allow the operators to work safely, maintain situational awareness and work at an appropriate work rate in an operational context
The HMI was a composite of user interfaces, the workstations and the operations room, all of which had been redesigned as part of the programme. In order to assess how well the HMI allowed the user to work safely, the users had to have a representative amount of workload placed upon them. It is obviously easier for a user to work safely when they have only one tenth of the workload they are used to. Also, being able to work safely requires the user to be able to obtain and maintain an appropriate level of situational awareness, i.e. the user must understand what is going on around them for them to be proactive. It is not appropriate to assess how well the HMI is allowing the operator to work safely if the operator is not requested to obtain and maintain his or her situational awareness. These three elements (safety, situational awareness and work rate)

as described are not mutually exclusive and can have a large impact on each other.

Provided effective training materials
If the training is not effective, (for example, is not complete or does not accurately relay the underlying concepts of the design) then the operators may be disadvantaged. These ineffective training materials can initially cause the operators to make errors or interact with the system inappropriately. It is only with experience or when serious safety incidents come to light that the ineffectiveness of the training materials is identified.

It was necessary for the user tests to allow for qualitative and quantitative data to be gathered. The IBM safety team did not want to solely rely on subjective data to form their arguments. The qualitative data was collected during expert walkthroughs and large-scale multi-user user tests. [The expert walkthroughs will not be described in this paper.] The quantitative data was gathered during the multi-user user tests using the Human Error Assessment Reduction Technique (HEART).

Planning and Logistics

The Purpose section above described the types of data that were required to be collected. IBM had then to create the plan that would describe how that data would be collected. The HF team had to consider such issues as:

- What type of test could be conducted to gather realistic contextual data?
- How could the test be structured to gather specific types of data (for example, data related to hazards)?
- Where and when would the test be conducted?
- What system components were required for the test?
- What was the required level of maturity of the system components?
- Who was required to support the test?
- What were the logistical issues in organising the tests?

The ADCC system originally required thirty-two different roles to operate the system. Because of the introduction of new technology and improved HMI design on the new system this number was reduced to twenty-two roles. However, assessing twenty-two roles still posed an enormous challenge, especially when we would require a few users from each role to collect appropriate statistical data.

The team decided to assess the roles that performed tasks that had significant safety or workload implications. These were also the same roles that the designers had focused and optimised the HMI design on; these roles were deemed key to the successful operation of the system. However, it was still important to demonstrate that although the design had been optimised for a few roles, the other roles would still be able to use the HMI effectively. Therefore, two types of test were designed: user tests for the more safety- and workload-related roles (i.e. key roles), and expert walkthroughs for the other roles. [This paper will not discuss the expert walkthroughs.]

The team also calculated whether the roles for the user tests would allow enough of the appropriate types of data to be collected. This was a simple mapping of the roles' tasks to the types of data such as requirements and hazards. However, when the team looked at gathering data related to workload and safety they saw that the key roles would need to have an enormous amount data and input from other team members to be realistic.

The team eventually concluded that some of the key roles required an entire team of operators to produce an environment that would be realistic and provide an appropriate workload. For example, the team would have to perform activities such as: keeping the system running, producing a clear geographical picture, taking instructions from the participant, communicating with participants on emerging issues, driving simulated military aircraft, relaying instructions to pilots, and communicating with other ADCC operators in the UK and Western Europe. Having an appropriate workload was essential as this has a significant impact on determining how safely an operator can perform. It was determined that the user tests required sets of up to twenty-eight operators to work simultaneously in order to create an appropriate environment. This still included some supporting operators playing multiple roles within the user tests.

Typically user tests are conducted individually one after the other. However, the team did not have the time to perform twenty user tests consecutively. The team therefore looked into the possibility of replacing some of the supporting operators with participants from the key roles. This plan allowed the team to perform fewer user tests but created other problems. For example, observing five key roles performing together would require five sets of observers. Also, the chance of repeating the scenario exactly the same each time was reduced because the participants' actions were not scripted. Each participant became a supporting operator for the other participants and thus could influence the scenario in slightly different ways. All of the operators inevitably work slightly differently even though they are trained to follow standard operating procedures.

One of the main constraints on the design of the user test was that there was only one location where such a user test could be performed. This was in the actual operations room which had the appropriate workstations, geographical display, voice communications and environmental features (for example, lighting). All of these HMI elements were essential to collect the appropriate types of data. For example, if the seating or lighting was not appropriate (for example, uncomfortable or dim lighting) this would only be identified by prolonged use in a realistic context. This would also allow the HF team to observe how this impacted the users' performance.

In order to conduct the dry run and the final user tests a total of seventy-two RAF operators were required. This included RAF Subject Matter Experts (SMEs), supporting operators, dry run participants (reused in the final user tests as cover for illnesses) and the final user test participants. Certain supporting operators had to play many roles to help reduce the number required. For example, one operator played some of the Western European air traffic control centres and was able to put on different accents, thus adding a further level of realism.

Once the location was agreed, the master project plan was inspected to determine when the operations room would be ready, i.e. workstations built, appropriate software builds installed and with environmental functions such as heating. Two slots were created in the plan to allow for the user tests. The first slot (eight days) was identified for the dry run of the user tests and the second slot (twelve days) for the final user tests. These two slots were positioned five weeks apart allowing for defects to be fixed and rework of the user test materials. In order to gather an appropriate amount of the types of data required, sixteen of these large-scale multi-user tests were conducted in the second slot.

As the user interface development had not been completed (i.e. less than 95%) there were some known defects in the system. Each of these had to be identified and a rationale provided on the potential impact that it would cause to the users during the test. It is easier to argue that a defect was the cause of a problem if it has already been identified as a potential problem rather than retrospectively looking for reasons when a problem surfaces.

There was an additional logistical challenge to find sufficient numbers of RAF operators with the required skills just prior to the Gulf War and Firefighters' strike (2003) (the RAF are used as a backup for the UK Firefighters). Due to the RAF's conflicting requirements on external and internal activities, there were only just enough operators remaining to call upon. Contingency plans were put in place to recall ex-RAF personnel to play supporting roles.

Once the personnel had been identified, the RAF had to move these operators from wherever they were in the world to the RAF base where the user tests were going to be conducted. Accommodation had to be found and booked, and coaches had to be arranged to collect and return personnel at the appropriate times.

Preparation

Once a high level user test plan was agreed, preparation of the user tests could proceed. The preparation included creating realistic scenarios, questionnaires, checklists and training materials as well as developing new logging software and providing the HF observers with domain training.

As previously mentioned the content of the scenarios was driven by the roles, their tasks, various requirements, hazards and contractual deviations. Four two-hour scenarios were produced that allowed for this type of data be collected. The purpose of having four scenarios was to allow each participant to practise with the first three scenarios before being formally assessed in the fourth scenario. These practise scenarios allowed the users to gain experience and confidence using the HMI, and practise working in the new environment with their colleagues. Even though they were practise scenarios they were observed as if it was a formal user test. This was for two main reasons:

1. To reduce the impact that the observers would have on the participants in the final scenario. Ordinarily, it is not advisable to subject user test participants to this level of scrutiny but these RAF operators were experienced in being observed as this is part of their training and certification.
2. To allow the observers to practise collecting appropriate data. This data also showed how well the participants had improved in their performance with each additional scenario.

The four scenarios required two types of information: background simulation files and a Master Events List (MEL).

Background simulation files

These files contained information such as the details of 1500 separate aircraft (there were at least fifty details per aircraft to specify, for example, speed, height, flight plans, call signs, etc), radars (for example, types of output), areas (for example, danger areas), and supporting data (for example, weather, airfields, missions) for sixty totes (i.e. windows of information). All this information had to be created and populated into files and databases; this was an enormous task.

The RAF SMEs were requested to assist in the creation of this data to ensure that it was as realistic as possible. Figure 1 shows a sample of the types of information required to be displayed on the geographical display.

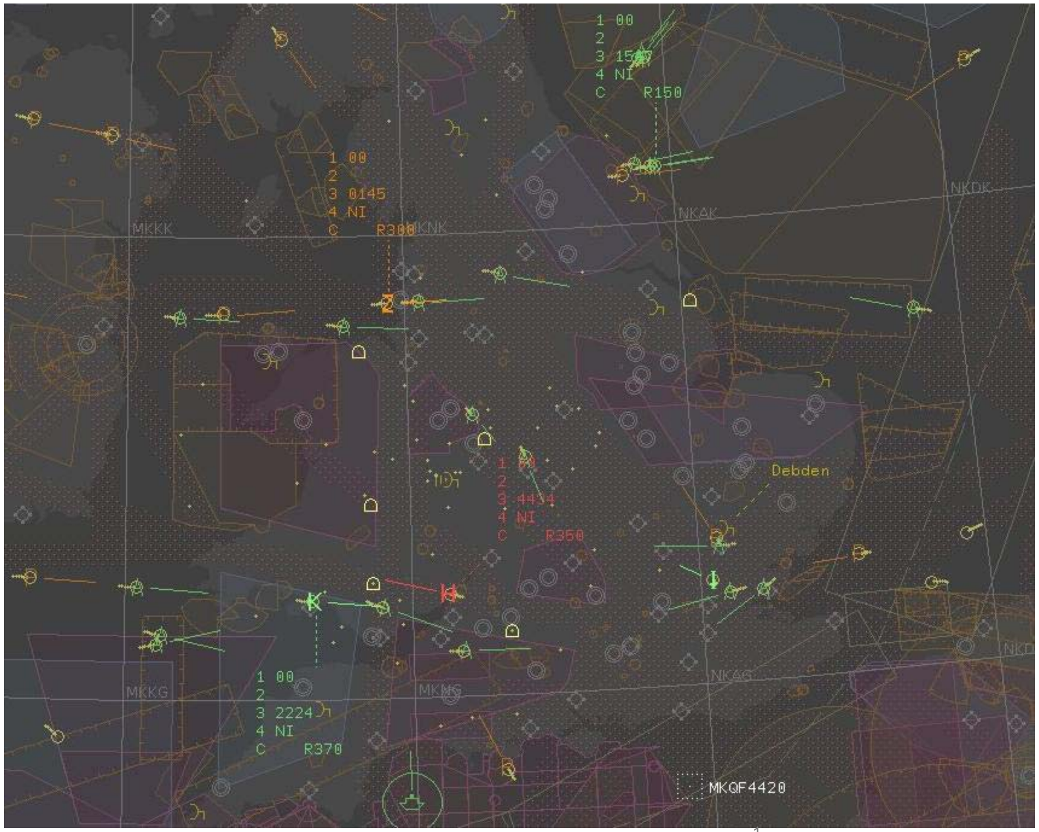

Figure 1. Sample of the geographical display[1]

Master events list (MEL)

The RAF produced an MEL to ensure that every planned event was carried out by the supporting operators at the correct time. Each of these scheduled events was designed to place a particular amount of workload on the participants. The MEL specified such things as: when a certain supporting operator would call a participant on the telephone; what they would say; when an emergency mayday was called; when a radar would be disconnected; and when the participant's console would crash (the plug was pulled from the back of the console in order to test procedures for the coping with such an eventuality).

[1] The colours and sizes of the symbols are not accurate in this format

There were two types of workload scripted into the scenarios: system-imposed workload, such as the amount of air traffic; and personnel-imposed workload, such as taking phone calls from other parties. It was the personnel-imposed workload that drove the need for a large number of personnel to support the test. There were hundreds of scripted events that would occur during each of the scenarios, and they had to be repeatable as several runs of the same user tests were required.

The communications systems had to be modified to ensure that when a participant contacted, for example, London Air Traffic Control, the call was routed to the appropriate supporting operator, and that the supporting operator knew which participant was calling and thus knew what to say.

There were many different parties employed to observe the user tests for different reasons. One of the challenges was being able to facilitate so many observers in such a small space, and to not interfere with the participants. [Figure 2 shows where each participant was positioned.] Each participant was observed by (at least):

- RAF SME – observing participants operational performance,
- IBM HF specialist – observing for specific types of data,
- HEART data collector – collecting metrics related to common observable tasks,
- MoD Safety representative – observing for safety issues,
- MoD HF specialist – ensuring the user tests were fair and accurate,
- RAF officer – ensured that the MEL was followed.

Observer etiquette instructions were created that explained what observers, supporting operators and SMEs were allowed to do and not do during the user tests. For example, they were told not to interfere with participants (for example, by providing advice) at any time during the user tests. This was necessary as many of the observers had not seen or participated in a user test of this nature and did not know how easily the results from a user test can be skewed by inadvertent actions.

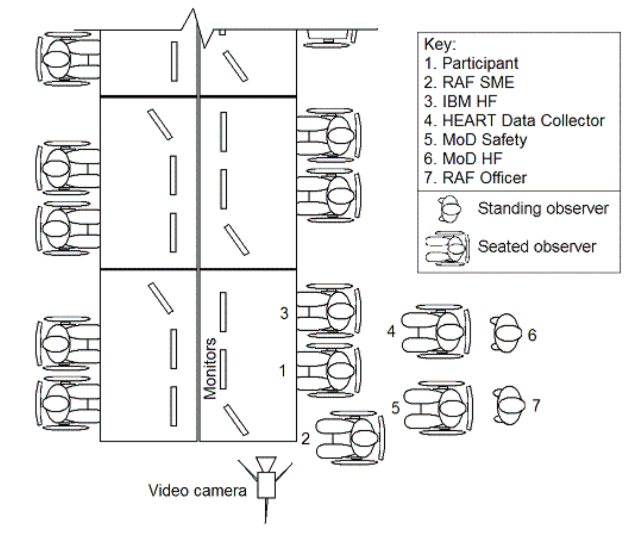

Figure 2. Observers' positions

A number of documents were produced to facilitate the smooth running of the training and user tests, such as timetables and seating plans. This was particularly important as there were a few peculiarities to consider. For example, because the operations room was in an underground nuclear-bomb-proof bunker there were only a few toilets available. One problem that was anticipated was the availability of these facilities if over fifty RAF operators were all to have a break at the same time. To overcome this problem the breaks were staggered so that operators could use the local facilities in turn. It was important to prevent the personnel from needing to use the alternative facilities as they were located above ground, thirty minutes away (due to travel and security procedures).

Information packs were also produced for all seventy-two operators. Each pack contained a user manual, timetable, seating plan, the roles that they were to play, observation etiquette instructions, observation forms (for example, for defects), and a pen. In addition to this paperwork the supporting operators and SMEs were supplied with the scenario-specific instructions (parts of the MEL) at the beginning of each scenario.

All of this planning had to be agreed and signed off by the MoD Integrated Project Team (IPT) before the test could proceed. The documentation for these large-scale multi-user user tests ran to hundreds of pages and took four people approximately six months to collate and prepare.

Training

It was necessary for the operators to be trained on the entire ADCC system in order for them to perform their tasks in the user tests. However, the only location that had all of this equipment installed and working at that time was the operations room. The biggest challenge was that the layout of the room had been designed optimally for ADCC operations and not for teaching. Workstations were not all facing the same way and monitors obscured eye contact.

Once the location had been established, the training materials were designed accordingly. The intention was to modify the training team's train-the-trainers (TTT) materials. The training team had planned to deliver the training over three days, but because of the time constraints and of the number of operators the team had to train for the user tests, there could only be one day's training for each operator. However, in addition to this one day's training the operators would also have the three two-hour practise scenarios to get up to speed. Based on the experience gained in previous user tests on this programme the team had confidence that this amount of time would be sufficient. However, if it was not enough and the users performed poorly it could be argued that the participants would have performed better had they had the full three days' training.

The training materials had to be significantly modified for the user test. The TTT materials did not contain standard operating procedures and contextual information which were important to the operators. This is something the RAF trainers were scheduled to add at a later date. Also, to reduce the content of the materials, the operators would only be trained on the functions that they needed to know for the scenarios. Further, in cases where there was more than one method of performing a task, the operators would be shown only the most efficient method. They were also asked not to attempt tasks they had not been instructed to perform or interact with user interface dialogs they had not been trained to use.

A plan was devised to train the operators in two groups. The first group was required for the dry run user tests and was by far the larger group. This consisted of SMEs, supporting operators and dry run participants. The SMEs and supporting operators would then be re-used in the final user tests. The second group was trained five weeks later and mainly consisted of new participants. The SMEs and supporting roles were provided with refresher training and education on any new or enhanced features that had been implemented since the dry run.

As previously explained, the operations room was not ideal for teaching and was too large to enable one person to present and be seen and heard by all the trainees. Dividing the room into four or five independent training areas was possible, but as the operators all had to interact with the same system it would prove very difficult. For example, if each group was allowed to use the system in an uncoordinated manner they would directly affect the other groups' usage. The final training plan required a master instructor to narrate the presentation (via microphone and speakers) and group instructors to change the slides at appropriate times. The trainees were divided into four groups (noted as A to D in Figure 3), and each group had two projectors, two projector screens, and one group instructor. One of the projectors permanently displayed the geographical display (as shown in figure 1) and the other displayed one of three screens (using a three-way switch): the totes (the windows of information), the presentation slides and the communication (radio and telephone) user interface. This configuration was necessary to allow for the correct combination of screens to be displayed during the training.

It was important that the trainees had as much time using the system as possible, so hands-on exercises were inserted throughout the presentation and the group instructors provided local support. The master instructor was required to coordinate the activities and work with the other group instructors to monitor the timings of each section in the presentation.

The training presentation was designed to cover the most common aspects first. The roles that carried out relatively few and common tasks were allowed to finish early and practise with their group instructor. Therefore, the seating plan tried to ensure that the trainees representing the same roles were seated together (for example, group A, B, C or D in figure 3); this organisation allowed groups to practise in isolation and not distract other groups. This also enabled the group instructors to specialise in only the roles they were assisting. There were a few trainees who required training on unique aspects in addition to the common aspects. These trainees were taught in small groups by their instructors after they had completed the training for the common aspects.

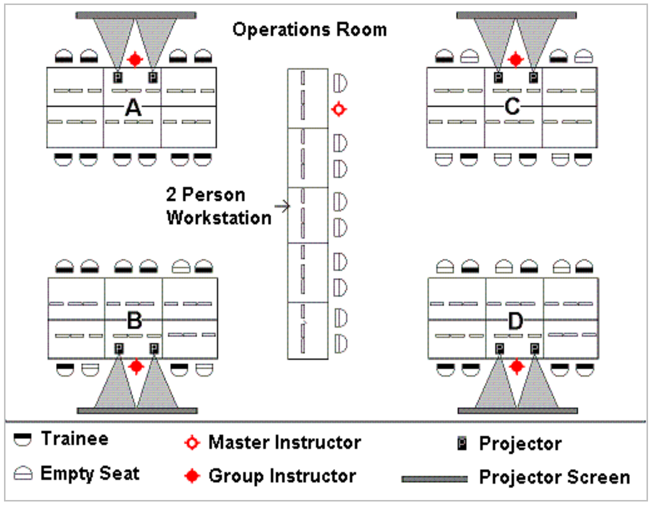

Figure 3. Seating plan for training

A further complication in the training plan was the scheduling of breaks. Due to the large number of trainees, instructors, technical support staff, and the management team it was not possible for them all to go for breaks at the same time. For example, the canteen could not cope with fifty people arriving together as the RAF site was being refurbished and was not fully staffed. The solution to this problem was to stagger the breaks for the different groups and provide many personnel with packed lunches.

On completion of each of the three practise scenarios, the participants were allowed to have refresher training on any item they required. This training was recorded and passed on to all the other participants undertaking the same role in subsequent user tests. This process was used to minimise the difference in the amount of training each participant received, thus helping improve the reliability of correlated data.

It was also important to control how much training and advice was passed on to the participants by their peers. The user test team (IBM and RAF personnel) were requested not to give guidance to the participants and to point the participants to the HF team for help. Where possible the HF team would point to the appropriate page in the user manual or if necessary provide one-on-one tuition. Again any extra advice was recorded and similar advice passed to the other participants. If a participant forgot how to do something in the user test then they were allowed to speak with a peer in their team. This of course would be recorded on tape and used to examine, amongst other things, whether the training materials could be improved.

Conduct

The process for conducting the user tests was to train one group of operators on the first day, let them practise (using the three two-hour practise scenarios) on the second day, and then let them perform the final scenario (the important one) on the third day. This process was repeated twice in the dry-run and four times in the final user test phase. Figure 4 shows where the operators were positioned in the operations room during the user tests.

At the beginning of every user test the HF team set up the workstations according to the recommended layout (for example, the monitor was placed 600mm from the front of the desk). If the participant chose to adjust the workstation (for example, push the monitor backwards to 900mm away from the front of the desk) then this was noted. If the participant later made a comment that the text was too small or the SME reported that they had worked unsafely then the team would be able to use this information as mitigation.

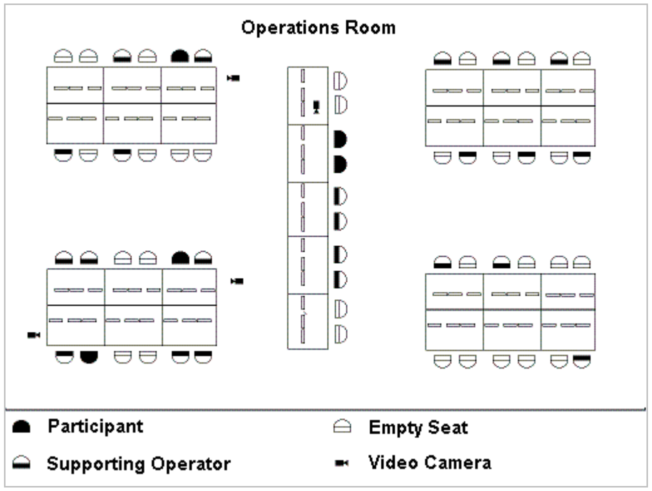

Figure 4. Layout of the operations room for the user tests

The SMEs were requested to observe the participants' ability to work safely, obtain and maintain situational awareness and work at an appropriate work rate. Although this might seem very subjective, this is what the SMEs are accustomed to doing in the normal operations. Each operator has to have a license to use the ADCC system, and one of the elements in obtaining a license is a subjective assessment by an SME. Therefore, the team were confident that an SME's opinion of a participant's performance would be accurate. It should be noted that the SMEs were given training on how to use the HMI to provide more awareness for them on how well the participant was using the HMI. For example, an SME might have observed a participant working at a lower work rate than expected due to inappropriate use of the HMI (e.g. not using fast keys).

The HF personnel collected data in relation to defects, known hazards, system and user requirements, and usability issues. They used a logging tool specifically created by IBM for this task. To reduce the time it takes to log findings during a test a hierarchy of short codes was created. For example, for the alert mechanism there were codes related to items such as distraction (flashing), not noticing alerts, reacting in an appropriate time, and accidental deletion. These codes facilitated the rapid entry and correlation of findings. The HF team practised with the logging tool until they could use it quickly and remember the codes.

Video cameras (as shown in figure 4) were used to record the operators' working envelope, looking for things such as inappropriate posture and environmental effects. The intention was to only refer to these tapes if an operator made a complaint in respect to these aspects. If the HF team had relied on using only video for collecting data (i.e. had not manually collected data at the time) then the schedule would have had to have been increased by hundreds of hours for the examination of videos.

On completion of each scenario the participants were interviewed. First they had a debriefing by the SME who had been observing them. By careful questioning the SME could determine whether the participant had or had not seen particular features. It was important that the SME did not criticise the participant and bias any comments the participant might make during the HF specialist interview.

The HF questionnaire was structured around the types of data the test was looking for. For example, questions such as 'Did you find anything distracting?' were asked. The HF team was looking for comments on areas such as the environmental noise levels, flashing alerts, and telephones ringing. The questionnaire was created using a mixture of open and closed questions requiring explanations or responses on a Likert scale.

All of the observers were instructed on user test etiquette to minimise disruptions and the contamination of data. However, as one of the observers found out to his embarrassment, the instruction 'Do not fall asleep whilst observing as you may fall into the participant' had not been included. When the participant was asked if he had found anything distracting, he replied "The man falling into me didn't help". As a serious point, it should be noted that the observers are as important in user tests as the participants, and their concentration levels should also be considered.

With any formal user test it is essential that a dry run (or pilot) user test is performed first. On this project it would have been catastrophic had the team not performed this activity. The dry run allowed the team to see whether the scenario performed as expected (for example, whether it was realistic and accurate), that the system performed adequately, that the timetables and seating plans were correct, and that the questionnaires and other paperwork were understandable.

The dry run user test did not go according to plan. This was the first time that the system had been performance loaded, e.g. by having every terminal logging on at the same time. This caused the system to gradually degrade over the first thirty minutes on every attempted run, and so the team never got to see a full scenario in operation in the dry run.

This was a risk the team had to accept for the final user tests five weeks later. As a small mitigation to this risk we were confident that the scenarios would be appropriate as the RAF had assisted in their creation and they had many years' experience in organising exercises of this nature.

A number of defects were observed in the dry run and were recorded and prioritised for rectification prior to the final user tests. Those defects that could not be rectified before the final user tests were recorded and added to the training materials to notify the operators. A rationale was also provided that explained what impact each defect could cause to the participants' performance.

Results

The main objective of the results section is to describe what the elicited data was used for, rather than relaying the qualitative and quantitative results.

The team was able to obtain a very large amount of useful data from this activity. The data was provided by the participants (via questionnaires), the supporting operators (via observation forms), the SMEs (via their reports) and the HF team (via the logging tool and HEART forms).

The data was used to help show that:

- Many of the user, system, and safety requirements had been met,
- Potential hazards had not caused confusion,
- The novel designs and deviations from contracted standards had not caused confusion,
- The vast majority of the operators could work safely, maintain effective situational awareness and work at an appropriate work rate, and
- The training was effective.

This data was used as evidence to help form arguments in what eventually became an acceptable safety case. However, not everything about the HMI was positive. A few improvements were called for and implemented later; for example, one type of aircraft symbol was not bright enough and hampered the operator from finding the object quickly when visually scanning the display. Also, it was not possible to gather data for some of the requirements or hazards. For example, there was a function that operators could use if they wanted to adjust their display. The premise was that it could confuse the operators. However, none of the operators chose to use this function and thus no data could be collected on this hazard.

The video was used to capture operator positioning and posture in case they made negative comments on the workstation or environmental factors. The captured video was referred to on two occasions. One of these was when a participant complained of back ache. On examination of the video it was shown that the participant was continuously sitting incorrectly and was not using the HMI functions appropriately. No more than thirty minutes of the 160 hours of recorded video were ever required to be replayed.

Lessons learned

There were many lessons learnt on this project: in fact, too many to list. However, here are a few of the lessons that should be taken on board if this type of large-scale multi-user user test is performed again.

The training presentation could have been better coordinated. As previously mentioned, the master instructor narrated the common sections of the presentation and group instructors controlled the presentation slides and gave demonstrations. However, it would have been easier if the master instructor had control of all of the common slides rather than relying on his colleagues to stay coordinated. Ideally, if there is money in the budget then plan for a separate purpose-built training room where all of the necessary systems are available.

When test plans are created they are supposed to be made repeatable in case they have to be run again. It quickly became apparent that these multi-user user tests were not going to be repeatable for two main reasons.

- There were five participants working simultaneously, each of whom works in a slightly different way. Even though the operators are all trained the same and follow the same standard operating procedures, no two operators will do the same actions at the same time. Therefore, when five participants are working together it is virtually impossible to expect the sequence of actions to be identical on each run of the user test.
- Due to the size and complexity of the user test the team did not have the time or resources to conduct the test again. Further it would have been very difficult to find more participants who had not been exposed to the system in some capacity.

When seventy-two military personnel are gathered to one location in a secluded part of the country they will want to do some serious socialising. Many of them will have been separated from their friends for many months, and they will have a lot of catching up to do. What the team did not want was seventy-two operators turning up with hangovers. To their credit, the team showed their true professionalism and were fully fit for duty throughout the test period.

Conclusion

Overall this large scale user test proved well worth the effort and expense of running it. The whole user test programme (including planning, preparing, training, conducting and analysis) took IBM the equivalent of two person-years in effort. In addition to this, the RAF contributed at least this amount again. However, the benefits were enormous. IBM captured a mass of data that was used to satisfy requirements and mitigate hazards, as well as demonstrate that the HMI would allow the operators to work safely, with good situational awareness and at an appropriate work rate. Of the twenty participants that were used in the assessment of the HMI, eighteen of them worked appropriately according to the SMEs. This was after just one day's training, i.e. a third of the training they were scheduled to get. When these training times are compared with the old system, the operators would have required three weeks' training and a period of practise before they could have achieved this level.

In addition to achieving the primary objective of collecting data for the safety case, the RAF team was extremely happy with the findings and became very enthusiastic about the new system. Had the data been gathered using small user tests and mathematical models (as was originally planned), the RAF would not have been as confident or as enthusiastic. The feel-good factor cannot be underestimated. If the customer and users are happy, and they have seen the system successfully used in a realistic operational setting before they accept it, then this is bound to improve the working relationship.

Practitioners' takeaways

Although the client may specify or mandate a UCD approach (for example, ISO 13407 [3] or Defence Standard 00-25 [2]) to be followed, they may not realise how expensive it is to support this approach. The reduction in whole life cycle costs of using a UCD approach are well documented, but the costs of supporting the development from the customer's perspective are sometimes less explicit. Therefore, it is worthwhile discussing this with the client early in the programme, and if necessary, updating the risk register accordingly.

The summative user tests are an excellent mechanism for collecting data and evidence for a safety case. However, to ensure that appropriate safety-related data can be collected on a safety-related system the participants must be placed in an appropriate working environment under a realistic amount of workload.

Do not rely solely on a summative user test to produce all the necessary operational safety-related data for a safety case. A good design process that includes activities such as formative user tests and heuristic evaluations, as well as the demonstration of effective training materials can also produce an enormous amount of safety-related data [4].

The production of the training materials must be aligned with the user tests. The Training team must be notified as early as possible on the required roles and tasks for each user test. This will ensure that the appropriate training materials are available at the right time. Further, Train The Trainer materials are not usually appropriate for a user test. They are often lacking in context and standard operating procedures as these are usually added later by the client's Trainers once they have been trained. On this project, it took the HF team 60 days to convert the materials into the correct format. Therefore, design the training materials so they can be used effectively for both purposes (i.e. TTT and user tests).

Limit the number of observers to only those that are absolutely necessary. When these tests were carried out, the project was nearing completion and everybody wanted to see the system in action. This was a cause of concern for the HF team as the more people that are around the more chance of distraction and skewing of results, albeit accidentally. Therefore, create and distribute observation etiquette instructions to everybody entering the user test area. Even attach copies of the instructions to the test room entrance, in the canteen and in congregation areas.

Acknowledgements

I would like to thank the team that helped plan and execute these user tests. Particularly Dave Roberts, Emma Reeves, Mark Tibbits, Mark Farmer, Jo Stray, Paul Hubbard and Paul Pople from IBM; the RAF team; and Mark Upton from Electronic Data Systems (EDS).

References

[1] Defence Standard 00-58 (2000), Hazop Studies on Systems Containing Programmable Electronics
[2] Ministry of Defence Standard 00-25 (2004), Human Factors for Designers of Equipment.
[3] ISO 13407:1999 Human-centred design processes for interactive systems.
[4] Hey, E (2004). Human Factors Contribution to a Safety Case. BCS HCI 2004.

The Author

Elliott Hey is a Senior User Experience Consultant in IBM Business Consulting Services. His main experience is in the design of user interfaces for military and office applications, web sites, PDAs and mobile phones, for clients in Banking, Insurance, Retail, and the Public Sector.