THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY, *14*(4), 395–420
Copyright © 2004, Lawrence Erlbaum Associates, Inc.

Applicability of Usability Evaluation Techniques to Aviation Systems

Michael Clamann
Booz Allen Hamilton
Mclean, Virginia

David B. Kaber
Department of Industrial Engineering
North Carolina State University

Research in the field of human–computer interaction (HCI) has shown that early usability evaluation of human interfaces can reduce operator errors by optimizing functions for a specific population. HCI research has produced many methods for evaluating usability, which have proven effective in developing highly complex computer systems. Given the importance of the human in the loop in aviation systems, it is possible that advanced commercial cockpit and air traffic control systems may benefit from systematic application of usability research. This article identifies the special requirements of the aviation domain that will affect a usability evaluation and the characteristics of evaluation methods that may make them effective in this context. Recommendations are made of usability evaluation techniques, or combinations of techniques, most appropriate for evaluating complex systems in aviation technology.

The role of any system interface is to provide a dialog between the operator and the device. This dialog directs the actions of the operator to the device in the form of controls and from device to operator by converting raw data to useful information. If a system is *useable*, the dialog is intuitive and natural and allows the user to work in harmony with the system. There are numerous ways to evaluate and enhance the usability of a system. Some are empirical; others require the support of a usability expert, and some are best performed by groups. Ideally, these methods are applied dur-

Requests for reprints should be sent to David B. Kaber, Department of Industrial Engineering, North Carolina State University, 2401 Stinson Drive, 328 Riddick Labs, Raleigh, NC 27695–7906. E-mail: dbkaber@eos.ncsu.edu

ing the development of the system to ensure a usable product. Conducting a usability evaluation for any system poses unique challenges.

In the complex, dynamic, tightly regulated environment of aviation, the challenge of performing a usability evaluation expands considerably in comparison to evaluation of traditional human–computer interaction (HCI) applications. Dramatic advances in technology and pressures to increase the number of commercial flights are driving factors in the evolution of interface controls on the flight deck and in air traffic control (ATC) workstations. As technological functions are added to already highly complex interfaces, which may increase operator workload, it is important that the user be considered in the design of the system. Furthermore, as the consequences of an error in aircraft piloting or ATC may be catastrophic, the importance of the human operator in the decision process is even more evident. For example, in December 1995, the pilots of an American Airlines Boeing 757 bound for Cali, Colombia, altered their landing approach from one that required them to pass over the airport and reverse course to one that involved a shorter, straight-on approach. In the process of interacting with automated systems used to navigate the aircraft's flight path, the pilots became so distracted by display interfaces and functions that they failed to notice they were heading directly toward El Deluvio, a mountain peak 10 miles east of the airport in the San Jose mountain range. At 9:38 p.m., the 757 struck the mountain at an elevation of 8,900 ft, killing all but 4 of the 163 passengers on board (Aeronautica Civil, 1996; Strauch, 1997).

In an industry where the human operator plays such a crucial role in the system, the effectiveness of systems usability evaluation is of significant importance. Current avionics manufacturers realize the importance of addressing usability issues in the systems design process, and they have become more sophisticated in their capabilities to deal with usability problems within the recent past. These developments are in large part due to federal regulations requiring human factors test plans for the development of new avionics systems. For example, Rockwell Scientific and other aircraft cockpit systems designers and developers are integrating sophisticated human factors evaluation methods, comparable to experimental methods used in human factors lab research, into their design processes to develop new advanced cockpit weather displays (W. E. Kelly, Research Scientist, Rockwell Scientific, personal communication, November 18, 2003). Despite this fact, there is very little organized information on methods that designers and manufacturers like this can use for evaluating the usability of highly complex systems (Abbott, Wise, & Wise, 1999). The purpose of this article is to identify usability evaluation techniques, or combinations of techniques, most appropriate for evaluating the complex systems used in aviation technology.

In this article we present an overview of two major types of aviation systems: those designed for aircraft cockpits and ATC. This includes a discussion of the environments and usability issues common in existing aviation systems. Potential problems in the development of new systems using existing evaluation processes are also discussed. We identify the existing requirements of regulators in terms of human factors test plan development for avionics certification and workload (usability) assessment. Beyond this, we review several current usability evaluation testing techniques, including usability inspection methods and usability testing. Some of the more widely used methods are presented, along with their advantages and disadvantages, to provide a basis for comparisons of the different techniques. In a final section, the information on aviation technology and usability evaluation techniques is integrated to determine which of the techniques best fits the context of current aviation systems development. We conclude with a summary of how manufacturers can use the recommendations on formal usability evaluation methods to develop effective and cost effective human factors test plans for avionics systems design and development.

THE AVIATION SYSTEMS ENVIRONMENT

> On April 14, 1990, an Indian Airlines Airbus A320 aircraft crashed just short of the runway at the Bangalore, India airport destroying the aircraft and killing 90 people on board. The investigators determined that the probable cause of the accident was the failure of the pilots to realize the gravity of the situation and immediately apply thrust. The pilots spent the final seconds of the flight trying to understand why the autoflight system was in idle/open descent mode rather than taking appropriate action to avoid impact with the ground. (p. 863)

As demonstrated in this example, described by Funk, Suroteguh, Wilson, and Lyall (1998), aviation systems are complex, dynamic environments where the consequences of errors can be disastrous. Interface design problems include the amount of information being presented and integration of interfaces. In aircraft, the number of stimuli in the cockpit, such as the densely packed controls and the output produced by warning indicators, status displays, flight path displays, ATC data links, weather information, navigation information, and communications, combined with a constant requirement for situation awareness (SA), contributes to the possibility of excessive mental workload for the pilot (Stokes & Wickens, 1988). When new technology is added to a flight deck system that actually reduces pilot workload, that reduction is often exploited by the introduction of another new system (Billings, 1997; Smith, 1999). Billings said the number and choice of displays present in the advanced commercial aircraft cockpit may only be restricted by the amount of available space, and the expansion of automation in the cockpit seems to be only limited by the imagination of designers with little consideration for pilot performance consequences.

Air traffic controllers also have their own specialized problems. Although in most conditions pilots receive instant feedback from their environment, air traffic controllers work through a system representation and experience a delay when observing the effects of their command sequences (Billings, 1997). They also face their own cognitive challenges. To forecast numerous flight paths simultaneously, controllers construct complex three-dimensional mental models or “pictures” from a two-dimensional graphical and text-based representation of current aircraft layout combined with the expectations of their routes. If this model is lost, it must be reconstructed again through available tools, including computer systems, paper flight strips, and memories—a process that can take the controller as much as 15 to 20 min (Billings, 1997; Garland, 1999). The consequences of losing this model are described in a summary of events from Garland.

On February 1, 1991, a USAir Boeing 737–300 collided with a SkyWest Fairchild Metroliner on the runway at Los Angeles International Airport. The 737 was attempting to land while the smaller Metroliner was on the same runway waiting for takeoff clearance. The crash and the resulting fire destroyed both airplanes, and many passengers on both planes were killed or injured. The cause of the accident according to the NTSB was “controller error.” More specifically, it was due to the local controller forgetting she had cleared the Metroliner into position on the same runway that she subsequently cleared for the 737.

Despite accidents like this one, there is still relatively little research on working memory in ATC (Garland, 1999) and the implications of failures on ATC performance. Furthermore, very little research, which was published at the time of the design of the current ATC system, was consulted during the system development process (Benel, 1998).

In the near future, air traffic controllers can expect to see additional automation of their workstations, as the ATC system is updated (Hopkin, 1998). Many of the design decisions and methodologies for how to deliver these updates will come from lessons learned from changes made to other complex systems, such as the aircraft cockpit. If these new systems are designed based on past conventions (technology-centric automation design practices), it will be important to perform exhaustive human factors evaluations during the design process.

In 1996, the Federal Aviation Administration (FAA) Human Factors Study Team Report on the interaction between flight crews and modern cockpits cited that about 70% of the fatal accidents involving new generation aircraft are due to human error potentially induced by the design of cockpit interfaces (Singer, 1999). According to Woods and Sarter (1993), problems that occur in interactions between users and complex systems are due in large part to inadequate feedback on status, behavior, and intentions of the system; lack of visibility that prevents the user from constructing an effective model of the system’s behavior; and designs that replace, rather than reduce, cognitive load. These problems can produce failures in attention allocation that can ultimately contribute to a loss of SA (Sarter, Woods, & Billings, 1997). Advanced ATC and flight deck systems tend to consist of multiple layers of automated technology that separate the operator from the device. Like similar complex systems (e.g., nuclear power plants), a fundamental problem with aviation systems is that the operators do not fully understand what their automated systems are doing, or even why (Billings, 1997). This may also lead to SA errors and overall system failures.

One interface that appears frequently in the human factors literature, when examining weaknesses in the usability of cockpits, is the Flight Management System (FMS). The FMS has been shown to be one of the weakest links in high workload, dynamic operational situations (Singer, 1999). This system was originally designed to optimize flight paths, but over its 20-year history the features of the FMS have grown substantially (calculations for position, speed, and wind; management of navigation data sources; trajectory calculation; flight path predictions; remaining fuel calculations; flight time calculations; error determination; steering and control command generation; map computations, etc.; see Billings, 1997). Despite the complexity of the information, the interface itself has remained relatively static, resulting in a device that alone requires about 1,200 hr of experience for a pilot to achieve a real understanding (Riley, 2001). In a survey conducted by Sarter and Woods (1992), 55% of pilots with over 1 year of experience reported unexpected behavior from the FMS and 20% did not understand all the modes and features. Even though pilots can make the FMS work, they are often not able to explain why.

One of the main problems identified with the FMS is that its command sequences do not integrate with the pilot's procedures (Riley, 2001). They must be considered and executed separately in a distinct series of operations in order to prepare a single flight plan. This can take 10 to 15 min, even for an experienced pilot. In general, the FMS is a good example of a system that has features defined from an engineering perspective rather than a user's perspective; that is, all of the required features are available, but they can be difficult to execute. The design of the FMS is only a single representation of the usability flaws in some aviation systems. Funk et al. (1998) also specifically cited autopilot and auto-thrust systems, controls for the brakes and spoilers, and the electronic horizontal situation indicator as devices playing a role in recent incidents and accidents.

Contemporary ATC workstations have significantly less automation than flight deck systems. However, given the workload requirements experienced by controllers, several opportunities exist for the application of new and emerging technologies. Benel (1998) listed new decision aids to support strategic planning, global positioning system technologies for more accurate aircraft position information, digital data links to the flight deck, voice as an input and output device, routing calculations based on fuel consumption, automated problem detection and resolution, airspace and runway sequencing and spacing (including runway conflict detection), and new methods for data visualization as prospects for enhancing ATC. Any number of new problems could emerge from these new systems if not examined carefully. For example, introducing a data link to automate information transfer between pilots and controllers may offload some controller workload, but it may also remove important nonverbal cues from the transmission that communicate urgency in a situation (Benel, 1998).

Lack of standardization of aviation system interfaces also contributes to errors (Singer, 1999). Interfaces can vary from one environment to another, making it difficult for pilots and air traffic controllers to transfer their skills to different cockpits or ATC sectors or stations. For example, different versions of the FMS software may exist across aircraft of one type or across types of vehicles, presenting a major challenge for pilots in terms of adapting to different interface states (Kaber, Riley, & Tan, 2002). In addition, the automation in many aviation systems does not necessarily transfer to a variety of types of operators and skill levels working under varied conditions. This lack of automation accomodation of all operators may be attributed to the fact that systems are generally evaluated by experts with several years' experience, so the needs of the less experienced operator may not be considered. With such high consequences for errors in system operation, combined with the ultimate responsibility being given to the operator, it is imperative that new systems and new components for existing systems be evaluated not only for their functions but also for their relative usability.

The Design Process

According to Sexton (1988), until the last 30 years, the modernization of aviation systems has often been a process in which controls and displays are updated piecemeal for individual systems. More displays are added to cockpits without re-engineering the control array and are fit where there is still available space. Over time, this has resulted in a complex array of knobs, switches, and displays that does not necessarily integrate intuitively. Until the 1970s, flight deck technology did not vary dramatically from that of the 1930s.

Due to the capabilities of new electronic technologies to integrate these displays and controls ("glass" cockpit technologies), a rapid increase in changes to cockpit controls and displays began in the mid to late 1970s. Despite this sudden increase in technological advances, the design and implementation methodologies employed for introducing new components and retrofitting systems by many aircraft manufacturers during this period were similar to those used in the 1960s (Sexton, 1988). Complex integrated displays were designed using methodologies developed for analog gauges. These methods produced cockpit designs like that found in the Boeing 757 and 767 that use six, integrated, cathode ray tube displays instead of the clusters of single instruments of older models, but, according to Sexton (1988), they did not demonstrate any substantial reduction in pilot workload.

The same development environment also produced systems such as the FMS. Although the FMS usability problems just identified have existed for some time and still represent concerns for crew and human factors and safety professionals working in the aviation industry, several manufacturers and the FAA have recognized the importance of dealing with such usability issues, particularly as a means for managing crew workload and are currently conducting research and developing enhanced design processes for future cockpit development (e.g., Feary, Sherry, Polson, & Fennel, 2003).

ATC has experienced a similar design evolution. As described by Benel (1998), until the 1970s, enhancements to the ATC system took the form of a series of upgrades, including radar, flight strip printing, and automated processing of flight data. A traditional linear development cycle was used to design these legacy systems, which addressed the designers' requirements to produce a system efficiently rather than addressing the needs of the users to have a useable system. Task analyses were based on user statements that the developers then used for the design with only limited ongoing feedback from operators. Furthermore, in the early decisions to allocate functions between the operator and system, human functions were not chosen because they were matched to human abilities but because they were infeasible for the existing technology (Hopkin, 1988).

With many traditional approaches to system design, one important factor that is frequently left out of the process is the early involvement of the end user (pilot or air traffic controller) or a usability expert in the design process (Stein, 2000). This is obviously not the case with all aviation systems manufacturers. Early involvement of users is crucial to the design process because the cost of making revisions increases dramatically as the development of the system proceeds. If domain experts find usability problems with the system after functional prototypes have been completed, it is much less likely that problems will be corrected, and the burden will be added to the users during the training process (Sarter et al., 1997; Stein, 2000).

Several contemporary aviation system designers have realized the need to consider the user in the design process and have proposed participatory design methodologies. Advances in new technologies encourage participatory design through the availability of high-fidelity prototyping. Several methodologies for participatory design of aviation systems, including complete feedback loops and testing, do exist (Benel, 1998; Sexton, 1988; Williges, Williges, & Fainter, 1988); however, they are not applied consistently across manufacturers and domains. For example, beginning in the 1980s, Lockheed-Georgia Company made use of its own participatory design process for new systems, retrofits, and simulators of modern aircraft. The design process, described by Sexton (1988), follows this methodology:

1. Design team selection: Include a pilot familiar with the proposed system, human factors engineers, mechanical design engineers, and avionics engineers. Each member of the design team must be involved full time with the project.
2. Mission analysis: Obtain, forecast, and determine information on the proposed aircraft with respect to user needs, operating environment, procedures, and

new technologies. Engineers from each area present forecasts for technology levels for their area of expertise. This information is then used in preparation of documented mission scenarios, which detail the environments and situations in which the new equipment will be used.

3. Design: Team members conceptualize the aircraft design by using the mission scenarios as drivers. The forecasts are also used as design considerations. By involving all members of the team, technological trade-offs can be agreed on. For example, the human factors engineer can ensure that the design is not developed so as to benefit hardware efficiency at the cost of the interface. A full-scale mock-up is developed in this stage so the team can easily reconfigure it.

4. Test: The configuration is tested with real pilots using scripts that expose subjects to prescribed situations. The tested design is then integrated in a flight simulator.

The availability of powerful computer hardware and rapid prototyping tools has simplified the process of designing mock-ups. Foam and paper interfaces and scripts have been replaced with graphical simulations and virtual reality scenarios during the design and test phases (Sarter & Woods, 1992).

Despite the existence of such contemporary design processes that prescribe the input of experts, such as the one used by Lockheed, and the usefulness of high-fidelity prototypes, in general it is not apparent from the designs of contemporary cockpit and ATC interfaces, and the number of aviation mishaps and accidents attributed to pilot error, that formal usability methods are very widely used throughout industry. On the basis of published research, some other manufacturers do use rapid prototyping of user interfaces; walk-through evaluations with pilots; and more formal user evaluations, including human factors experiments on control configuration, panel layout, and menu design (e.g., Honeywell; Riley, 2001). NASA recently formulated a rapid usability evaluation approach for Boeing's future cockpit interface development (Feary et al., 2003). However, in those companies that do conduct usability testing, unfortunately, the process may be seen as one of the reasons for slow adoption of new technologies in the flight deck because of the time required for a thorough assessment of an interface. This further supports the need for effective and economical approaches for application in the aviation context.

Certification: Requirements of Regulators

Because of the public interest in civil aviation, the current certification requirements for aviation systems are fairly strict. According to Singer (1999), most countries follow the rules established by the FAA and the Joint Airworthiness Authorities. Both the airworthiness requirements (regulating design) and the operational requirements (defining proper use) must be fulfilled to legally operate a commercial aircraft. The documented set of airworthiness requirements is called *Federal Aviation Requirements, Part 25* (U.S. Department of Transportation/FAA, 1995), and aircraft certification is a well-structured process summarized by the following steps (Singer, 1999):

1. A *test description* is provided to the certification authority. It defines new functions, lists the design requirements they fulfill, and lays out how compliance will be shown (e.g., simulation, flight test, pilot evaluation).
2. Iterative testing is performed in the *development testing* phase. Pilots and engineers test functions, system logic, and the display.
3. When testing is complete, the designers must submit a *certification test plan* describing the evaluation process.
4. Test pilots evaluate the *fit, form,* and *function* of the system, including the ergonomics of the interface and overall layout. This process is very subjective.
5. A *certification test report* is then prepared based on the certification test plan, showing how each item has been successfully tested. Compliance can be shown either by testing in a flight simulator or by testing individual components in a lab.
6. *Crew evaluation certification reports* are then prepared for qualitative requirements not covered by the engineers.

When all of these documents are approved, the system is declared *airworthy.* One of the main reasons for the design of the current certification process is that the FAA is attempting to lead manufacturers toward more rigorous "usability" assessments of cockpit interface designs. There is some evidence of this in current certification process documents, which identify common usability issues (accessibility of controls, ease of use, salience of displays) as factors in pilot workload.

Although the airworthiness and operational requirements cover a broad scope of topics, the verbiage of the paragraphs defining human factors requirements is very general. In many cases, these paragraphs were written for cockpits consisting of levers and buttons, not dynamic electronic screens. Furthermore, the requirements primarily address concentration and fatigue issues but do not present tolerable levels for either, and they do not address the higher levels of pilot information processing that may be compromised by unusable interface design. Overall, the requirements are general and do not provide much assistance to making engineering decisions for designing aviation interfaces and do not establish quantitative levels for compliance.

Singer (1999) said, *workload, fatigue, ease,* and *simplicity* are meaningless without a frame of reference to define their tolerable levels. Based on the content of the FAA's Advisory Circular (AC) 25.1523–1 (U.S. Department of Transportation/FAA, 1993), the agency considers workload to be defined by the functional requirements of flight, including the number, urgency, and complexity of procedures as well as the duration of mental and physical effort in a normal operation. In addition, the FAA considers (usability) issues of accessibility, and ease of use, of controls to be factors in workload. Specific measures for assessing these facets of cockpit workload are not provided and there are no criteria in terms of the number of tasks or levels of, for example, concentrated mental effort by which to make decisions regarding interface designs toward moderating pilot load. It is possible that there is a lack of definition of threshold values for workload, concentration, fatigue, and so on, by the FAA because there are few, if any, measures of these factors, which are completely theoretically and empirically defensible, for which threshold values can be set.

Although the certification process is defined for systems examined in isolation, it does not provide structure for including the human operator. Currently, certification authorities evaluate flight deck workload in comparative terms based on earlier airplane certifications, a process that is carried out by experienced FAA certification pilots, who consider multiple-flight conditions (Billings, 1997; U.S. Department of Transportation/FAA, 1993). This is the primary approach defined for establishing minimum flight crew requirements (and related human factors issues) as part of the certification process. However, the FAA does permit multiple methods for manufacturer compliance to allow a range of companies and methods to achieve the requirements. For example, analytical methods (e.g., task analyses, timeline evaluations) can be used to assess new devices primarily impacting flight operation procedures, and simulator demonstrations are considered acceptable for showing compliance, provided there is sufficient substantiating data developed on the validity and reliability of the assessment approach. Final decisions on workload implications of new models of equipment are normally based on testing with qualified and trained pilots ("line pilots," who fly the same aircraft, are recommended by the FAA as they can make direct comparison with operational experience), unless traditional testing methods cannot be used with the new design.

Certification requires that a piece of hardware or software perform its intended function. This does not necessarily mean that the equipment is user friendly. Meeting minimum requirements does not ensure a good or even a safe system (Stein, 2000). Even though the requirements for performance of the system itself are well established, the usability evaluation of aviation system interfaces is not well defined, and each vendor is free to adapt an independent philosophy in showing compliance. The requirements do not identify specific usability evaluation methods, metrics, or acceptable and unacceptable levels of performance. Appendix D of AC25.1523–1 does provide some additional detail on what the FAA considers criteria significant for determining crew make-up—for example, ensuring conspicuity of instruments, ensuring adequate instrument feedback to direct pilot behavior, identifying additional duties of crew members removing them from their normal duties, and identification of the level of automation or operator monitoring required to guard against flight control and essential system failures. There is little supplemental guidance on methods to use and specific evaluation criteria to assess workload and usability to prevent human error with new cockpit devices.

In the United States, the FAA also has the responsibility of certifying ATC systems. Stein (2000) provided an overview of the certification process in which experienced ATC specialists and evaluation personnel perform the certification process with little input from human factors experts. If human factors experts are employed, it is usually toward the end of development when they can do little to affect the system. The responsibilities of certification for ATC systems are defined in FAA Order 6000.39, which states that certification only means the required system functions have been accomplished. It does not ensure well-designed or usable equipment or requirements for human factors involvement. This does not mean that human factors input was not used in the design of some systems, but it does show that the burden of defining measures for usability will ultimately fall on those who design the system. To date, as with commercial cockpit systems design, specific usability measures have yet to be determined (Stein, 2000).

The certification process consists of numerous general guidelines for designing technologies that are expected to be valid for all types of aviation systems. The information presented really represents (workload/usability) guidelines rather than regulations, as there are no specific goals defined for new cockpit technologies as part of the "requirements." To be fully valid, the design guidelines should be proposed in the context of the particular system being designed to meet specific requirements, subject to specific constraints (Billings, 1997). Such requirements would be based either on established human factors guidelines or on system guidelines confirmed to be effective for human-centered design. In either case, the constraints for the requirements should be evaluated using validated metrics.

USABILITY EVALUATION TECHNIQUES

The role of a usability evaluation is to confirm that an interactive system is behaving as expected in a way that makes sense to the user. Typically, this means finding usability problems in an interface (Nielsen & Mack, 1994). Current literature on HCI describes many methods for this type of evaluation. In general, techniques can be grouped into two major classes, *usability testing* and *usability inspection* (Virzi, 1997). Testing involves formal experimental evaluation of an interface, whereas inspection covers a variety of informal methods that can also be classified as *walkthrough* and *nonwalkthrough* (Newman & Lamming, 1995). Walkthrough techniques typically analyze sequences of events in an interface, determined by user goals. Examples include the cognitive walkthrough and model-based evaluation. Nonwalkthrough techniques do not require an evaluation sequence based on the task. Review-based evaluation and heuristic analysis are included in this category. Usability tests may include aspects of both walkthrough and nonwalkthrough techniques, depending on the design of the experiment. Here we review the more commonly used techniques.

The cognitive walkthrough is a technique developed to show how easy a system is to learn. It is typically applied to systems with the expectation that users have little or no prior training (Newman & Lamming, 1995); however, the methodology has previously been adapted to the aircraft cockpit (Polson & Smith, 1999) by identifying and attempting to analyze categories of domain activities. The method coordinates analysis by several experts on user's task procedures with a focus on interface usefulness, usability, visibility, and feedback (Dix, Finlay, Abowd, & Beale, 1998; Virzi, 1997). Experts simulate expected behaviors of users, interpret system responses to task performance, and determine if users would choose the correct course of interface actions (Newman & Lamming, 1995). The method also has some variations—for example, the cognitive jogthrough, which streamlines the normally time-consuming process of evaluating the interface by using groups of experts in the process (Rowley & Rhoades, 1992).

Model-based evaluation techniques, such as Goals, Operators, Methods, and Selection rules (GOMS), use models of the user to predict performance on specific tasks using an interface (Virzi, 1997). A GOMS model is the result of a task decomposition fit into a specific structure. The models make assumptions about user information processing and are effective for describing expert performance in routine tasks (Dix et al., 1998). GOMS models have been used successfully to define improvements to task training protocols and interfaces in aviation systems (Irving, Polson, & Irving, 1994).

Heuristic evaluation involves a systematic inspection of an interface, usually early in the design cycle by a small group of evaluators. The evaluators use a general set of design principles (Dix et al., 1998), such as

1. Visibility of system status.
2. Match between system and real world.
3. User control and freedom.
4. Consistency and standards.
5. Error prevention.
6. Recognition rather than recall.
7. Flexibility and efficiency of use.
8. Aesthetic and minimalist design.
9. Help users recognize, diagnose, and recover from errors.
10. Help and documentation.

The evaluators produce lists of usability problems with the interface by comparison to the set of principles or heuristics (Nielsen & Phillips, 1993). This technique was originally designed as a means for nonexperts to perform a usability analysis, but the results can be enhanced by the involvement of a usability expert (Virzi, 1997). It can also be performed at any point within the development life cycle, with or without the use of the actual interface, although the more realistic the information, the more relevant and valuable the results. Heuristic evaluations have been used to produce what are called "cold," "warm," or "hot" estimates of system conformance with usability principles (Nielsen & Phillips, 1993), depending on the level of access the evaluators have to the system. Hot estimates are performed on the implementation, or final version of a system. Warm estimates require a functioning prototype, and cold estimates can be conducted on a written specification. Because usability problems can be examined early in the design cycle, changes to a design may be less costly for a team, as compared to using other methods, like GOMS, that may require an actual interface prototype (Feary et al., 2003).

Of course, usability can also be evaluated through direct observation of users while they interact with the system (Dix et al., 1998). Users may be asked to describe interface actions to nonexperts during task performance (verbal protocols), as in cooperative evaluation, or after a task, as in posttask walkthroughs. protocol analysis can also be used to evaluate behavior by capturing the users' actions through a variety of recording media including audiovisual recordings, automated computer logs, or even paper-and-pencil notes and user diaries (Dix et al., 1998).

Differences Among Inspection Methods

Virzi (1997) categorized inspection methods along three dimensions: the characteristics of the evaluators, the number of evaluators required, and the goals of the inspection. With respect to the first dimension, the level and type of expertise of the judges affects the outcome of the evaluation. Factors include the amount of usability expertise, the degree of domain knowledge, prior design experience, and academic training. In all cases, a greater level of experience, specifically in the particular domain, will result in a greater value of the results of the inspection. Experts with usability experience and domain knowledge may be able to find more usability problems than experts with only usability experience. However, there is a trade-off in that the level of expert training usually has a direct effect on the cost of the evaluation.

Techniques such as GOMS, heuristic evaluation, and cooperative evaluation do not rely on a usability professional with expert-level experience, whereas the cognitive walkthrough and usability tests (discussed next) do require the involvement of one or more experts. Within these methods, especially those requiring less evaluator experience, expertise can also be combined to affect the outcome. In at least one study, researchers found that the involvement of a usability expert greatly improved the results of a usability inspection method designed for nonexperts (Jeffries, Miller, Wharton, & Uyeda, 1991).

The number of evaluators working during a single session is Virzi's (1997) second dimension. Some evaluation techniques require independent work, whereas others use groups. Combining the analyses of several evaluators after completing individual sessions can make the results of a usability inspection far more comprehensive, especially when the level of experience of the evaluators is lower. Nielsen (1994) said the number of errors observed during a heuristic evaluation increases as the number of expert evaluators increases; two evaluators will find 50% of the usability problems, and 3 will find about 60%. This figure starts to level off at about 5 evaluators, so to locate 90% of problems, approximately 15 expert evaluators would be required (Nielsen, 1994). The degree of complexity of a system will also affect the number of evaluators involved in an analysis with more evaluators needed with a more complex interface.

It can also be advantageous to combine the opinions of professionals in separate areas of a development process, such as computer scientists, graphic artists, and engineers, to see how the different aspects of a system interact. In group design reviews, such experts work together to locate usability problems (Virzi, 1997).

The last of Virzi's (1997) dimension is the goal of the inspection. Many types of usability inspections attempt to uncover usability problems, whereas others focus on the extent to which the system supports specific aspects of use. For example, heuristic evaluation and cooperative evaluation are designed specifically to find usability problems. Heuristic evaluation, however, is traditionally not effective at evaluating user performance (Nielsen & Phillips, 1993). Walkthroughs, in contrast, are designed to evaluate ease of learning, whereas GOMS models evaluate performance times and disregard errors and learnability issues. The selection of an inspection method should be determined by the type of usability data required. This point demonstrates the importance of establishing usability goals early in the process, prior to selecting the technique. The measurements that define the system's usability should determine the methods employed and the evaluation team selected (Virzi, 1997). Selecting the method in advance will likely constrain the results.

Despite the advantages of usability inspections, there is no replacement for a formal laboratory test. Because inspections may not include the actual users, there is an important contributor missing in this category of evaluation technique. Usability tests are more efficient in terms of cost. Compared to usability testing, inspections cost more on a per-problem basis, although they are more effective at finding smaller usability problems. Virzi (1997) said that instead of being used exclusively, usability inspection should be either a precursor or an enhancement to an empirical usability test.

Usability Testing

Testing involves a controlled experiment complete with a working hypothesis, test participants, formal procedures, and analysis of statistics (Wixon & Wilson, 1997). The imagination and resources of the experimenter determine the scope and level of detail. Testing can be conducted on system use in either a lab or real-world setting (Wixon & Wilson, 1997). The main purpose of testing is to answer a specific question about the interface, such as the amount of time required to perform a task, or the number of errors encountered. Subjective data, such as user satisfaction, can also be collected and evaluated.

An experiment is designed to evaluate whether the interface meets the criteria established in the form of a usability goal defined at the outset of the project (e.g., satisfaction, learnability). Prior to beginning, all goals are assigned quantitative levels and matched to usability metrics in the context of the system being tested (Wixon & Wilson, 1997). The primary steps for conducting a usability test are very similar to those followed in human factors experiments (cf. Sanders & McCormick, 1993, chap. 2; Wixon & Wilson, 1997).

There are some important disadvantages to usability testing that must be considered in deciding whether it is appropriate for the task, including that the technique requires an experimenter to observe the user and/or record performance data. The method can also be quite intrusive to performance (Dix et al., 1998). Lab studies require staffing and equipment, which deter some companies from formal evaluation. Furthermore, the testing cycle places additional time at the end of the development process, which may delay the release of a system. In some cases, testing can be integrated in the development process, thereby reducing any unreasonable extension of time beyond the completion date (Wixon & Wilson, 1997).

Comparisons of Inspection and Testing Methods

Opinions on which evaluation methods are the most effective for finding usability problems, or fostering usability solutions, are mixed. Several studies have compared different techniques in terms of effectiveness, cost, and accuracy (Doubleday, Ryan, Springett, & Sutcliffe, 1997; Jeffries et al., 1991; Kantner & Rosenbaum, 1997; Nielsen & Phillips, 1993; Virzi, 1997). The following list includes examples of some of the results:

- Usability tests can find more problems than the cognitive walkthrough and identify problems believed to be more severe (Jeffries et al., 1991; Virzi, 1997).

- Heuristic analysis performed by experts can find more problems, faster than think-aloud (observational) techniques (Virzi, 1997).
- Usability testing is most effective at identifying serious and recurring problems but not low-priority ones (Virzi, 1997).
- Usability testing produces the most accurate performance data, followed by GOMS, and then by heuristic evaluation (Nielsen & Phillips, 1993).
- Usability testing is more expensive than cognitive walkthroughs, heuristic evaluation, or GOMS modeling (Nielsen & Phillips, 1993).

Despite the distribution of results, these studies unanimously agree that (a) no single technique is superior for all usability evaluation tasks, (b) any single technique is better than none at all, and (c) the ideal combination is to use a usability test combined with another technique. For development teams interested in conducting a usability evaluation, there are several issues that need to be considered before selecting a method, or combination of methods. Each technique varies along several parameters, including the optimal phase in system development at which it can be applied, the training and expertise of the person who will administer it, the size of the team conducting the evaluation, the amount of time required, the type of information and comprehensiveness of the results, the subjectivity or objectivity of results, the degree of intrusiveness, and of course the cost of the evaluation (Dix et al., 1998). When comparing the techniques, no single method is superior in all areas. If a development team were to select only one technique, some aspect of the evaluation would ultimately be compromised. However, in many cases, usability evaluation techniques can be effectively combined to produce more thorough results. (We address this in the next section.)

The techniques reviewed here represent only a small number of the available methods. There is no shortage of techniques for evaluating the usability of an interface, but each method has specific strengths and weaknesses. Despite this, there has been only limited application of usability evaluation in aviation systems (Kaber et al., 2002).

USABILITY IN AN AVIATION CONTEXT

In general, use of contemporary usability techniques is often perceived as an effort-intensive activity for a limited return on investment (Ivory & Hearst, 2001). Aviation systems design teams need methods which are relatively inexpensive to use and provide some aid in HCI-related decisions (Feary et al., 2003). More specifically, current development procedures, and the certification process, dictate that effective techniques for evaluating usability in aviation systems have the following characteristics:

- Rapid execution: Technology is advancing faster than the capability to add it to existing or even new aircraft. Any usability evaluation technique should not slow this process further.
- Cost-effective: Current requirements in aviation can increase the cost of designing a new component by a factor of three in some cases (Abbott et al., 1999). The usability evaluation should not pose an excessive additional financial burden on vendors.
- Integrated in the full development life cycle: The need to build usability into the development process has been emphasized by several publications, including aviation human factors research (Abbott et al., 1999; Singer, 1999; Smith, 1999; Stager, 2000; Stein, 2000; Williges et al., 1988; Wixon & Wilson, 1997).
- Input from a variety of domain experts: As in designing any interactive system, it is crucial that aviation systems be designed for the intended user population. Pilots should be included early and continually in the development process (Stein, 2000; Virzi, 1997).
- Transferability and scalability to cockpit design: Any usability evaluation technique must be adaptable to the context of cockpit interface design or ATC workstation interface design and should be scalable for application to a single new component or an entire system.

Applicability of the Evaluation Techniques

On the basis of the criteria just presented, we analytically evaluated the applicability of the various evaluation techniques reviewed. There has been some support for usability testing in aviation systems design because of the quality of the results, the flexibility of the application, the preferences of usability experts, and the potential for wide acceptance (Williges et al., 1988). Although more expensive than inspection methods, the advantage of accurate, quantifiable results is often considered to outweigh the costs. However, if cost becomes crucial to a project, there are procedures for reducing the money and time spent on a formal experiment.

For example, Miller and O'Donnell (1993) developed a methodology for usability testing that reduces the expected overhead for a lab environment: Low-overhead usability testing is designed to be executed in 1 or 2 days for a cost of, at most, $500. The procedure requires a spacious office or conference room and 12 (or more) participants, who are divided into two groups performing tests in two separate sessions. The test environment is still treated as formal by the evaluators, who continue to behave as if they were running a formal laboratory experiment. During the test itself, participants respond to questions on interface tasks using a predesigned, electronic feedback form. At the same time, evaluators use an "observer" to record information on participant behaviors and performance. Both the feedback and observation forms contain subjective and objective data. After participants complete the evaluation session, they answer subjective questions as a group, and evaluators consolidate the data from all the forms and the group interview on an end-of-day checklist.

There are many parallels between the low-overhead method and the elaborate usability testing methodology described by Wixon and Wilson (1997), such as employing an administrator and observers, organizing prescribed test materials in advance, controlling the environment, selecting the participants, and basing data on the participants' performance in a defined set of tasks. Although they have cut several corners, Miller and O'Donnell (1993) adhered to the basic constructs of formal usability testing. The low-overhead method may be effective when the time to complete a project or the budget is constrained. The trade-off is that the limited time spent in the lab results in a corresponding reduction in the amount of data and the accuracy of the evaluation, as compared to actual usability testing. It is up to the aviation systems design team to determine which is more important.

For over 10 years, heuristic evaluation has been used effectively to perform rapid usability inspections. Heuristic evaluation fits all the unique requirements for usability evaluation in the cockpit. It can be performed at any phase of the development process, either with or without an interface, and it can be performed both faster and with less expense than the other inspection techniques. It can be performed by experts or nonexperts, but in the context of aviation systems it would likely need to be overseen by an expert trained both in usability and in piloting and so forth.

It is also likely that heuristic evaluations are transferable to aviation. Kaber et al. (2002) conducted a warm heuristic-based analysis of the Multi-Control Display Unit (MCDU) component of the FMS to identify usability issues. In their study, the evaluators observed the use of an MCDU based on the design implemented in the MD–11 passenger aircraft in a hypothetical flight task. Seven experts participated in a group evaluation of the MCDU based on principles under the major headings of learnability, flexibility, and robustness. The evaluators concluded that the MCDU violated several of the principles in these categories.

We can also use the example of the reported problems with the FMS to conduct a cold heuristic evaluation. In review, Sarter and Woods (1992) cited weak feedback, incomplete mental models, and imprecise expectations of the outcome of some command sequences, whereas Riley (2001) noted inconsistency with the pilot's language and procedures as current problems in the FMS interface. When comparing these issues to the list of Nielsen's (1994) heuristic principles presented earlier, there appears to be some overlap. Table 1 shows one possible matching of usability problems with the FMS to specific heuristics.

It would be too simplistic to say that the existing problems with the FMS would have been prevented by application of a heuristic evaluation. However, given the overlap between the problems with the FMS and the usability principles, as demonstrated by Kaber et al. (2002) and shown here, it is likely that heuristic evalua-

TABLE 1
Matching of FMS Usability Problems With Heuristics

FMS Problem	*Applicable Heuristic*	*Definition of Heuristic*[a]
Weak feedback	Visibility of system status	"The system should always keep system status users informed about what is going on, through appropriate feedback within reasonable time" (p. 30).
Inconsistency with pilot language	Match between system and the real world	"The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order" (p. 30).
Incomplete mental models	Recognition rather than recall	"Make objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate" (p. 30).

Note. FMS = flight management systems.
[a]Heuristic definitions are from Nielsen (1994).

tion could have added some value to the design process and that design recommendations based on this type of evaluation could be useful for improving FMS usability.

It may be determined that current design heuristics for usability may not be directly transferable to an aviation context. There may be a need to adapt existing heuristics and develop new ones for evaluations of aviation systems. Research has already been performed to identify new guidelines for modern aviation systems. Billings (1997) formulated 15 guidelines that could benefit the existing certification requirements for systems, including some degree of automation. It may be possible to use these contemporary human factors principles as a basis for heuristic evaluation.

The direct observation technique, as part of a cooperative evaluation, also fits with the requirements for evaluating aviation interfaces. This method is quick, is inexpensive, and utilizes the input of users and multiple experts by design. This method has also been used for development of aviation systems (Williges et al., 1988), so its transferability and scalability may not be in question.

Combining Techniques

With respect to combining multiple techniques as part of a usability evaluation, given the importance of cost in developing aviation systems, the completeness and accuracy of results should justify the specific approach. Because of the high level of expert involvement, combining a formal usability test with a cognitive walkthrough, for example, would be fairly expensive and should only be considered if the results are far superior to less costly combinations. Given the scope of the walkthrough technique, it may not be necessary when usability tests are already being conducted. For similar reasons, a detailed GOMS model is not advantageous when usability testing is possible. As stated earlier, the cognitive walkthrough specifically evaluates ease of learning and the GOMS model produces performance data. Performance and learnability can be evaluated through empirical testing. That is, usability tests can be adapted to include the benefits of the cognitive walkthrough and GOMS model without the additional (and costly) involvement of experts required to perform the cognitive walkthrough, or the additional time required to develop a detailed GOMS model. In one experiment, the performance data derived from a usability test were determined to be more accurate than the data produced by a GOMS model, which supports the use of the formal test (Nielsen & Phillips, 1993). In contrast, neither the cognitive walkthrough nor the GOMS model can be adapted to produce results equivalent to usability testing.

Usability Trends in Aviation

There is some additional evidence to support the transfer of usability evaluation techniques to aviation interface design. In formulating suggestions for new certification requirements, Singer (1999) consulted existing HCI guidelines for ways to improve displays, such as using natural dialog; speaking the user's language; minimizing the user's memory load; being consistent; providing appropriate feedback, shortcuts, clear exits to processes, and understandable error messages; and preventing errors. He also suggested that pilots of varying backgrounds be used in simulation programs and tests instead of using seasoned testers to review new system prototypes. This evaluation approach is consistent with aspects of heuristic evaluation and cooperative evaluation.

Williges et al. (1988) also reviewed HCI guidelines and the software design process when proposing a development methodology for aircraft and ATC software. In their chapter, the authors described a 14-step design process (see Table 2). The methodology is similar to models recommended for user-centered iterative design, such as those described by Gould, Boies, and Ukelson (1997), which advocate an early and continual focus on users, validated by ongoing user testing, and an integrated and iterative design that reflects the user's input. Because of the comprehensive scope of the development process and application to software design in the context of aviation systems, methodologies like these provide a good framework for determining the stage of development where usability evaluation will have the greatest impact.

TABLE 2
Development Methodology for Aircraft and ATC Software (Williger et al., 1988)

Step	*Description*
Design objectives	The design goals of the aviation system must be stated before the software is designed. This includes user-oriented metrics, measures of learnability, usability, and human factors design principles.
Task-function	This step includes identification of inputs analysis and outputs; specifications for dialog sequences; and the control structure for interfaces, dialogs, and computations.
Focus on users	The initial design should involve input from end users of the system. This can include interviews, questionnaires, statistics, and literature reviews.
Dialogue design guidelines	Because each design situation can vary within a system, the structure of the human–computer interaction should also be defined in advance. To ensure consistency and improve the initial design.
Structured	The final step of the initial design stage walkthroughs involves combining all the information into a description of the proposed control structure. Users and designers then walkthrough actual aircraft scenarios using the proposed design, either formally or informally.
Initial design modifications	Using feedback from the walkthrough, changes are often made to the initial design. The above steps are repeated until no modifications are required.
Rapid prototyping	Software prototypes are drafted using the initial design specifications.
User-defined interfaces	As with the initial design stage, the end users should be involved during the design of the software prototypes.
User-acceptance testing	Once prototypes are completed, they should be tested by the pilot or controller who will be using them. These evaluations can be done one on one, in small groups, or compared to field data.
Iterative redesign	As with the initial design, feedback from acceptance testing should be applied directly to the prototype, which is modified further until it receives approval from the operator.
Operational software interface	Production software is completed.
Benchmarking	The simulated tasks representing actual users' tasks should be used in conducting a summative evaluation.
Formal experimentation	An experiment testing the interface based on the requirements and the benchmark values should be conducted to validate the design.
Feed-forward results	The data obtained on the existing interface can be used in designing future interfaces.

DISCUSSION AND RECOMMENDATIONS

Based on our comparisons of the advantages and disadvantages of the various evaluation techniques and our matching them to the needs of aviation systems design, usability testing, heuristic evaluation, and cooperative evaluation may be most applicable to the aviation context. These three techniques applied at strategic intervals within the development process may greatly reduce the potential for a system to go into production with usability flaws.

Heuristic Evaluation

A usability domain expert should perform a cold heuristic evaluation early in the design process during the task-function analysis step before any prototypes are developed. This can expose flaws in the design before the design team begins drafting the first prototype, resulting in less time lost due to revisions. The use of a usability/domain expert can increase the quality of the evaluation.

Another additional heuristic evaluation could be conducted each time a prototype is developed for review by the design team in order to validate the design. The results of all the evaluations should be communicated within the design team and carried forward to the next phase of evaluation.

Cooperative Evaluation

Similar to Williges et al. (1988), we recommend conducting cooperative usability evaluations, attended by representatives from various system development areas (computer programmer, graphic artist, human factors expert, etc.), throughout the development process subsequent to use of heuristic evaluation. Cooperative evaluations should be conducted after the results of the heuristic evaluations have been applied to the design or the prototypes to limit the amount of time the larger group will need to spend in evaluation meetings.

Usability Testing

Once a working model of the interface is complete and the inputs from the heuristic and cooperative evaluations have been applied, a formal usability test, involving a sample of the user population, should be conducted. Ideally, the project team should directly observe the process. In general, pilot performance data and errors should be recorded as well as data relevant to the new aspects of the interface. It is important to define usability metrics to assess the design principles of concern in advance of performing any tests. As noted by Singer (1999) in the domain of aviation, and Wixon and Wilson (1997) in describing the usability engineering framework, usability measures are most useful when combined with a quantitative frame of reference.

Last, any aviation systems usability testing protocol should be designed to reflect both the component that is changing as well as its affect on the system as a whole (Billings, 1997; Stein, 2000; Woods & Sarter, 1993). In addition to any performance data collected on the new component in isolation, there should also be a test that gathers quantitative data comparing overall system performance with the new component to performance in the same environment without it. Stein offered that conditions such as motivation, mental fatigue, and the complexities of human information processing may not be an issue when designing a piece of hardware or software in isolation but will be apparent when the system is used in the real world. When the evaluation is complete, results should be preserved to provide benchmark data for the next generation of system. Several databases already exist that can give values for errors per flight hour, which could be used as benchmarks for initial testing (Singer, 1999).

CONCLUSIONS

Performing usability evaluations of highly complex and dynamic systems, such as flight decks or ATC workstations, poses significant challenges for usability experts. The classic problems of collaborating with a resistant design team, convincing managers of the importance of usability evaluation despite tight budgets, and struggling to assemble participants for empirical tests are still present but are overshadowed by the need to evaluate a very complex interface with narrow margins for error and high private and public visibility. Furthermore, the environment created by the certification process makes the evaluation process even more difficult by restricting expense and applying outdated and ambiguous human factors guidelines (Singer, 1999).

More usability problems can be found in a system by combining a mix of complementary usability evaluation techniques, each of which specializes in a different method of exposing usability problems. The effectiveness of these methods can be increased even further through the involvement of human factors and domain experts early in the development life cycle. Formal usability testing, heuristic evaluation, and cooperative evaluation can be applied to aviation systems in this manner.

As technology continues to drive changes in aviation, there is a need for detailed usability evaluations of new components added to aviation systems. At present, technology is evolving more rapidly than the certification process can accommodate. Usability evaluations of aviation systems need to be just as dynamic as the technology itself to meet the demand.

Finally, this research may have relevance to other contexts, including manufacturing organizations attempting to improve their human factors design process (e.g., advanced machine technologies, medical devices, etc.) in response to increasingly sophisticated human factors/ergonomics regulations and principles of good professional practice. As discussed, regulations in the avionics industry now require human factors test plans for certification, and they ask for information on specific measurement and evaluation techniques to be used, including objective human performance measures to address design features of concern. The formal usability evaluation methods we describe in this article may be applicable to the manufacturing domain, and companies should consider the advantages and drawbacks of each in developing test plans that are both effective and economical.

ACKNOWLEDGMENTS

Michael Clamann's work on this project was supported in part by a NASA Graduate Student Researchers Program Fellowship (Grant NGT–1–01004). The technical point of contact was Lance Prinzel. David Kaber's work on this article was supported by NASA Langley Research Center (Crew/Vehicle Integration Branch and the Psychological and Physiological Stressors and Factors Program) through Grants NAG1–01–002 and NAG–1–02056. The grant monitors were Anna Trujillo and Lance Prinzel. The views and conclusions presented in this article are those of the authors and do not necessarily reflect the views of NASA.

REFERENCES

Aeronautica Civil of the Republic of Colombia. (1996). *Aircraft accident report: Controlled flight into terrain, American Airlines Flight 965, Boeing 757–223, N651AA, near Cali, Colombia, December 20, 1995.* Santafe de Bogota, Colombia: Author.

Abbott, D. W., Wise, M. A., & Wise, J. A. (1999). Underpinnings of system evaluation. In D. J. Garland, J. A. Wise, & V. D. Hopkin (Eds.), *Handbook of aviation human factors* (pp. 51–66). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Benel, R. (1998). Workstation and software interface design in air traffic control. In M. Smolensky & E. Stein (Eds.), *Human factors in air traffic control* (pp. 341–390). San Diego, CA: Academic.

Billings, C. E. (1997). *Aviation automation: The search for a human-centered approach.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Dix, A., Finlay, J., Abowd, G., & Beale, R. (1998). *Human-computer interaction* (2nd ed., pp. 406–441). London: Prentice Hall Europe.

Doubleday, A., Ryan, M., Springett, M., & Sutcliffe, A. (1997). A comparison of usability techniques for evaluating design. In *Proceedings of DIS '97: Designing interactive systems: Processes, practices, methods, & techniques* (pp. 101–110). New York: ACM.

Feary, M., Sherry, L., Polson, P., & Fennel, K. (2003). Incorporating cognitive usability into software design processes. In J. Jacko & C. Stephanidis (Eds.), *Human–computer interaction: Theory and* *practice (Part I). Proceedings of the 10th International Conference on Human–Computer Interaction* (pp. 427–431). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Funk, K., Suroteguh, C., Wilson, J., & Lyall, B. (1998). Flight deck automation and task management. *Proceedings of 1998 IEEE International Conference on Systems, Man, and Cybernetics* (Vol. 1, pp. 863–868). San Diego, CA: IEEE.

Garland, D. (1999). Air traffic controller memory: Capabilities, limitations, and volatility. In D. J. Garland, J. A. Wise, & V. D. Hopkin (Eds.), *Handbook of aviation human factors* (pp. 455–496). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Gould, J., Boies, S., & Ukelson, J. (1997). How to design usable systems. In M. Helander, T. Landauer, & P. Prabhu (Eds.), *Handbook of human-computer interaction* (pp. 231–255). New York: Elsevier.

Hopkin, V. D. (1988). Air traffic control. In E. Wiener, & D. Nagel (Eds.), *Human factors in aviation* (pp. 639–663). San Diego, CA: Academic.

Hopkin, V. D. (1998). The impact of automation on air traffic control specialists. In M. Smolensky & E. Stein (Eds.), *Human factors in air traffic control* (pp. 391–419). San Diego, CA: Academic.

Irving, J., Polson, P. G., & Irving, J. (1994). *Applications of formal methods of human computer interaction to training and use of the control and display unit* (Tech. Rep. No. 94–08). Boulder: University of Colorado.

Ivory, M. Y., & Hearst, M. A. (2001). The state of the art in automating usability evaluation methods. *ACM Computing Surveys, 33,* 470–516.

Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. M. (1991). User interface evaluation in the real world: A comparison of four techniques. In S. P. Robertson, G. M. Olson, & J. S. Olson (Eds.), *Proceedings of ACM CHI '91 Conference on Human Factors in Computing Systems* (pp. 119–124). New York: ACM.

Kaber, D., Riley, J. & Tan, K.-W. (2002). Improved usability of aviation automation through direct manipulation and graphical user interface design. *The International Journal of Aviation Psychology, 12,* 153–180.

Kantner, L., & Rosenbaum, S. (1997). Usability studies of WWW sites: Heuristic evaluation vs. laboratory testing. In *Proceedings of SIGDOC '97: ACM 15th International Conference on Systems Documentation* (pp. 153–160). New York: ACM.

Miller, M., & O'Donnell, C. (1993). Usability testing on a shoe string. In S. Ashlund, K. Mullet, A. Henderson, E. Hollnagel, & T. White (Eds.), *Interchi '93 Adjunct Proceedings* (pp. 85–86). New York: ACM.

Newman, W., & Lamming, M. (1995). *Interactive system design.* Boston: Addison-Wesley.

Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen, & R. Mack (Eds.), *Usability inspection methods* (pp. 25–62). New York: Wiley.

Nielsen, J., & Mack, R. (1994). *Usability inspection methods.* New York: Wiley.

Nielsen, E., & Phillips, V. (1993). Estimating the relative usability of two interfaces: Heuristic, formal, and empirical methods compared. In S. Ashlund, K. Mullet, A. Henderson, E. Hollnagel, & T. White (Eds.), *Proceedings of ACM INTERCHI '93 Conference on Human Factors in Computing Systems* (pp. 214–221), New York: ACM.

Polson, P., & Smith, N. (1999). The cockpit cognitive walkthrough. In R. S. Jensen (Ed.), *Proceedings of the Tenth International Symposium on Aviation Psychology.* Columbus: Ohio State University.

Riley, V. (2001, Spring). A new language for pilot interfaces. *Ergonomics in Design,* pp. 21–26.

Rowley, D. E., & Rhoades, D. G. (1992). The cognitive jogthrough: A fast-paced user interface evaluation procedure. In P. Bauersfeld, J. Bennett, & G. Lynch (Eds), *Proceedings of CHI '92* (pp. 389–395). New York: ACM.

Sanders, M. S., & McCormick E. J. (1993). *Human factors in systems design* (7th ed.). New York: McGraw-Hill.

Sarter, N., & Woods, D. (1992). Pilot interaction with cockpit automation: Operational experiences with the flight management system. *The International Journal of Aviation Psychology, 24,* 303–321.

Sarter, N., Woods, D., & Billings, C. (1997). Automation surprises. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (2nd ed.) (pp. 1926–1943). New York: Wiley.

Sexton, G. A. (1988). Cockpit-crew systems design and integration. In E. Wiener & D. Nagel (Eds.), *Human factors in aviation* (pp. 495–525). San Diego, CA: Academic.

Singer, G. (1999). Filling the gaps in the human factors certification net. In S. Dekker & E. Hollnagel (Eds.), *Coping with computers in the cockpit* (pp. 87–107). Brookfield, VT: Ashgate.

Smith, C. (1999). Design of the Eurofighter human machine interface. In *Air and space Europe* (Vol. 1, No. 3, pp. 54–59). Amsterdam: Elsevier Science.

Stager, P. (2000). Achieving the objectives of certification through validation: Methodological issues. In J. Wise & V. Hopkin (Eds.), *Human factors in certification* (pp. 91–104). London: Lawrence Erlbaum Associates, Ltd.

Stein, E. S. (2000). A critical component for air traffic control systems. In J. Wise & V. Hopkin (Eds.), *Human factors in certification* (pp. 57–63). London: Lawrence Erlbaum Associates, Ltd.

Stokes, A., & Wickens, C. (1988). Aviation displays. In E. Wiener & D. Nagel (Eds.), *Human factors in aviation* (pp. 387–431). San Diego, CA: Academic.

Strauch, B. (1997). Automation and decision making—Lessons from the Cali accident. In *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting* (pp. 195–199). Albuquerque, NM: Human Factors and Ergonomics Society.

U.S. Department of Transportation/Federal Aviation Administration. (1993). *Advisory circular (AC) 25.1523–1.* Washington, DC: Author.

U.S. Department of Transportation/Federal Aviation Administration. (1995). *14 CFR Part 25—Airworthinesess standards: Transport category airplanes* (Chap. 1, Subchap. C). Washington, DC: Author.

Virzi, R. A. (1997). Usability inspection methods. In M. Helander, T. Landauer, & P. Prabhu (Eds.), *Handbook of human-computer interaction* (pp. 705–715). New York: Elsevier.

Williges, R. C., Williges, B. H., & Fainter, R. G. (1988). Software interfaces for aviation systems. In E. Wiener & D. Nagel (Eds.), *Human factors in aviation* (pp. 463–493). San Diego, CA: Academic.

Wixon, D., & Wilson, C. (1997). The usability framework for product design and evaluation. In M. Helander, T. Landauer, & P. Prabhu (Eds.), *Handbook of human–computer interaction* (pp. 653–685). New York: Elsevier.

Woods, D., & Sarter, N. (1993). Human interaction with intelligent systems in complex dynamic environments. In D. Garland & J. Wise (Eds.), *Human factors and advanced aviation technology* (pp. 107–110). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

Manuscript first received June 2002

Copyright of International Journal of Aviation Psychology is the property of Lawrence Erlbaum Associates and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.