Designing and Testing a Tool for Evaluating Electronic Flight Bags

Divya Chandra[1], Michelle Yeh[1], and Vic Riley[2]

[1]United States Department of Transportation Volpe National Transportation Systems Center
55 Broadway, Cambridge, MA 02142, USA
{chandra, yeh}@volpe.dot.gov

[2]User Interaction Research and Design, 2125 Whalen Drive, Point Roberts, WA 98281, USA
vic@uird.com

ABSTRACT

The Federal Aviation Administration (FAA), system designers, and customers all recognize that Electronic Flight Bags (EFBs) are sophisticated devices whose use could affect pilot performance. As a result, human factors issues have received considerable attention from the EFB community. In addition, the FAA's Advisory Circular (AC) on EFBs (AC 120-76A) identifies a need for evaluating EFBs from a human factors perspective and contains a list of human factors considerations for review. However, the AC does not specify exactly how to do the field human factors evaluation.

Our research is directed at developing tools and procedures that could be used by FAA field evaluators in conducting structured and comprehensive, yet practical, EFB usability evaluations. The tools and methods were developed and refined over the course of several tests with real EFB systems. In this paper, we describe the evolution of one promising tool into its latest, relatively mature, format. We also present our test procedure and methods of processing the resulting data into feedback for the manufacturer. Our next step is to expose more potential users, especially those in the FAA, to the tools and methods to determine if these products are useful in practice.

Keywords

Electronic Flight Bag, EFB, usability, evaluation, human factors tool, system design, design

INTRODUCTION AND PROJECT GOALS

The Electronic Flight Bag (EFB) industry is flourishing. As of September 2003, Chandra, Yeh, Riley, & Mangold list fifteen companies that supply EFB systems, plus others that supply either EFB software or hardware products [4]. The Federal Aviation Administration (FAA) defines an EFB as any "electronic display system intended primarily for cockpit/flight-deck or cabin use." [6] In practice, the term "EFB" describes a wide variety of devices. Some common functions include electronic documents, electronic charts, and flight performance calculations. Other available functions include cabin video surveillance and surface moving map displays. (See [6] for a more complete definition and more examples of EFB capabilities.)

There are two main reasons for the rapid development and implementation of EFBs. First, they appeal to a wide audience because of their flexibility and cost. EFBs come in a variety of form factors and support a range of functionality that can be customized for any type of operator—general aviation, charter/business, or air transport. Many EFBs are based on commercial-off-the-shelf computer technology that is customized for the flight deck environment, so they are not as costly as traditional installed avionics. Second, the March 2003 Advisory Circular (AC) 120-76A issued by the FAA allows a streamlined field approval process for EFBs [6]. The EFB AC also gives industry designers and government regulators a clearer understanding of issues to be reviewed.

A significant portion of AC 120-76A (Section 10) addresses human factors considerations for EFBs [6]. The FAA, system designers, and customers all recognize that EFBs are sophisticated devices whose use could affect pilot performance [1–4, 6, 8]. In particular, many EFBs make extensive use of graphical user interfaces and can support multiple new functions, some of which may impact operating procedures. As a result, human factors issues have received considerable attention from the EFB community. Some of the specific issues called out in the EFB AC include user interface consistency, legibility, error potential, and workload. The EFB AC refers to a more comprehensive document on human factors considerations for the design and evaluation of EFBs [3]. That document, from 2000, has since been updated and superseded by Chandra et al., 2003 [4]. For a brief overview of these lengthy technical reports, see Chandra, 2002 [1].

The need for evaluating EFBs from a human factors perspective is identified by the FAA [6], but the procedure for doing this evaluation is not specified. In fact, translating the general human factors guidance into a thorough yet practical EFB evaluation is a non-trivial task. Our research is directed at developing new tools and methods that could be used by FAA field evaluators in conducting structured and comprehensive EFB usability evaluations in the field. Two key field constraints are that the evaluations are brief—just 2 to 4 hours long, and they may be performed by staff who are not human factors experts. The purpose of the evaluation tools and methods is to help conduct a thorough, systematic evaluation to identify major system weaknesses. The products of this research will be publicly available, so system manufacturers will know what to expect in advance. Manufacturers could use the tool and procedure in-house to anticipate general results of a future regulatory inspection.

In the next section, we describe the evolution of the usability assessment tool towards its latest version. Note that a preliminary version of the tool and a detailed description of

its origins are presented in Chandra, 2003 [2]. That paper is directed at an audience of system manufacturers; it includes *general* advice for conducting in-house evaluations. Here, we briefly review the earlier versions of the tool but focus on the current version. We show how the tool has evolved *and why*, from a research perspective. After presenting the tool, our test procedure is detailed. We describe how the test evaluations were performed, who performed them, and what data were collected. After describing the test procedure, we describe the data analysis and synthesis steps that were followed to prepare feedback for the system manufacturers. Finally, we briefly discuss preliminary feedback from manufacturers on the tools and our plans for follow-on work.

EVOLUTION OF THE TOOL

This research builds upon earlier work in which human factors considerations related to EFBs were identified and prioritized. The results of this work were documented in a lengthy and detailed set of guidance for EFB design and evaluation [4]. While this guidance is comprehensive and informative, the document is cumbersome to use in a brief field evaluation. We quickly realized that a short paper-based tool that could serve as a "guide for usability assessment" would be more practical. The tool would list usability and design topics to be evaluated.

In Chandra, 2003 [2], we reported on preliminary versions of the EFB assessment tool. In that paper, we presented two sets of items for the assessment tool. One was a short high-level list of user interface topics with about 20 items, and the other was a long list of over 180 items created by condensing the full-length document [4]. Samples from these lists and some alternative formats for presenting the items are illustrated in Chandra, 2003 [2].

We have since tested a variety of formats for the tool. One version, for example, is contained in Appendix B of Chandra et al. [4]. Appendix B is an 11-page summary of roughly 100 pages of equipment requirements and recommendations. The format of Appendix B, and sample guidance, are shown in Figure 1. Each item is a paraphrased version of guidance from the main document. When viewed electronically, links to the full topic description are active. The paraphrased guidance is most useful to readers who are already familiar with the structure and general content of the version of guidance from the main document. The format used in Appendix B of Chandra et al. [4] is easily transferred into that of a detailed usability assessment tool (see Figure 2). The format for the guidance in Figure 2 is tighter than that used for Appendix B, but it is still lengthy. For a generic EFB system, the tool is five pages long; including the topics for specific applications adds another five pages.

Note that Appendix B and the detailed tool format shown in Figure 2 provide heuristics only. There is no designated space for recording evaluator comments or ratings, which are important products of an evaluation. We considered many different ways of incorporating space for comments, but in the end decided to leave them separate because of the added flexibility. For example, notes can be recorded directly into an electronic file using a word processor, handwritten on a paper copy of the tool, or written into a separate notebook.

The rating scale was an open issue for the tool in Chandra, 2003 [2]. We considered several options, such as a 3-, 4-, or 5-point acceptability scale. We expected that higher-resolution scales would provide designers with more

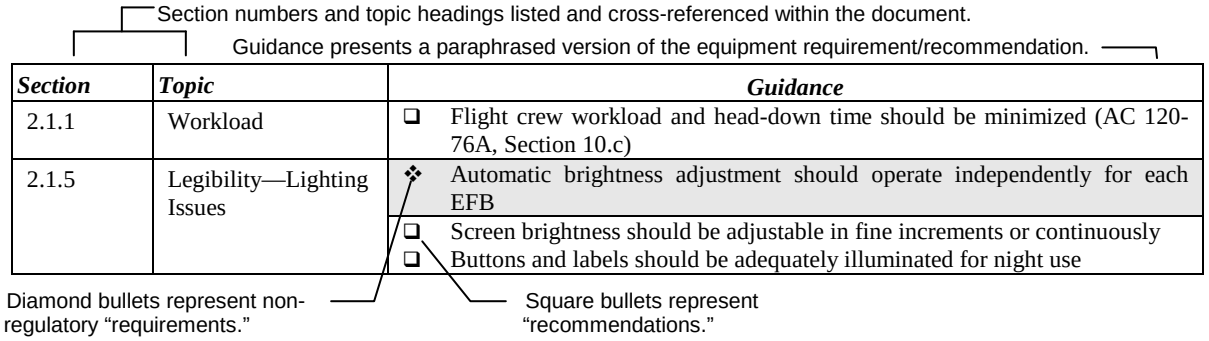

Figure 1: Format of EFB summarized equipment requirements and recommendations from Chandra, Yeh, Riley & Mangold, 2003 [4]. The structure is intended to support quick review with pointers to more detail when needed.

2.4.3 General Use of Colors

- ❖ Red and amber should be reserved for highlighting *warning* and *caution* level conditions respectively (AC 120-76A, 10.d(1))
- ❖ Color should not be sole means of coding important differences in information; color should be used redundantly
- ❖ Color-coding scheme should be interpretable easily and accurately
- ❑ Each color should be associated with only one meaning
- ❑ No more than six colors with assigned meanings should be used in a color-coding scheme
- ❑ EFB colors should not conflict with flight deck conventions
- ❑ For Part 121 and 135, default colors that represent different types of data should be customizable only by an appropriately authorized administrator
- ❑ If colors are customizable, there should be an easy way to return to default settings

Figure 2: Format of an EFB usability assessment tool based on Appendix B (Summary of Equipment Requirements and Recommendations) in Chandra, Yeh, Riley & Mangold, 2003 [4].

detailed information about the quality of the system, but they would also increase the time required for the evaluation.

For the purposes of a brief regulatory review, the rating scale was expected to be more coarse, e.g., acceptable or not acceptable. In the latest version of the assessment tool, we decided not to suggest a rating scale. Doing so gives the evaluator more flexibility in deciding how to designate severity ratings and accommodates individual rating preferences and styles. In addition, severity ratings can be assigned post-hoc based on evaluators' notes on the impact, frequency, and persistence of problems [7].

Figure 3 shows the latest version of the high-level EFB usability assessment tool. This is simply a list of topics to consider for the evaluation; evaluators are asked to go through the list commenting on each item. The comments, which could be either positive or negative, can actually be more valuable to a designer than severity ratings because they give the designer insight into the cause of any difficulties, not just their severity. In some cases, topics in the tool will not be relevant, but it is important to consider every item to ensure a thorough evaluation. As the evaluators comment on each item, they provide supporting examples, and, if they choose, preliminary assessments of problem severity. The one-page version in Figure 3 is for a generic EFB system. A 2.5-page version contains additional customized guidance for four applications (electronic documents, electronic checklists, electronic charts, and flight performance calculations).

Through tests of the tool against real systems, we honed its content, item order, and language. The content of the tool was generated from a generic high-level list of user interface dimensions (see Chandra [2] for the full initial list). This list was fleshed out by adding items that represented themes (i.e., groups of items) in the detailed tool generated from the full-length EFB document [4]. The net result is a good blend of high-level and somewhat more specific topics.

Our philosophy for item order was to go from concrete to abstract or local to global. Because evaluators may still be familiarizing themselves with the system early in the evaluation, we expect that they will find it easier to start by commenting on concrete aspects of the design (e.g., icons and formatting). As they build up experience with the system, they will be able to comment on more abstract, potentially global, aspects of the design (e.g., error potential, consistency across applications, or workload).

One topic, error handling and prevention, is brought up more than once in the tool. It is listed as a concrete subtopic in some cases (e.g., as "potential for errors," under the Hardware topic, and as "confusability" under the Symbols and Graphical Icons topic), and it is called out as a more general topic overall. By listing it both ways, we are more likely to capture detailed comments regarding error potential, which may be especially important to regulators.

Language was a significant issue in earlier versions of the tool. Terms that were familiar to some human factors experts were not always intuitive for non-human factors experts; some terms were not even clear among human factors experts. Our latest tests show that the current language is understandable, or at least not distracting, to evaluators. During the evaluations, if the evaluators did not understand a term, they were asked to guess at its meaning and then use their own definition to complete the tool.

EVALUATION PARTICIPANTS AND PROCEDURE

We developed and refined the EFB usability assessment tool over the course of several evaluations with realistic systems that were volunteered by vendors for the purpose of trying out the test procedure and draft assessment tools.

In most cases, a team of two evaluators worked together to complete the evaluation through co-discovery with a talk-aloud protocol. We preferred the co-discovery technique because it fits with the typical FAA evaluation process in which teams of two to four evaluators review a system together. Also, co-discovery is useful because evaluators working in teams often discover more about a system than evaluators working alone do.

The dialogue between the two evaluators was transcribed by a note-taker/observer. Having a dedicated person to observe and take notes is not standard in a regulatory evaluation, in which evaluators are generally responsible for taking their own notes. Our early tests revealed, however, that evaluators were distracted by having to take their own notes and note-taking disrupted the flow of the open-ended discussions in progress. Also, notes taken by the evaluators tended to be incomplete and not especially useful to anyone but their author (if that). In contrast, notes from a dedicated note-taker were relatively complete transcripts of the sessions, recorded directly into an electronic document in our tests.

The evaluators in our tests were researchers with an aviation and/or human factors background. Some were licensed pilots and/or experienced system designers, but they were not FAA personnel or air transport pilots (the intended end users). To give them a sense of the FAA perspective, we sent evaluators materials in advance, including copies of AC 120-76A [6], Appendix B from Chandra et al. [4], AC 25-11 [5], and a draft copy of the assessment tool (Figure 3). FAA staff would definitely be familiar with the two ACs and may have seen the assessment tools before the evaluation as well. In addition, it is helpful for tool users to (1) have enough general knowledge of user-interface components to be able to articulate their impressions of a device, and (2) expect that they will encounter problems and realize that these problems are not their "fault."

The evaluation sessions lasted 3 to 4 hours total. Chandra, (2003) [2] was published based on findings from two EFB evaluations. Since then, we have tested two more EFB units with four evaluation sessions for each unit. We revised the assessment tools between sessions based on participant feedback. The evaluation consisted of the following stages, which are described below:

1) Introduction (15 min)
2) Task-Based Exploration (1 to 1.5 hour)
3) Tool-Based Review
    a. High-Level Tool (up to 1 hour)
    b. Detailed Tool (up to 1 hour)
4) Feedback on tools and wrap-up (15 min)

EFB Usability Assessment Tool

HARDWARE CONSIDERATIONS

- Physical Ease of Use
  — Input devices and display, accessibility of controls
- Labels and Controls
- Lighting Issues (day vs. night use)
  — Brightness adjustment, illumination of labels
- Amount of feedback, potential for errors

SOFTWARE CONSIDERATIONS

Symbols and Graphical Icons

- Clarity of intended meaning, confusability
- Legibility and distinctiveness

Formatting/Layout

- Fonts (size, style, case, spacing)
- Arrangement of information on the display
  — Consistency with user expectations and internal logic

Interaction (Accessing functions and options)

- Home pages and ease of movement between pages
- Number of inputs to complete a task
- Ease of accessing functions and options
- Feedback (system state, alerts, modes, etc.)
- Responsiveness
- Intuitive logic

Error handling and prevention

- Susceptibility to error (mode errors, selection errors, data entry errors, reading errors, etc.)
- Correcting errors (e.g., cancel, clear, undo)
- Error messages

Multiple Applications

- Consistency and compatibility across applications
- Identifying current position within system
- Ease of switching between applications

Automation (if any)

- Is there enough? Too much?
- Is it disruptive/supportive? Predictable? User control over automation? (e.g., manual override)

General

- Consistency of controls/elements; are they distinctive where appropriate?
- Visual, audio, and tactile characteristics
- Use of color (especially red and amber) and color-coding
- Amount of feedback (system state, alerts, modes, etc.)
- Clarity and consistency of language, terms, and abbreviations
- End-user customization (if any)

WORKLOAD

- Problem areas

OTHER

Figure 3. High-level EFB usability assessment tool.

Introduction to the Test

During the introduction, the experimenter explained that the purpose of the test was to evaluate the EFB usability assessment tools and gave a brief introduction to the EFB system and applications that would be reviewed. The brief introduction provided context on the application(s), the system, and their intended use.

The introduction was not intended to mimic formal system training. As a result, the evaluators were possibly less prepared than EFB end users might be with the system. However, we felt that this was an appropriate worst-case scenario to consider because (1) EFB end-users may see some system features infrequently, or under stressful conditions, where intuitiveness could be an important factor in actual performance, and (2) the typical FAA evaluator may not receive full training with the system prior to reviewing it. In addition, manufacturers strive to build systems that require minimum training. Our protocol put this theory to the test.

Task-Based Exploration

The task-based exploration phase was effectively a self-paced familiarization period with the system. Participants stepped through a set of tasks, which were custom designed for the system in advance. The tasks were designed to have a beginning state and a desired goal. They were open-ended enough that users could digress for a while, but specific enough that participants knew when they had successfully satisfied the goal. It is important to let evaluators perform the tasks without assistance, even if they stray from the manufacturer's intended path toward a goal during this phase. Unintended digressions can help evaluators develop an internal model of the interface structure, which can help identify where the user interface structure is non-intuitive or inefficient.

Participants were asked to talk aloud as they performed the tasks, stating their expectations and rationale for the steps they tried. These spoken comments were transcribed by a note-taker/observer in real-time. (In some tests, the experimenter and note-taker were the same person; in other tests, they were separate individuals.) The notes captured the entire discussion, including any dead-ends that the evaluators encountered. The note-taker/observer could also ask for clarifications and/or examples as needed. In general, however, the evaluators were not interrupted.

Tool-Based Review

The high-level tool and the detailed tool were described earlier. A sample version of the high-level tool is shown in Figure 3, and a sample from the detailed tool is shown in Figure 2. The high-level tool was typically completed within one hour. The detailed tool, however, took longer, especially if the system consisted of multiple applications. In earlier evaluations, we varied the order of the tools [2], but in the latest two evaluations, we always presented the high-level tool first because it was the main tool we were assessing. Evaluators were given only one hour to work on the detailed tool, even if they had not finished. Again, evaluators were asked to talk aloud as they worked through the tool, and a note-taker transcribed their comments. Clarifications and/or examples were solicited as necessary.

Feedback on Tools and Wrap-up

The last step in the test was to obtain feedback from the participants on our tools and methods. We used a written questionnaire to structure the comments. Responses to the questionnaire helped us to identify changes to be made to the tool prior to the next evaluation session; aggregate results from the questionnaires are not meaningful so they are not presented here.

DATA ANALYSIS, SYNTHESIS, AND WRITEUP

We collected many pages of notes from each evaluation session. The notes were in two separate sections, one from the task-based exploration phase, and the other from the tool-based review. Either of these sets of notes could be analyzed independently, but the tool-based review produced notes that were easier to use as a starting point. We first collated the notes from the tool-based review across the different evaluation sessions. This produced a file that used the section headings from the tool, with comments on each aspect from every evaluation team below. In practice, evaluators did not proceed through the tool items in order; they often started from one topic and then mentioned other associated topics, but this was not a problem for the data analysis because the overall quantity of data collected from the tool was relatively small (but dense), and related issues could easily be identified.

The first step in processing notes from the task-based exploration was to clean them up by deleting incomplete thoughts, repeated comments, and any other uninformative material. It was then possible to analyze and synthesize the findings by looking across the issues to (1) identify specific difficulties encountered when using the device, (2) look for relationships between the difficulties that were encountered, trying to gather related problems under a single topic heading, and (3) determine problem severity by noting frequency of occurrence, impact, and persistence. Note that relationships between problems and problem severity may become more clear as the findings are drafted and revised into feedback for the manufacturer. For example, a set of error-related issues may appear at first to be unrelated, but may all arise from a single root cause.

Over the course of these evaluations, we developed a standard format for written feedback to the manufacturer. It included an overview of the evaluation protocol and purpose, and a table of contents, which provided an overview of the topics to be discussed. The individual topics were assigned high, medium, or low priorities. High priority issues were those that either (1) violated known FAA regulations and/or guidance, or (2) were global and, in our opinion, had a potentially significant performance impact. Low priority issues were areas we felt could use improvement, but did not appear to have a significant performance impact. The bulk of issues were neither high nor low priority, and so were given a default label of “medium” priority.

In the feedback, we grouped results into topics specific to that EFB. In other words, it was not always appropriate to use generic headings from the tool; doing so can produce feedback that is not always specific enough to act upon. We recommend that feedback be given in terms of functional user interface components. Each topic area began with a

statement of the difficulties encountered, along with information about the frequency of occurrence. *Specific* examples were provided. Where appropriate, we made suggestions for design changes that could address the issue. Often, these were suggestions made by the evaluators during the session, but sometimes they were suggestions based on the synthesized findings across evaluation sessions. In addition, we sometimes provided *observations* (e.g., regarding operational acceptability), which were items that evaluators identified but which may not have a direct human factors impact.

PRELIMINARY FEEDBACK FROM MANUFACTURERS

Although the tools were designed for FAA users, they may also be useful to manufacturers. In fact, some manufacturers are beginning to try out the tools and are considering how to fit the tools into existing design and development processes. Using the tools and methods described here could be a relatively inexpensive way to catch significant problems early on and to track progress on addressing these problems. Preliminary feedback from manufacturers suggests that this is the case. However, it is also clear that the tools are not a substitute for more formal human factors testing.

In particular, there are three shortcomings of the Volpe usability assessment tool. First, the design of an EFB, from the manufacturer's viewpoint, must satisfy not only FAA regulators but also their intended customers. Therefore, testing with end users is necessary, and many operational complexities (e.g., regarding work flow) must be understood in order to optimize the design. Second, the Volpe tool does not provide quantitative results (e.g., time to complete a task, number of steps, or number of errors made). The only quantitative results from using the tool come in the form of frequency of problem occurrence, and those are only available if multiple evaluations are conducted. Quantitative results from usability testing are important for justifying the cost of resources to address problems, and could also be important data for more formal regulatory evaluations (e.g., for installed EFB systems [6]). Finally, results from the Volpe tool highlight problems in the design but do not specify solutions. Additional usability tests will be necessary to choose between design options.

PLANS AND CONCLUSIONS

The EFB usability assessment tool has matured significantly over the past year. Our next step is to expose more potential users, especially those in the FAA, to the tools and methods to determine if these products are useful in practice as designed, or with minor modifications. Note, however, that there is *no* requirement for either the FAA or industry to use these EFB usability assessment tools. The tools and methods will stand on their own merit. If they are useful, they will be adopted, and if they are not useful, they will not be adopted.

In summary, this effort was directed at developing tools and procedures that could be used by FAA field evaluators in conducting structured and comprehensive, yet practical, EFB usability evaluations. Results from several evaluations have yielded a tool and procedure that show great promise. In addition, while these products were designed for EFB evaluations, a good portion of the tools consists of generic guidance that could have much broader applicability.

ACKNOWLEDGMENTS

This work is being conducted by the Volpe Center's Operator Performance and Safety Analysis Division and was funded by the FAA Human Factors Research and Engineering Division (AAR-100). We would like to thank our FAA sponsor, Tom McCloy, as well as the many other FAA staff who have given us feedback and suggestions. Many thanks also to the manufacturers who volunteered their units for testing. We have learned much from you all. And finally, thanks to the participants for their time and valuable suggestions about our evaluation tool and methods. Vic Riley's participation was funded by the Volpe Center (DTRS57-03-P-80181). Thanks to Joan Cahill (Trinity College/Aircraft Management Technologies) and Debbie Matuskevich (Jeppesen) for feedback on the tool.

The views expressed herein are those of the authors and do not necessarily reflect the views of the Volpe National Transportation Systems Center, the Research and Special Programs Administration, or the United States Department of Transportation.

REFERENCES

1. Chandra, D. C. (2002). Human Factors Evaluation of Electronic Flight Bags. *Proceedings of HCI–Aero 2002*, pp. 69–73. Menlo Park, California: AAAI Press.
2. Chandra, D. C. (2003). A tool for structured evaluation of electronic flight bag usability. *Proceedings of the 22nd Digital Avionics Systems Conference*, pp. 13.E.2.1–13.E.2.8. Madison, Wisconsin: Omnipress.
3. Chandra D. C. & S.J. Mangold. (2000). *Human factors considerations in the design and evaluation of electronic flight bags (EFBs) Version 1: Basic functions*. (Report No. DOT-VNTSC-FAA-00-22). Cambridge, MA: USDOT Volpe Center.
4. Chandra, D., Yeh, M., Riley, V., & S.J. Mangold (2003). *Human factors considerations in the design and evaluation of Electronic Flight Bags (EFBs), Version 2*. DOT-VNTSC-FAA-03-07 and DOT/FAA/AR-03/67. Cambridge, MA: USDOT Volpe Center. Available at www.volpe.dot.gov/opsad/efb.
5. Federal Aviation Administration, Advisory Circular AC 25-11. July 16, 1987. *Transport Category Airplane Electronic Display Systems*.
6. Federal Aviation Administration, Advisory Circular AC 120-76A, March 17, 2003. *Guidelines for the certification, airworthiness, and operational approval of electronic flight bag computing devices.*
7. Nielsen, J. & R.L. Mack (Eds.) (1994). *Usability Inspection Methods*, New York, John Wiley & Sons, Inc.
8. Shamo, M. (2000). What is an electronic flight bag, and what is it doing in my cockpit? *Proceedings of HCI–Aero 2000*. Toulouse, France.