

# USABILITY METHODOLOGY APPLIED TO ON-BOARD GRAPHICAL WEATHER DISPLAYS

Kimberly R. Raddatz, John Uhlarik, & Peter D. Elgin  
Kansas State University  
Manhattan, KS

Technology that provides graphical weather is increasing in its availability in the cockpit. However, if considerations are not made regarding the accessibility and display of weather information, the potential exists for that technology to be detrimental. The present research combined components from several different usability methods commonly used in the area of human-computer interaction (i.e., heuristic evaluation, cognitive walkthrough, observation, structured interview, and subjective ratings) to assess and evaluate cockpit weather displays. When applied to graphical weather displays, these usability methods provide important information regarding human factors issues and bottlenecks. Two expert evaluators conducted subjective usability assessments on two commercially available weather displays while six GA pilots participated in user testing. Potential human factors issues or problems leading to sub-optimal performance were identified through the construction of “error” plots using user-testing data from benchmark tasks and classified into heuristic violations. The findings from cognitive walkthroughs, structured interviews, and subjective ratings generally provided explanations for the problems identified through the error plots.

## Introduction

Weather information plays an integral role in supporting the pilot’s ability to strategize during flight planning and supporting tactical cockpit weather decisions that ensure safe and efficient flight. However, if considerations are not made for important human factors issues (e.g., clutter, symbology, color, information currency, etc.) regarding the display of weather information, the presence of weather in the cockpit has the potential to be detrimental, resulting in increased mental workload, pilot confusion, and inappropriate usage. These issues could lead to such consequences as poor weather-related decision-making or even total disregard for weather information.

In an attempt to systematically identify problematic human factors issues inherent in weather display avionics and their consequences, a usability assessment was performed on two commercially available weather avionics systems. To maximize the utility of the results, the usability assessment methodology consisted of components from several different standard usability methodologies that provided both *objective* and *subjective* evaluations. Thus, the goal of this project was to integrate objective and subjective usability evaluation methodologies to more effectively evaluate the system and to more precisely identify and diagnose likely causes of human factors bottlenecks.

Graphical weather information can be shown on a variety of flight deck displays (e.g., flight management systems, multi-function displays (MFDs), externally mounted LCDs). Most avionics systems capable of presenting graphical weather

information do so through a panel-mounted MFD averaging 5 in. diagonally. These MFDs are often capable of presenting other flight relevant information in addition to weather, such as navigation information, airport information, terrain, and traffic. Information selection is typically accomplished through the use of line-select keys (i.e., unlabeled keys next to the interface that correspond to specific menu items on the display) and/or dedicated function keys (i.e., keys whose function is labeled on the key surface and remains invariant regardless of display mode).

Graphics-based displays are capable of presenting a wide range of weather products and information, most commonly graphical weather radar and lightning information (either via data transmission or onboard sensors). Other data-linked graphical weather products for cockpit display include ceilings, visibilities, wind speeds and direction, temperatures and dew points, and NEXRAD. Typically, graphical weather displays rely on unique symbology and/or different colors to indicate severity and/or intensity of weather. Most weather products may be displayed either individually or overlaid with other non-weather information (e.g., terrain).

### *Usability Evaluation Background*

Usability refers to how easy it is for users to learn a system, how efficiently users can use the system once it is learned, and how pleasant the system is to use (Mack & Nielsen, 1994). Usability is one of the core constructs in human-computer interaction research (Gray & Salzman, 1998). The general goal of

usability is to identify potential design problems and suggest guidelines to avoid these problems. A usability problem can be defined as “an aspect of the user interface that may cause the resulting system to have reduced usability for the end user” (Mack & Nielsen, 1994, p. 3).

Although various usability evaluation methods have been promoted as ideal or optimal tools for evaluating and improving interfaces, each method has advantages and disadvantages (cf., Gray & Salzman, 1998). As a result, the present usability assessment integrates several different types of usability methods. Examples and brief descriptions of each method follow.

- **Inspections:** Expert evaluators (EE) assess each system in terms of specific end-user tasks (e.g., displaying weather information) and unstructured exploration.
- **Heuristic Evaluation:** A small number of EEs examine the interface and judge its compliance with established usability principles, called heuristics (Nielsen, 1994).
- **Contextual Observation/Interview:** Researchers monitor user performance in a simulated work environment, observing how the user performs specific tasks, probing the users for details regarding their underlying goals, and eliciting users’ thought processes and actions while performing the tasks.
- **Cognitive Walkthrough:** For a given task, the EE investigates, in sequence, each of the steps necessary to perform the task and attempts to uncover design errors that would interfere with learning by exploration (Wharton, Rieman, Lewis, & Polson, 1994).

Specific elements of these four usability methods were employed throughout the three phases of the present assessment: *pre-usability inspection*, *usability testing*, and *post-usability inspection*. Pre- and post-usability inspections are based on rules of thumb and the general skills, knowledge and experience of the trained EEs, and consequently, yield subjective evaluation data based on expert opinion. Typically, evaluators have some experience with the system domain, human factors and usability guidelines, user testing, and user interfaces (Mack & Nielsen, 1994). For the present assessment, EEs were human factors specialists with considerable knowledge of weather avionics but no pilot experience. Conversely, usability testing affords empirical evaluations by testing the system with real end-users ecologically valid tasks, and consequently, yields objective data. Past research has shown that usability inspection methods discover many problems typically

overlooked by user testing, and that user testing identifies problems that are overlooked by usability inspections (Mack & Nielsen, 1994). Therefore, to achieve the most comprehensive and meaningful results, the present usability assessment combined usability testing and inspection methodologies. Six IFR-rated instructor pilots were used as subject matter experts for the user-testing phase, while two human factors specialists served as EEs during the pre- and post-usability inspections.

## Usability Assessment Methodology

### *Pre-Usability Inspections*

Pre-usability inspections are typically used to identify problems in a prototypical interface design and then make recommendations for improving the design (Mack & Nielsen, 1994). The first goal of the pre-usability inspection was to familiarize the EEs with the functions and capabilities of each system so that benchmark tasks could be generated for the user-testing phase. Benchmark tasks were ecologically valid tasks explicitly pertaining to weather information display or to display formatting. Several of the tasks were supported by both systems, allowing for comparison between the systems. A list of the benchmark tasks can be found in Figure 1. The EEs also identified the least number of actions needed to complete each benchmark task (i.e., the “gold standard” of performance). Thus, the actions initiated by pilots when completing each task during the user-testing phase can be compared to the gold standard for that task.

The second goal of the pre-usability inspection phase was to identify problems within the system from a perspective that incorporated human factors principles. The EEs evaluated each system in terms of specific tasks pertaining to the display of weather information and in terms of unstructured exploration of the system (i.e., inspections). The EEs also judged system compliance with established usability principles, called heuristics (i.e., heuristic evaluation). Heuristics were specifically defined categories of display, design, and user issues. This list of heuristics originated from several established human factors and usability sources, including Nielsen’s (1994) factor analysis of 249 usability problems and the 13 principles of display design (Wickens, Gordon, & Liu, 1998). In addition to these more general heuristics, specific heuristics were developed as supplements when needed to categorize usability issues specific to displaying weather.

## Usability Testing

The usability-testing phase assessed how well targeted end-users (i.e., GA pilots) were able to learn and utilize a system. Each system was configured as a stand-alone piece of equipment running demonstration programs that used simulated weather data. Pilots were asked to perform realistic information-gathering tasks while the EEs observed their performance. One EE sat next to the pilot, provided instructions before each task, probed pilots for details regarding underlying goals, and elicited pilot thought processes and actions while performing the tasks (i.e., contextual observation/interview). The other EE videotaped each session. From the videotapes, EEs recorded every action initiated by the pilots and the time it took to complete each task. Each input action performed during the user-testing phase was classified either as an *efficient action* (i.e., consistent with the task completion that required the least number of actions – the gold standard) or an *inefficient action* (i.e., unnecessary or inconsistent with the gold standard).

After usability testing for a system was completed, a structured interview was conducted to elicit comments and feedback about issues specifically regarding the efficacy of the system's interface. Each user-testing session took an average of three hrs to complete.

## Post-Usability Inspection

During the post-usability inspection, the EEs identified reasons why inefficient actions were performed and suggested guidelines for improving system usability. During this phase, the EEs transcribed all pilot comments and feedback elicited through debriefing and discussion questions posed at the conclusion of the user-testing phase. Then, for a given task, the researcher examined, in sequence, each of the steps necessary to perform the task. For each inefficient action, EEs assessed cue validity by asking two questions: 1) why wasn't the efficient action initiated; and 2) why was the specific inefficient action chosen? Thus, by answering these two questions, the EEs gained insight into why the pilot chose the inefficient action (i.e., cognitive walkthrough). In addition, the EEs classified the inefficient actions as violations of one or more display heuristics (i.e., heuristic evaluation).

## Information Accessibility Results

The goal of this assessment was to identify problematic human factors and usability issues in current weather avionics and suggest guidelines for resolving these issues. These issues were divided into two categories: 1) those that influence information accessibility, and 2) those that influence pilots' interpretation and/or subjective assessments of the weather information. Information accessibility results were based on three analyses of the user-testing data: 1) empirical analysis was conducted to identify those benchmark tasks that pilots had difficulty completing; 2) for each benchmark task, the problematic action sequences (i.e., bottlenecks) were identified; and 3) bottlenecks were explained in terms of violation(s) of general (theoretical) display heuristics. These three analyses provided the basis for a general list of guidelines for eliminating human factors issues and bottlenecks.

## Identification of Problematic Benchmark Tasks

Inefficient actions. The average number of inefficient actions for a benchmark task was calculated by subtracting the number of actions comprising the gold standard for that task from the average number of input actions initiated by pilots during user-testing. In general, more inefficient actions denoted the presence of usability issues/problems.

Task completion time. Task completion time was defined as the duration of elapsed time between the pilot's initiation of the first action and task completion (Figure 1). Longer task completion times denoted the presence of usability issues/problems, especially when controlling for the number of required actions. For example, displaying graphical ceilings and visibilities in System B required only two input actions but task completion time was 200 sec.

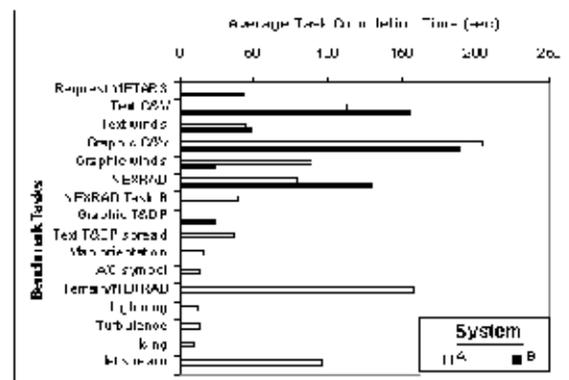


Figure 1. Average completion time per task.

### Identifying Cause(s) of Usability Issues.

Objective data such as task completion time and average number of inefficient actions per task were used to identify which tasks posed overall problems for the pilots. For tasks that elicited many inefficient actions and relatively long completion times, error plots were used to further identify exactly where in the system the bottlenecks occurred. Subjective data were used to diagnose why the bottlenecks occurred.

For each problematic benchmark task, error plots were created. Error plots represent the average number of inefficient actions performed in between each gold standard action (see Figure 2). The gold standard represents the optimal sequence of actions necessary to perform each task. Thus, optimal performance would be depicted as one input action for each required action sequence (denoted as the dashed line in Figure 2). The presence of a large number of inefficient actions between specific required actions represents pilot misunderstanding and/or confusion (i.e., a bottleneck in the system). Therefore, for the task displayed in Figure 2, (i.e., display graphical ceilings and visibilities for an area within 200 NM of current position), the error plot indicates the presence of bottlenecks when pilots transitioned between actions 4 and 5 and between actions 1 and 2 (denoted by the circles).

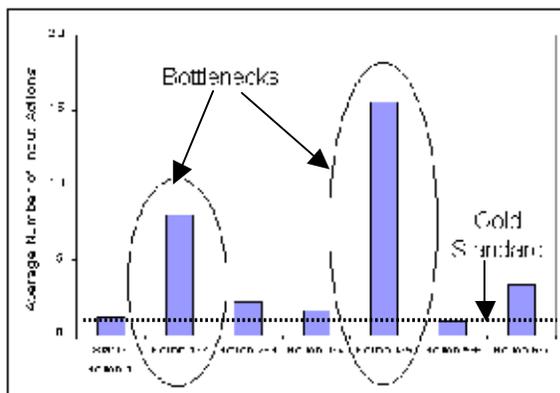


Figure 2. Example of an error plot (for displaying ceilings and visibilities within 200 NM of current position) showing the average number of initiated actions relative to the gold standard (depicted as the dashed line).

Thus, objective data were used to identify *where* bottlenecks occurred in the system. However, subjective data provided the information necessary to reveal *why* the bottlenecks occurred and suggested how they could be attenuated. For example, as previously noted, the error plot for the ceilings and visibilities task in Figure 2 identifies two potential

bottlenecks. The information gained during the pre-usability inspection phase (through inspection and heuristic evaluation), indicated that transitioning between action 1 and 2 on this system required the pilot to activate the “METAR” option to display graphical MET ARs. Contextual observation and pilot comments revealed that most pilots interpreted the “METAR” option to mean “text weather” and therefore it was not immediately apparent that the “METAR” option was the appropriate one for displaying graphical weather. Similarly, transitioning between action 4 and 5 required the pilot to activate the “200 NM display range” option, which was a hidden option not listed on the display range menu. This required the pilot to know that in order to access the “200 NM display range” option, they must first activate the “USNATL” range option. Thus, knowledge that the EEs gained from the subjective usability evaluations was essential to diagnose *why* the bottlenecks occurred.

### Classifying Bottlenecks in Terms of Violated Heuristics

A final analysis explained the bottlenecks (identified in the user-testing phase) as violations of one or more established display heuristics. This analysis was performed on two separate data sets. First, inefficient actions performed during the user-testing session were classified as the result of one or more violated design heuristics. Second, bottlenecks identified by error plots were classified as the result of one or more violated heuristics. The following are examples of heuristics most frequently used to classify human factors issues and bottlenecks:

- Visibility of system status. The display should keep pilots informed about the status of the system through timely feedback.
- Consistency and standards. There should be consistency in the symbology, terminology, and actions used throughout the display. Also, the general format of a display should not change in different display modes unless the changes are explicitly used to draw the attention toward some display variable.
- Recognition rather than recall. Objects, actions, and options in the display should be visible at all appropriate times. Pilots should not be responsible for remembering important information.
- Descriptive labeling. There should be enough description inherent in the labeling to specifically indicate the function and/or outcome of an option. Also, all labeling should be salient

on the screen (in terms of size, position, and/or color, etc.).

- Ambiguous symbology. The meaning of the symbology should be intuitive. The depiction of the symbology should be salient/noticeable (i.e., not easily overlooked) and easily discriminable from other types of symbology.

## **Pilots' Interpretations and Subjective Assessment Results**

Several different methods were used to collect pilots' interpretations, subjective assessments and comments about the displayed weather information. The methods included the following:

- § Validity of the interpretation: Upon completion of each benchmark task, pilots were asked to interpret the meaning of the displayed weather information. Each interpretation was classified as either *correct* or *incorrect* as defined by the system's user manual. Possible reasons for misinterpretation were classified in terms of violated display heuristics.
- § Questionnaire Results: Each pilot completed questionnaires regarding specific characteristics of and reactions to each system. The questionnaire was divided into four parts: 1) pilot's general response or reaction to the system, 2) training and ease of use of the system, 3) system information layout, and 4) weather information and content. In addition, pilots were asked general questions about their experience with and/or potential use of weather information in the cockpit. These questions were not specific to either system.
- § Pilot Comment and Feedback: Throughout the usability-testing phase, pilots were encouraged to verbalize their actions and thoughts regarding their interaction with each system.

Results based on these analyses were integrated with the results from the information access analyses to yield a more extensive list of human factors issues regarding the display of weather in the cockpit.

## **Partial List of Guidelines**

The following is a partial list of guidelines generated from the results of the usability assessment. Each guideline includes a brief description of a specific incident from the usability assessment.

### System organization must make sense.

- § The "WX" label on a menu should allow access to *all* types of weather information through one avenue or another. Deviations from intuitive organization (e.g., organizing text weather information under an "INF" option instead of a "WX" option) require even more descriptive labeling to guarantee the system organization is interpreted correctly.

Descriptive labeling. The label must intuitively indicate menu option and/or input device function.

- § In order to display graphical weather in System A, pilots must press the "EXIT" key. However, the term "EXIT" is not indicative of displaying graphical weather.

Hidden functions should be avoided, especially if the system organization does not follow pilots' logical expectations.

- § The only way to obtain a 200 NM view of ceilings and visibilities on System A is to activate the "USNATL" menu option. The "200 NM display range" option does not exist on the display range menu list (i.e., it can only be accessed through activating the "USNATL" display option).

Hidden function shortcut keys should not be the only way to retrieve certain information.

- § The only way to display graphical weather information on System B is to repeatedly press the "ENT" key, which is not labeled to support this task.

Consistency and standards. The meaning of label terminology should not change throughout the display.

- § In System A, the "EXIT" menu option sometimes cancels previous inputs/actions; other times it accepts previous inputs/actions; other times it actually displays graphical weather information.

Top-down processing. Display design and organization should account for pilots' logical expectations and capitalize on pilots' past experience. However, where the design deviates from pilot expectations, extra care needs to be taken to document this change.

- § Because of their previous experience, most pilots preferred to see weather information in text form. However, this does not necessarily mean weather information should not be displayed graphically. Rather, the symbology and the labeling of the graphical display must be

intuitive and meaningful, helping pilots overcome expectations and biases developed from prior experience (i.e., reduce negative transfer of training).

Incomplete information. The display should provide access to all weather parameters that are needed to make informed and safe weather decisions.

§ Both systems displayed ambient and dew point temperature spread information only. However, every pilot stressed the importance of knowing actual ambient and dew point temperatures, in addition to their spread.

BACK/UNDO function. Displays should always support some type of BACK/UNDO function.

§ Neither system provided a specifically labeled "BACK" button for pilots to use when they wanted to return to a previous screen.

## Summary

The overall goal of the avionics usability assessment was to develop a list of design guidelines for use by FAA certification specialists and by avionics vendors during the design process. The complete list of guidelines is intended to address issues related to the current state-of-the-art technology used for cockpit weather displays. Several usability assessment methodologies were incorporated to identify human factors issues and problems leading to sub-optimal pilot-system interaction in two currently available weather avionics systems. This usability assessment used objective data (e.g., inefficient actions, task completion time, and error plots resulting from the user-testing phase) to identify *where* bottlenecks occurred in the system. Subjective data (e.g., heuristic evaluation, inspection, and pilot comments and feedback from the pre- and post usability inspection phases) were used to diagnose *why* the bottlenecks occurred. Specific usability issues and bottlenecks identified through this assessment have implications to avionics displays in general as well. Based on the results of the usability assessment, guidelines were suggested to aid the general design and/or certification process of graphical weather avionics.

## Acknowledgements and Notes

Support for this research was provided by the U.S. Department of Transportation - Federal Aviation Administration (FAA) [Contract DTFA-02-02-R-03491]. The authors would like to acknowledge Colleen Donovan and Dr. Kevin Williams for their input and feedback concerning this project.

A more extensive version of this report is available from:

John Uhlarik  
Department of Psychology  
1100 Mid-Campus Drive  
492 Bluemont Hall  
Manhattan, KS 66506  
or  
email: uhlarik@ksu.edu.

## References

- Gray, W.D., & Salzman, M.C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction, 13*, 203-261.
- Mack, R.L. & Nielsen, J. (1994). Executive summary. In J. Nielsen & R.L. Mack (Eds.), *Usability inspection methods* (pp. 1-23). New York: John Wiley & Sons.
- Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen & R.L. Mack (Eds.), *Usability inspection methods* (pp. 25-62). New York: John Wiley & Sons.
- Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The cognitive walkthrough method: A practitioner's guide. In J. Nielsen & R.L. Mack (Eds.), *Usability inspection methods* (pp. 105-140). New York: John Wiley & Sons.
- Wickens, C.D., Gordon, S.E., & Liu, Y. (1998). *An introduction to human factors engineering*. New York: Addison Wesley Longman.