VALIDATING THE USER-CENTERED HYBRID ASSESSMENT TOOL (USER-CHAT): A COMPARATIVE USABILITY EVALUATION

by

PETER D. ELGIN

B.A., Loras College, 1996
M.S., Drake University, 1999

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychology
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2007

Abstract

Usability practitioners need effective usability assessment techniques in order to facilitate development of usable consumer products. Many usability evaluation methods have been promoted as the ideal. Few, however, fulfill expectations concerning effectiveness. Additionally, lack of empirical data forces usability practitioners to rely on personal judgments and/or anecdotal statements when deciding upon which usability method best suits their needs. Therefore the present study had two principal objectives: (1) to validate a hybrid usability technique that identifies important and ignores inconsequential usability problems, and (2) to provide empirical performance data for several usability protocols on a variety of contemporary comparative metrics. The User-Centered Hybrid Assessment Tool (User-CHAT) was developed to maximize efficient diagnosis of usability issues from a behaviorally-based perspective while minimizing time and resource limitations typically associated with usability assessment environments. Several characteristics of user-testing, the heuristic evaluation, and the cognitive walkthrough were combined to create the User-CHAT. Prior research has demonstrated that the User-CHAT supports an evaluation within 3-4 hrs, can be used by individuals with limited human factors / usability background, and requires little training to be used competently, even for complex systems. A state-of-the-art suite of avionics displays and a series of benchmark tasks provided the context where the User-CHAT's performance was measured relative to its parent usability methods. Two techniques generated comparison lists of usability problems – user-testing data and various inclusion criteria for usability problems identified by the User-CHAT, heuristic evaluation, and cognitive walkthrough. Overall the results demonstrated that the User-CHAT attained higher effectiveness scores than the heuristic evaluation and cognitive walkthrough, suggesting that it helped evaluators identify many usability problems that actually impact users, i.e., higher thoroughness, while attenuating time and effort on issues that were not important, i.e., higher validity. Furthermore, the User-CHAT had the greatest proportion of usability problems that were rated as serious, i.e., usability issues that hinder performance and compromise safety. The User-CHAT's performance suggests that it is an appropriate usability technique to implement into the product development lifecycle. Limitations and future research directions are discussed.

VALIDATING THE USER-CENTERED HYBRID ASSESSMENT TOOL (USER-CHAT): A COMPARATIVE USABILITY EVALUATION

by

PETER D. ELGIN

B.A., Loras College, 1996
M.S., Drake University, 1999

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychology
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2007

Approved by:

Major Professor
John Uhlarik, Ph.D.

Copyright

PETER D. ELGIN

2007

Abstract

Usability practitioners need effective usability assessment techniques in order to facilitate development of usable consumer products. Many usability evaluation methods have been promoted as the ideal. Few, however, fulfill expectations concerning effectiveness. Additionally, lack of empirical data forces usability practitioners to rely on personal judgments and/or anecdotal statements when deciding upon which usability method best suits their needs. Therefore the present study had two principal objectives: (1) to validate a hybrid usability technique that identifies important and ignores inconsequential usability problems, and (2) to provide empirical performance data for several usability protocols on a variety of contemporary comparative metrics. The User-Centered Hybrid Assessment Tool (User-CHAT) was developed to maximize efficient diagnosis of usability issues from a behaviorally-based perspective while minimizing time and resource limitations typically associated with usability assessment environments. Several characteristics of user-testing, the heuristic evaluation, and the cognitive walkthrough were combined to create the User-CHAT. Prior research has demonstrated that the User-CHAT supports an evaluation within 3-4 hrs, can be used by individuals with limited human factors / usability background, and requires little training to be used competently, even for complex systems. A state-of-the-art suite of avionics displays and a series of benchmark tasks provided the context where the User-CHAT's performance was measured relative to its parent usability methods. Two techniques generated comparison lists of usability problems – user-testing data and various inclusion criteria for usability problems identified by the User-CHAT, heuristic evaluation, and cognitive walkthrough. Overall the results demonstrated that the User-CHAT attained higher effectiveness scores than the heuristic evaluation and cognitive walkthrough, suggesting that it helped evaluators identify many usability problems that actually impact users, i.e., higher thoroughness, while attenuating time and effort on issues that were not important, i.e., higher validity. Furthermore, the User-CHAT had the greatest proportion of usability problems that were rated as serious, i.e., usability issues that hinder performance and compromise safety. The User-CHAT's performance suggests that it is an appropriate usability technique to implement into the product development lifecycle. Limitations and future research directions are discussed.

Table of Contents

List of Tables

List of Figures

List of Acronyms

ACO	Aircraft Certification Office
FAA	Federal Aviation Administration
FGCP	Flight Guidance Control Panel
GA	General Aviation
HCI	Human Computer Interaction
HF	Human Factors
HSI	Horizontal Situation Indicator
IFR	Instrument Flight Rules
MFD	Multi-function Display
PFD	Primary Flight Display
UEM	Usability Evaluation Method
User-CHAT	User-Centered Hybrid Assessment Tool
VFR	Visual Flight Rules

Acknowledgments

I would like to gratefully acknowledge the many individuals who have helped me along throughout the dissertation process. First, I want to thank my wife, Michelle, for her continued and unwavering support. On days when the project seemed insurmountable and there appeared to be no end in sight, she provided much needed and appreciated encouragement and motivation. When I needed her the most, she was there. I am eternally blessed to have her in my life.

I would like to express sincere gratitude to my dissertation committee members, Drs. John Uhlarik (major advisor), James Shanteau, Richard Harris, and Doug Goodin. Their guidance and insightful feedback has proved to be invaluable. Special thanks go to Dr. Uhlarik for providing direction with the project and helping me see the 'big picture' when I was engrained and swamped with data analysis. Gratitude is expressed also to the FAA, especially Tom McCloy, Colleen Donovan, and Kevin Williams who were instrumental in making this line of research possible over the last few years. Their support provided a unique opportunity to acquire valuable real world usability experience and allowed me to hone my skills as a human factors / usability professional.

Graduate school is much more enjoyable when one is surrounded by exceptional graduate students. I learned a great deal from my fellow graduate students – I will miss our interactions. In particular, much appreciation is expressed to, Shawn Farris, Tuan Tran, and Jason Ward, who assisted in my dissertation with data collection and analysis. I would like to especially thank Kim Raddatz. Over the past few years, Kim and I have worked together on many research projects (grant and non-grant related) and have debated on a variety of psychological topics. I am thankful for her astute commentary, which contributed substantially to my dissertation. I am grateful for her friendship and wish her well in her future endeavors.

Finally, my family has been a pillar of support. To my parents, Dan and Grace, my brother Mike, and my two sisters, Sara and Molly, thank you for all you have done for me and for not giving up on me. I hope I can reciprocate the support they have given me.

Dedication

To Michelle and Parker

Chapter 1 -- Introduction

As manufacturers continue to deal with increasing competitive pressures, the need for usable commercial products becomes more evident (Desuvire, 1994). As a result, there are growing demands on human-computer interaction (HCI) professionals to develop effective usability assessment methodologies (Andre, 2000) in order to facilitate the identification and resolution of usability issues in a product's interface. While many contemporary usability evaluation techniques have been championed as the ideal (Gray & Salzman, 1998) and promise to offer large cost savings (Desuvire, 1994), their acceptance as an integral component to the product development lifecycle has been met with some resistance. Nielsen (1994a) postulates that this resistance is attributed to the perception that many usability techniques are too intimidating, too expensive, and/or too difficult and time consuming to use. In fact, Ivory and Hearst (2001) contend that contemporary usability methodologies are often perceived as effort-intensive activities with limited returns on investment. In order to make usability assessments more accessible and desirable for manufacturers, Nielsen (1994b) advocated the development of "discount" usability methods that are inexpensive, fast, and easy to use. While the notion of creating discount usability methods sounds appealing, many of these techniques, however, fall short in terms of overall effectiveness. Therefore, current demand necessitates continued research that strives for developing usability assessment protocols that maximize returns on investments.

Additionally, usability practitioners do not have adequate empirical data available to decide which usability method is most appropriate or of optimal relevance to their specific product assessment environment (Andre, Williges, & Hartson, 1999). As a result, many practitioners must rely on anecdotal evidence or personal judgments and/or experience to decide which usability method is appropriate. Empirically-derived validation data made available for the multitude of usability techniques can aid decisions for determining which method is most effective and most suitable given the practitioner's needs and goals (e.g., detecting lots of usability problems that impact system performance). However, few comparison studies report appropriate validation measures for most usability techniques (e.g., Andre, 2000). Consequently an empirical comparison study is needed to add to the growing HCI literature of validation metrics for usability assessment methods.

While the demand for effective usability methods and validation data is shared by usability practitioners in all product industries, according to Clamann and Kaber (2004), human factors professionals associated with the aviation industry have an especially acute need for efficient usability method(s) to incorporate into the avionics assessment environment. Aviation usability practitioners, namely those charged with ensuring that a system is safe for flight operation, have the added responsibility of assessing highly complex avionics systems used in dynamic environments where operator error can be disastrous[1]. This need is exacerbated due to recent technological advances in data-linked information, graphical computing capabilities, and the capacity to support the display of many types of information previously unavailable in cockpits. The introduction of additional flight-relevant information (e.g., text and graphic weather, moving maps, terrain, traffic, flight planning, etc.) has the potential to increase pilot awareness, flight safety and efficiency by supporting strategic planning and improving risk assessments. However, if important human factors principles and issues are not addressed, the presence of all of this information can be detrimental (Clamann & Kaber, 2004). For instance, lack of concern for such elemental issues as top-down processing, display clutter, symbols and color usage, menu structure, and information validity have the potential to result in confusion and misunderstanding, possibly leading to unsafe decision-making or complete disregard of the displayed information. Thus, it is important that the certification process for any complex candidate system assess important human factors issues regarding new technologies, ensuring safe yet efficient flight.

The Federal Aviation Administration (FAA) is the governing body that certifies all avionics displays mounted in a cockpit's flightdeck. The purpose of certification is to evaluate candidate systems, ensuring that the system complies with FAA regulations (i.e., minima) and that the system does not adversely affect pilot performance and safety. Additionally, the FAA ensures that candidate systems are usable by the target user population (e.g., transport, general or commercial aviation pilots).

[1] While the remaining context focuses primarily on usability assessment issues associated with the aviation industry, specifically the process of determining whether a candidate avionics system is certifiable (i.e., safe for use in cockpits), many aspects of the certification process may mirror usability evaluation environments in other domains. Usability practitioners operating in non-aviation industries, e.g., website design, software development, mobile devices, construction equipment, etc. may relate their own usability assessment setting to the constraints of the certification process and may readily draw upon and apply many discussion points to their specific situation. That is, while the remaining scope is limited to the certification process for candidate avionics systems, many of the ideas and conclusions can be easily applied by human factors professionals to other usability evaluation environments.

Utilizing a variety of assessment environments (e.g., bench-top test, simulator, in a plane on the ground or in the air), flight test engineers, who are skilled pilots, often differ in the procedures by which they certify avionics systems. Some flight test engineers familiarize themselves with the functionality of the candidate system, its information organization and display symbology by systematically initiating every menu option and input action; others perform a series of tasks (e.g., insert a waypoint into a flight plan, display a specific weather product, etc.) intended to exercise the various functions of the candidate system. Whatever the evaluation strategy, the goal of a flight test engineer is the same – to gather enough information and evidence in order to determine whether the candidate system conforms to FAA design guidelines and certification regulations and to ensure that the system is functioning properly. Once it is documented that a candidate system complies with these standards, the system is then certified.

Typically the goal of human factors practitioners in the aviation industry (or any other industry) is typically to optimize the human-machine interaction. However, the certification environment employs human factors principles and practices for goals with slightly different emphasis. Rather than focusing on how to optimize the human-machine relationship, the FAA is primarily concerned with the extent to which minimum human factors standards are satisfied. Assessing compliance using minima criteria differs from assessing compliance using optima criteria. Minima (e.g., requirements) are FAA regulations that must be satisfied in their entirety in order for a system to be certified. Violating any minimum regulation indicates non-compliance and can thereby justify non-certification. Minima represent issues that, if not addressed, may compromise pilot performance and safety (e.g., misinterpreting information, causing disorientation/confusion, increasing workload, etc.).

Conversely, optima (e.g., recommendations) are FAA guidelines that are intended to improve the human-machine interaction, but whose violations do not necessarily have negative consequences. Thus, while it is highly desirable to optimize an interface, optimization is not necessary from the FAA's perspective. The FAA mandate of minima over optimal is conveyed systematically as the difference between "thou shall" and "thou should." For instance, while the FAA supports the use of color in order to optimize an interface, when information is color-coded, the FAA has minima concerning how specific meaning shall be conveyed through the use of color (e.g., red denotes warnings while yellow/amber denotes cautions).

Certification Environment Constraints

While the complexity and diversity with which information can be displayed in the cockpit is rapidly increasing, the FAA's certification procedures for assessing complex systems have not advanced at the same pace. As avionics technologies continue to advance, the need to modify the current certification process for evaluating these highly complex systems becomes crucial. In addition, the characteristics of the certification environment itself present unique challenges and constraints for the procedure to be used in the evaluation process. Therefore, efforts to update the certification process must recognize four major constraints to ensure that modifications are both constructive and appropriate.

One omnipresent constraining characteristic of the certification environment is the limited availability of FAA time and personnel available to assess the candidate system. As the complexity of candidate systems increase, the complexity of the certification process increases also. For instance, the time and effort required for evaluating a complex multi-functional display (MFD) is considerably greater than for a simple, single function "round dial" display, largely because of differences in the amount of functionality that must be assessed. As most FAA Aircraft Certification Offices (ACOs) will attest, time and personnel are limited commodities. Therefore, certification personnel would benefit from a tool that enables them to perform a comprehensive evaluation of system functionality while requiring minimal time and effort.

Secondly, evaluations currently performed by certification officials are often based on subjective evaluations of the candidate system and interpretations of the meaning and importance of FAA regulations. The subjectivity of the evaluations often results in contentious interactions with manufacturers. Often missing from evaluations are objective criteria that provide indisputable justification of identified human factors issues and demonstrable evidence of their importance. While the judgment, expertise, and reasoning ability of the ACO flight test engineer(s) is certainly invaluable, the ability to also generate objective data during certification would be beneficial, especially for the ensuing discussions between the FAA and manufacturers regarding those issues that pose significant obstacles to certification. Specifically, if there is disagreement between the FAA and the manufacturers regarding the results of the evaluation or the importance of identified certification issues, objective performance-based data could be used to justify and validate disputed issues as actual human factors problems. Thus, the

certification process would benefit from protocols that supplement evaluators' judgments with performance-based data that highlight the frequency and severity of identified human factors / certification issues.

Third, as noted earlier, the responsibility for evaluating a candidate system typically falls upon flight test engineers. While the flight test engineer may approximate the targeted user population, he/she may lack the human factors background necessary to proficiently conduct an evaluation or more appropriately, a usability assessment. Nevertheless, the flight test engineer, who may lack the necessary background, is asked to conduct what is essentially a human factors evaluation[2]. The lack of human factors knowledge may be exacerbated when evaluating complex systems such a MFDs under time pressure (i.e., some important human factors issues may be overlooked). Therefore, the certification process could be improved with tools that enable individuals with limited human factors background to adeptly conduct a human factors evaluation.

Lastly, the variability in the current certification process and the lack of human factors background in most flight test engineers have both contributed to a process that lacks standardization across ACOs. An inconsistent certification protocol results in considerable latitude and variation regarding the level of detail and the thoroughness of an evaluation. Standardizing the certification process would ensure that the FAA and the manufacturers are in agreement regarding the criteria against which the candidate system are evaluated and the degree of detail of the evaluation. Explicit standardization of the certification process can provide a greater degree of predictability for manufacturers regarding the characteristics of their system that must be certified.

Considering these aforementioned issues[3], the certification process would benefit substantially from the development of a certification procedure that effectively supports individuals with limited human factors backgrounds to conduct an objective performance-based evaluation of a complex candidate avionics system without increasing the time and personnel demands. If such a protocol were developed

[2] Some ACOs are fortunate to have a human factors professional on staff and available during the certification process. However, this is the exception rather than the rule across ACOs.
[3] The characteristics associated with the certification environment may describe constraints common to usability assessment situations in other industries (e.g., limited time and personnel resources, un-standardized evaluation process, etc.). Thus, usability practitioners outside of the FAA may recognize many restrictions in their own assessment settings that are inherent with the certification process.

and implemented by the FAA for use in all ACOs, then the variability with which the certification process is applied across certification offices would be reduced substantially.

The HCI literature identifies many usability evaluation methods (UEM) as possible techniques to evaluate complex systems during certification. The most prevalent UEMs include *user-testing*, the *cognitive walkthrough* and the *heuristic evaluation*. These UEMs are described and discussed later. However, a review of these traditional usability methods reveals that, while each provided specific useful components that would satisfy some of the aforementioned certification environment constraints, none of the UEMs were deemed appropriate for satisfying all the constraints. Further, applying multiple UEMs in their entirety to the certification process would obviously be impractical, considering the certification constraints. Therefore, the technique deemed most suitable is one that integrates unique components gleaned from user-testing, the cognitive walkthrough, and the heuristic evaluation that appropriately satisfies the certification constraints.

Several years of FAA-funded research have led to the development of such a technique – the User-Centered Hybrid Assessment Tool (User-CHAT) (Uhlarik, Raddatz, & Elgin, 2002; 2003; Uhlarik, Elgin, & Raddatz, 2004). The User-CHAT is intended for relatively quick and easy use like the heuristic evaluation, offers performance-based data from target end-users completing representative tasks to substantiate identified usability problems like user-testing, provides structure to the assessment and supplies diagnostic information like the cognitive walkthrough and supports efficient evaluations by individuals with limited human factors backgrounds. Research has been conducted to evaluate the User-CHAT as an appropriate usability method that satisfies the constraints imposed by the current FAA certification environment (e.g., Uhlarik et al., 2004). However, in order to fully realize the potential of the User-CHAT as the preferred technique for cost-effectively identifying human factors issues that may compromise certification of a candidate system, a comparative usability evaluation study needs to be conducted whereby the performance of the User-CHAT is directly compared to performances of more established usability methods like the heuristic evaluation, cognitive walkthrough, and user-testing. Therefore the purpose of the present study is to compare the amount and type of usability problems the User-CHAT detects relative to the amount and type of usability problems found through user-testing, the heuristic evaluation, and the cognitive walkthrough.

Certification and Usability

The process of evaluating the human factors issues associated with candidate avionics systems during the FAA certification process is analogous to conducting a usability assessment. Usability is one of the core constructs in human factors and HCI research (Gray & Salzman, 1998). The general goal of usability is to identify design problems and suggest alternatives for resolving these problems. A usability problem is typically 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).

Usability assessment is a standard term that signifies the evaluation of an interface by trained usability specialists with or without the input of representative end users (Mack & Nielsen, 1994). Usability metrics can provide quantitative and/or qualitative data regarding a user's experience when interacting with a system. A variety of dimensions can be used to assess the degree of usability for an interface (Wickens, Gordon, & Liu, 1998).

1. *Ease of Learning*: refers to the speed with which novices can learn to use a system.
2. *Ease of Use*: refers to the degree to which the system supports user performance once the user has become familiar with it.
3. *Ease of Remembering*: refers to how well the user remembers how to use the system after a long absence.
4. *Frequency of Errors*: refers to the number of errors the user commits when interacting with the system.
5. *Overall Subjective Satisfaction*: refers to the user's subjective experience when interacting with the system; users who enjoy using the system will continue to use it in the future.

In order for systems to achieve high degrees of usability (e.g., easy to learn, easy to use, few errors, etc.), "at any given time, users should understand what is being presented, what they are required to do or have the option of doing, what they must do to accomplish their current goal, what would happen if they choose a particular option, and what the system is currently doing" (Wickens et al., 1998, p. 463). Through usability assessments, bottlenecks in the human-machine interface can be identified, analyzed, and resolved.

Usability Evaluation Methods

Based on the nature of the data accumulated during the assessments, UEMs are commonly grouped into two categories: user-testing and usability inspection methods. User-testing is a UEM where

researchers monitor representative end-users completing a series of specific tasks in a simulated work environment. This assessment typically entails observing user performance, probing the users for details regarding their underlying goals, and eliciting users' thought processes and actions while performing the tasks (Mayhew, 1999). A usability inspection method, on the other hand, is a generic name for UEMs in which evaluators themselves inspect the usability-related aspects of a system without the input of representative end-users. The defining characteristic of usability inspection methods (e.g., heuristic evaluation, cognitive walkthrough, etc.) is that they require expert evaluator judgments when assessing a system (Dutt, Johnson, & Johnson, 1994; Mack & Nielsen, 1994). Past research has shown that usability inspection methods discover many problems typically overlooked by user testing and vice versa (Mack & Nielsen, 1994). Although various UEMs have been promoted as ideal or optimal tools for evaluating and improving interfaces, each method has advantages and disadvantages (Gray & Salzman, 1998). However, regardless of whether the UEM is user-based or expert-based, the goal of each is the same – to improve the overall usability of a system (Rosenbaum, 1989).

Because the present research compares the differential efficacies of several usability inspection methods, user-testing, and the User-CHAT as potential techniques for certification assessment, a quick review of usability inspection methods and user-testing is appropriate. Also this review identifies the distinctive components from each methodology that provide the foundation for the User-CHAT.

User-testing

A longstanding and popular UEM is user-testing (Tan & Bishu, 2002). User-testing is essentially a research tool, with roots in traditional methodologies of experimental psychology (Rubin, 1994; Shneiderman, 1998). User-testing is generally used for the purpose of generating a usability problem report used for system refinements. Therefore, unlike controlled experiments, user-testing generally requires fewer subjects (maybe as few as three) and the amount of quantitative performance data are often reduced so the focus can be on errors users initiate while completing tasks.

User-testing employs participants who are representative of the target population in order to evaluate the degree to which an interface meets specific usability criteria (e.g., complete task *X* in *Y* amount of time with *Z* number of errors). During a user-testing session presents the user with a series of tasks to complete, usually in a simulated work environment. The human factors specialist unobtrusively

and impartially observes the user while he/she completes the tasks. Typically the user is asked to "think aloud" while performing the tasks. Thinking aloud allows the human factors specialist to capture the users' thought processes, uncovering clues about expectations, misconceptions, and confusion which, in turn, can facilitate tracing the source of documented usability problems (Rubin, 1994; Mayhew, 1999).

While the user is completing the task, the human factors specialist documents any comments made by the user regarding the task and records specific observations regarding the user's performance. There is little structure as to what information should be documented; rather it is left to the discretion of the human factors specialist to decide what information is important to record (e.g., notes to help understand the root of a problem, user comments about a specific problematic aspect of the system, questions to ask at the end of a testing session, etc.). Upon completion of all tasks, a debriefing session is conducted to further elicit comments and feedback about issues regarding the system. Thus, the output from user-testing is a collection of notes, comments, observations, etc. from the human factors specialist that must later be transcribed and analyzed (Rubin, 1994). According to Jeffries, Miller, Wharton, and Uyeda (1991) and Wharton, Rieman, Lewis, and Polson (1994), usability problems documented from user-testing sessions tend to contain more usability problems that are rated as "severe" (e.g., adversely impact user performance) than "minor" (e.g., simple design recommendations).

User-testing offers at least two major strengths. First, user-testing is considered to be a valid indicator of relevant usability problems (Rubin, 1994) because the data represents actual user performance with the system. Thus, extensive speculation and filtering of problems according to their predicted impact on users by human factors specialists is not needed; the impact can be assessed empirically from the evaluation session (Jeffries & Desurvire, 1992). Second, user-testing can provide objective performance data (e.g., performance errors when completing a task, time to complete task, number of errors, etc.), thereby circumventing personal opinions and interpretations.

Criticisms of user-testing techniques generally fall into four areas: cost of the test, artificiality of the testing environment, selection of participants, and reliability of the test results. User-testing is considered one of the more expensive evaluation techniques (Jeffries et al., 1991) because it involves recruiting end-users and staff resources are required to conduct the evaluation sessions and analyze the usability data (Scerbo, 1995). As with controlled experiments, sometimes testing in usability laboratories

do not necessarily reflect the actual operational environment and therefore do not necessarily indicate that a system will work in real world contexts (Rubin, 1994). Some researchers also have doubts as to how well the selected participants represent actual end users. Additionally, Holleran (1991) points out that designers can overestimate the power and generalizability of user-testing based on a small sample of subjects. For these and other reasons, many developers tend to supplement user-testing with usability inspection methods.

Cognitive Walkthrough

The cognitive walkthrough is an extensively researched usability inspection method (Sears, 1997; Andre, Hartson, & Williges, 2003). Cognitive walkthroughs (Lewis, Polson, Wharton, & Rieman, 1990) focus specifically on evaluating a system for ease of learning, particularly through exploration. The narrow focus on ease of learning is motivated by the observation that most users prefer to learn a novel system through exploration (Wharton, Bradford, Jeffries, & Franzke, 1992) rather than through a formal training course or reading the accompanying user's manual. That is, users tend to incorporate a "label-following" strategy when interacting with a novel system (Wharton et al., 1992). Label following refers to the user's strategy of selecting a particular action if the label for that action correlates with the user's goal. For instance, a user given the task of "saving a document" might initiate an action with the label "save" or a "save" icon.

In contrast to user-testing, evaluators applying a cognitive walkthrough adopt the perspective of the user, thereby inferring users' cognitive processes while performing specific tasks. Human factors specialists examine each step necessary to successfully complete a task, attempting to uncover mismatches between users' and designers' mental models, in terms of labels for menu titles, buttons, knobs, etc. Additionally, human factors specialists investigate the quality of the system's feedback for each action (Wharton et al., 1992). In order to maximize the cognitive walkthrough's effectiveness, evaluators should have more knowledge of cognitive concepts (e.g., goals, problem solving, etc.) than is required by most other inspection methods (Lewis et al., 1990; Wharton et al., 1992).

During a cognitive walkthrough, evaluators inspect an interface in the context of specified tasks by adopting the role of the targeted end-user. Although originally conceived as an independent process, recent implementations of the cognitive walkthrough include groups of evaluators that pool their problems

into a single report (Lewis, 1997). During the walkthrough, evaluators consider each action necessary to successfully accomplish the task. That is, the cognitive walkthrough analyzes the correct action sequence (i.e., most efficient route), asking if the correct action sequence will actually be followed by users (Polson, Lewis, Rieman, & Wharton, 1992; Lewis, 1997) or if there are elements in the display that may prohibit the correct sequence from being followed.

For each step in the action sequence, the evaluators consider four primary questions intended to stimulate a story about a typical user's interaction with the system. In particular, the evaluators ask the following four questions (Wharton et al., 1994):

1. *Will the users try to achieve the right effect?* [For instance, if the task is to open a new document, then the first thing the user must do is open the word processing program. Will the user know that he/she should be opening the word processing program?]

2. *Will the user notice that the correct action is available?* [If the action is to select from a visible menu, is the action legible, located in an easily viewable location or a location where the user expects it to be, etc.? If the word processing icon is hidden or buried under many menu layers, then he/she may never see it as a possible action.]

3. *Will the user associate the correct action with the effect trying to be achieved?* [If there is a menu option that says, "word processor," the user should have little difficulty associating the option with the goal. If the menu option is not so obvious, the user may have difficulty.]

4. *If the correct action is performed, will the user see that progress is being made toward solution of the task?* [If after selecting the word processing program and the system provides a dialog that states, "word processor opening," the user will understand that the action initiation was successful. Confusion may ensue when there is no feedback.]

While answering each of the aforementioned questions, evaluators document any usability problems encountered and reasons for those problems (Lewis, 1997). Because of the systematic manner with which it allows a system to be assessed, a cognitive walkthrough can diagnose usability problems at a micro-level of detail and provide potential resolutions. The output from a cognitive walkthrough is either a success story constructed by the evaluators that explains why a user would choose a particular action or a failure story indicating why a user would not choose the action (Lewis, 1997; Wharton et al., 1994). If a system is effective, there should be a high degree of mapping between the users' intentions and the interface; the appropriate action should be readily visible. According to Wharton et al. (1994), usability problems identified from cognitive walkthroughs have approximately equal numbers of usability issues

rated as “severe” and “minor”. Dutt et al. (1994), however, found that cognitive walkthroughs identify more usability problems rated that hinder task completion (i.e., severe) than personal preference (i.e., minor).

Similar to user-testing, the primary strength of the cognitive walkthrough is its task-based approach. The focus on specific tasks helps evaluators assess how the features of an interface fit together to either support users’ successful task completion or fail to do so. Additionally, the focus on critiquing correct action sequences is intended to provide feedback regarding how reasonable it is for a user to follow a specific action sequence (Lewis, 1997). Another advantage is the cognitive walkthrough’s systematic approach to diagnosing usability problems. Unlike other usability inspection methods where the primary purpose is to simply identify usability problems, in addition to identifying usability problems, the cognitive walkthrough provides possible resolutions for those issues. Lastly, cognitive walkthroughs focus on assessing the “learnability” of a system. Thus, interface options tailored to enable expert or “power” users (e.g., short-cuts, hidden functions, etc.) instead of novices will be detected as usability problems by the cognitive walkthrough.

Some weaknesses of the cognitive walkthrough include the need for human factors specialists to have a background in cognitive psychology and the extensive time necessary to apply the technique effectively (Lewis et al., 1990; Rowley & Rhoades, 1992; Wharton et al., 1992; Huart, Kolski, & Sagar, 2004). Researchers also have pointed out concerns with the scope of the cognitive walkthrough. For instance, Jeffries et al., (1991) found that cognitive walkthroughs identified more specific problems (e.g., terminology issues) than high-level general problems (e.g., global consistency); the low-level focus on specific action sequences often leave high-level problems unrecognized.

Heuristic Evaluation

The heuristic evaluation is also a prevalent usability inspection method in the HCI literature (Andre et al., 2003). Nielsen and Molich (1990) initially popularized the heuristic evaluation technique in the early 90’s and marketed the method as a “discount usability method” (i.e., inexpensive, fast, and easy-to-use) for evaluating systems.

When conducting a heuristic evaluation, a few expert evaluators (3-5 for maximum efficiency and cost-effectiveness) examine an interface and evaluate its compliance with a list of recognized usability principles (Nielsen, 1992). These usability principles, called heuristics, were initially developed by Molich

and Nielsen (1990) and later were revised by Nielsen (1994a) based on a factor analysis of 249 common usability problems. Evaluators typically inspect each display element within a system's interface, keeping in mind the list of heuristics. The heuristics represent general display design rules that describe common characteristics of interfaces with high usability levels. The following is a list of Nielsen's (1994b) 10 usability heuristics:

1. *Visibility of system status* – the system should always keep users informed about what is going on, through appropriate feedback within reasonable time.

2. *Match between system and 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.

3. *User control and freedom* – users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.

4. *Consistency and standards* – users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.

5. *Error prevention* – even better that good error messages is a careful design which prevents a problem from occurring in the first place.

6. *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.

7. *Flexibility and efficiency of use* – accelerators – unseen by the novice user – may often speed up the interaction for the expert user to such an extent that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

8. *Aesthetic and minimalist design* – dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.

9. *Help users recognize, diagnose, and recover from errors* – error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.

10. *Help and documentation* – even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

Nielsen (1994b) recommended that each evaluator independently step through an interface using a two-pass approach. The first pass is intended to help the evaluator obtain a general understanding of (i.e., "get a feel for") the system. The second pass allows the evaluator to focus on specific interface

elements, judging their compliance with the aforementioned heuristics list. Nielsen recommended also that evaluators' findings should be aggregated after all assessments have been completed to ensure independent and unbiased evaluations.

The heuristic evaluation provides a list of usability problems in the interface and classifies those problems with reference to the heuristics that were violated. This list of usability problems rated on severity range from severe to minor, but most rated problems tend to be dominated by minor problems (e.g., inconsistent fonts between screens) (Nielsen, 1994b; Sears, 1997). In fact, heuristic evaluations identify twice (Wharton et al., 1994) or two-thirds (Jeffries et al., 1991) as many minor-rated problems as severe-rated problems.

The advantages of the heuristic evaluation are that it is relatively inexpensive, easy to use, and does not require extensive pre-evaluation set-up time (Nielsen & Molich, 1990). Additionally, the heuristic evaluation does not require intensive training and background required by many other usability inspection methods (Nielsen, 1994b). Nielsen (1994b), however, notes that a background in cognitive psychology, human-computer interaction, and/or human factors can improve the performance of the evaluator significantly. Also, heuristic evaluations typically detect lots of usability problems (Nielsen & Molich, 1990; Jeffries et al., 1991) due to the breadth of the analysis (Wharton et al., 1994).

The heuristic evaluation is not without its shortcomings. First, usability problems identified are subjective and not based on objective data (Doubleday, Ryan, Springett, & Sutcliffe, 1997). Second, some heuristics contradict other heuristics. For instance, if designers incorporate a *minimalist* approach, then some interface options may be hidden, which violates the *recognition rather than recall* heuristic. Third, heuristic evaluators are simply observing the system and are not as absorbed in using the system as users are during a user-testing session. This may contribute to why usability problems found through heuristic evaluations and user-testing typically do not overlap (Doubleday et al., 1997). Fourth, a heuristic evaluation often identifies a large number of specific, one-time (i.e., the usability problem occurred once throughout a system), and low-priority problems (Jeffries et al., 1991). Finally, though the heuristic evaluation finds lots of usability problems, it does not provide systematic ways to remedy detected usability problems (Nielsen, 1994b).

The following table summarizes the usability assessment procedure, advantages, and disadvantages associated with user-testing, the heuristic evaluation, and the cognitive walkthrough.

Table 1. Summary of the Process, Advantages, and Disadvantages of User-testing, the Heuristic Evaluation, and the Cognitive Walkthrough.

UEM	*Process*	*Advantages*	*Disadvantages*
User-testing	▪ Targeted end-users perform representative tasks while "thinking aloud." ▪ HF specialist observes/records user performance.	▪ Identifies usability problems that actually impact users. ▪ Provides objective data (e.g., input errors, task completion time, etc.). ▪ Detects more problems rated as severe.	▪ Expensive if users need to be recruited. ▪ Sessions are videotaped and reviewed later = increased evaluator time.
Cognitive Walkthrough *Inspection Method*	▪ HF specialist adopts the role of the targeted end-user and inspects an interface in the context of specific tasks. ▪ The correct action sequence is analyzed to determine if it will be followed by users or not.	▪ Task-based approach. ▪ Provides feedback regarding how reasonable it is for users to follow the correct path. ▪ Systematic approach to diagnosing usability problems. ▪ Equivalent proportion of problems rated as severe and minor.	▪ Subjective. ▪ Time consuming. ▪ Requires background in cognitive psychology.
Heuristic Evaluation *Inspection Method*	▪ 3-5 HF specialists examine an interface and evaluate its compliance with a list of heuristics (i.e., general rules of thumb for designing usable systems). ▪ Usability problems are classified into one or more heuristics.	▪ Easy to use. ▪ Does not require extensive training. ▪ Detects a lot of usability problems.	▪ Subjective. ▪ Detects and classifies usability problems only (does not provide fixes). ▪ Most problems detected are rated as minor.

UEM Characteristics Incorporated into the User-CHAT

While user-testing, the cognitive walkthrough, and the heuristic evaluation are among the most popular UEMs in HCI literature and each have characteristics that satisfy *some* of the aforementioned constraints associated with the current certification process, none however, satisfies *all* the constraints. Therefore, the usability method most suitable for the certification process is one that integrates the

characteristics from user-testing, the cognitive walkthrough, and the heuristic evaluation that aptly satisfies the certification constraints. Table 2 shows the unique components/ideas that were appropriated from user-testing, the cognitive walkthrough, and the heuristic evaluation and integrated into the User-CHAT and why these distinctive aspects were deemed appropriate to suit the existing certification environment.

Table 2. Characteristics of User-testing, the Cognitive Walkthrough, and the Heuristic Evaluation Incorporated into the User-CHAT.

UEM	*Characteristics Incorporated into the User-CHAT*	*Rationale for Incorporation into the User-CHAT*
User-testing	▪ Use representative end users to generate performance-based data ▪ Establish benchmark tasks that exercise system functionality ▪ Ask users to "think aloud" during task completion	▪ *Identify* usability problems from the users perspective (i.e., the problems that actually impact users)
Cognitive Walkthrough	▪ Conduct a systematic usability assessment ▪ Establish correct action sequences for each benchmark task (e.g., efficient routes for task completion) ▪ Discover the root cause of potential usability problems	▪ *Analyze* usability problems (i.e., initial establishment of root causes of the problems)
Heuristic Evaluation	▪ Adopt the notion of developing a tool that is quick and easy use and does not require a lot of training ▪ Utilize a list of display design heuristics to synthesize usability problems	▪ *Classify* usability problems (i.e., initial real-time synthesis of usability data)

Development of the User-CHAT

The User-CHAT is a usability assessment technique designed to facilitate and standardize the certification process. This method utilizes behaviorally based metrics for evaluating the degree to which a candidate display conforms to human factors guidelines and usability principles for presenting various types of information in the cockpit. The evaluation supported by the User-CHAT is based largely on behavioral data collected from end-users performing representative tasks, rather than based solely on the judgments and expertise of one or more human factors specialists. Thus, not only does the User-CHAT support FAA evaluators in the comprehensive identification of certification issues but it also provides behavioral data to substantiate the existence, incidence, and severity of those issues to the

manufacturers. Also, while the User-CHAT is intended for use by certification officials, the tool could be used also by avionics manufacturers (as well as usability practitioners in other industries) throughout product development phases to identify and resolve usability issues.

While the ideal certification assessment procedure would utilize several different types of methodologies to ensure a comprehensive evaluation of a multifaceted candidate avionics system (Clamann & Kaber, 2004), real world situations like the FAA certification environment do not afford such comprehensiveness, due to time and resource limitations. Therefore, as described above, a more pragmatic approach was taken with the development of the User-CHAT such that specific components from several UEMs were chosen that afforded efficient and comprehensive identification, diagnosis, and classification of interface problems from a behaviorally-based perspective while satisfying time and resource limitations typically associated with the certification environment. By adopting a more behaviorally-based perspective to the evaluation, the User-CHAT allows certification personnel to better identify and justify usability bottlenecks.

The User-CHAT requires 2-3 personnel. One person acts as the user (preferably a flight test engineer or a pilot qualified to operate the candidate system), who completes a series of benchmark tasks. A second person acts as the supervisor, who monitors and records user performance and compares user performance with the gold standard (i.e., the most efficient route for task completion) for each benchmark task. If a third participant is available, he/she acts as the observer, who takes notes on user performance and dialog. If a third person is unavailable, the observer's responsibilities fall to the supervisor.

The User-CHAT is comprised of two phases: unstructured exploration and structured exploration. During the unstructured exploration phases, the user spends approximately 10-15 min freely exploring the system in order to gain familiarity with the system and to examine specific aspects of the system (e.g., obtain awareness with the menu structure, how input devices interface with the system, how information is organized, etc.). During unstructured exploration, the supervisor observes and records any comments or initial evaluations made by the user about the system as well as their own observations about the system.

During structured exploration, for each benchmark task, the user attempts to complete the task and the supervisor compares the user's performance to the gold standard for that task (i.e., the most efficient task completion sequence). When performance deviates from the gold standard, the supervisor: 1) records the first inefficient action (i.e., the first action performed instead of the gold standard; 2) tallies the number of subsequent inefficient actions until the user initiates the next required action (the gold standard) toward completing the task; and 3) identifies when the user spends an inordinate amount of head-down time scanning the interface attempting to determine what action to take next.

After each benchmark task is completed, the user interprets the displayed information/symbology and the supervisor compares the user's interpretation with the correct interpretation (as supplied by the manufacturer). Once the displayed data is interpreted, the supervisor directs the user's attention to where the first inefficient action occurred and asks the user two questions modified from the cognitive walkthrough:

1. Why was the gold standard action not initiated?
2. Why was the specific first inefficient action initiated?

Upon answering these two questions, the user and supervisor classify the answers to the two questions above in terms of a violation of one or more display heuristics. This list of heuristics accompanying the User-CHAT originated from several established human factors and usability sources including Nielsen's (1994a) factor analysis of 249 usability problems and the 13 principles of display design (Wickens et al., 1998). In addition to these more general heuristics, aviation specific heuristics were added to supplement specific issues associated with avionics systems. The user and supervisor are given the list of heuristics and their definitions to consult during this classification. This list of heuristics is analogous to a list of relevant FAA regulations, as well as important human factors considerations.

The User-CHAT was derived originally from a user-testing based methodology that was employed in a full-scale usability assessment performed over a period of three weeks in 2002 on two commercially available multi-function displays (Uhlarik et al., 2002). The full scale usability assessment was modified into the User-CHAT to accommodate time and personnel constraints inherent in the current certification process and was structured to support FAA evaluators with limited human factors and/or

usability training to perform a usability assessment where the ability to identify and diagnose bottlenecks is maximized.

As part of the effort to confirm that the User-CHAT is an efficient, effective, and useful usability technique for implementation into the certification process, the following assertions must be validated empirically:

1. Support an evaluation of a system within 3-4 hours using approximately 15-20 benchmark tasks.
2. Support an evaluation of a system by evaluators with limited human factors knowledge or training.
3. Require little training of evaluators.
4. Support the evaluation of complex systems supporting multiple gold standards for benchmark task completion.
5. Provide a more efficient and pragmatic evaluation of a system than existing UEMs (e.g., user-testing, cognitive walkthrough, and heuristic evaluation).

Since the initial development of the User-CHAT, efforts have been undertaken to evaluate its efficacy and efficiency as a UEM. Research by Uhlarik et al. (2003, 2004) have demonstrated that the User-CHAT supports an evaluation within 3-4 hours (with 15-20 benchmark tasks), can be utilized by individuals with limited human factors background and does not require much training to be used effectively, even for complex systems that support multiple gold standards.

Additional modifications have been made to the User-CHAT as a result of these evaluations. First, a “head-down” time component was added to assess when the user was thinking about the next action to initiate (“thinking time” could also be considered a bottleneck to user performance, even though the initiated actions may still be consistent with the gold standard). Second, a memory component was incorporated to assess whether memory influenced performance. That is, was the inefficient action initiated because the next action was not clearly visible, descriptive, intuitive, etc. thereby causing the user to rely on memory in order to recall functionality? Lastly, upon discussion of the documented bottlenecks, the user and supervisor assign a severity rating (e.g., an indication of how much the problem affected the user) to the usability bottleneck.

Comparative Usability Evaluation

The User-CHAT was tailored for the typical certification environment. Therefore, given the constraints of this environment, the User-CHAT should be the ideal usability assessment protocol for evaluating candidate systems. The User-CHAT accounts for time and resource limitations by focusing diagnostic thinking on actual problems identified by representative end-users and not on potential problems that may or may not impact users. Also, it has been demonstrated through prior research that the User-CHAT can effectively capture usability issues when used by evaluators with limited human factors knowledge or training. Thus flight test engineers with limited human factors domain knowledge can utilize the User-CHAT effectively. Additionally, the User-CHAT supports real-time documentation, which is essential, because videotaping certification sessions is not allowed.

The next step in the User-CHAT evaluation effort is to provide the FAA with empirical evidence that the User-CHAT is a better alternative to other established UEMs, and given the constraints of the certification environment, the User-CHAT is the usability assessment method of choice for providing an efficient evaluation of a system while standardizing the certification process. *Therefore, the present research is intended to conduct a comparative usability evaluation whereby the efficacy of the User-CHAT is directly and empirically compared to the efficacy of other established usability methods, specifically user-testing, the cognitive walkthrough, and the heuristic evaluation.* While the primary purpose of this research was to determine the effectiveness of the User-CHAT when compared to other well-established inspection methods, *the secondary purpose is to provide usability researchers and practitioners with an approach for conducting UEM comparison studies using standardized measures and definitions.* Thus a review is needed to highlight some comparison metrics that have been postulated recently to assess UEM effectiveness.

UEM Performance Metrics

The purpose of comparative usability studies is to allow the effectiveness of several UEMs to be directly compared along similar performance dimensions. Researchers conducting such comparison studies, however, have not agreed on commonly used and understood metrics for assessing UEM performance (Andre et al., 2003). Indeed, Lund (1998) pointed out that no single metric exists for direct comparison, resulting in multiple non standardized performance metrics being employed in many previous UEM comparison studies. Andre (2000) asserts that difficulties in reliably comparing UEMs can

be attributable to three non standardized factors in the HCI literature: (1) lack of established comparison standards, (2) definitions for comparison metrics vary, and (3) lack of a standardized UEM comparison procedure.

Bastien and Scapin (1995) initially identified three metrics upon which UEM performance could be assessed: thoroughness, validity, and reliability, which were later modified by Sears (1997). According to Andre et al. (1999), thoroughness, validity, and reliability form the core metrics that should be utilized in UEM comparison studies.

Thoroughness

Thoroughness is the most attractive metric for comparing UEMs (Hartson, Andre, & Williges, 2001). However, there are two different methods for calculating thoroughness. Sears (1997) introduced thoroughness as the ratio of the number of "real"[4] usability problems identified by a particular UEM relative to the total number of "real" usability problems that exist[5]:

$$\text{"Real" Thoroughness} = \frac{\text{\# of real usability problems found}}{\text{\# of real usability problems that exist}} \quad (1)$$

For instance, suppose 50 total usability problems were identified through a given UEM. If 30 of those usability problems were also listed in the standard list of 40 real usability problems, then that UEM would yield a thoroughness score of 30/40 = 0.75. Using signal detection theory terms, a UEM's thoroughness index refers to its ability to detect real usability problems (i.e., hits) and avoid Type II errors (i.e., misses). UEMs with low thoroughness scores leave important "real" usability issues unidentified (Hartson et al., 2001).

Calculating the denominator (i.e., "the number of real problems that exist") in the ratio for equation (1), however, is difficult (Sears, 1997) and often not clearly defined in most UEM comparison studies (Hartson et al., 2001). As discussed in Appendix A, a variety of techniques are available to determine the number of real usability problem that exist in an interface. Landauer (1995) and Newman (1998) contend that user-testing is the *de facto* standard for determining the denominator (i.e., standard-

[4] According to Hartson et al. (2001), "a usability problem (e.g., found by a UEM) is real if it is a predictor of a problem that users will encounter in real work-context usage and that will have an impact on usability (user performance, productivity, or satisfaction or all three)" (p. 383).
[5] For additional information regarding the process of determining both the "real" usability problems found and the number of "real" usability problems that exist, please refer to Appendix A.

of-comparison) because the problem list is generated by representative end-users performing typical tasks in a simulated work context. When the standard-of-comparison has been established, real thoroughness can be computed by the following intersection:

$$\text{“Real” Thoroughness of } UEM_P = \frac{|P \cap A|}{|A|} \quad (2)$$

where A is the list of usability problems generated through user-testing (e.g., UEM_A) and P is the list of usability problems found by a particular UEM (e.g., UEM_P). Using the example above, the 50 hypothetical usability problems found in the earlier example by UEM_P represents P, and $P \cap A$ are the 30 usability problems shared by both lists (i.e., intersection). The Venn diagram in Figure 1 illustrates this intersection.

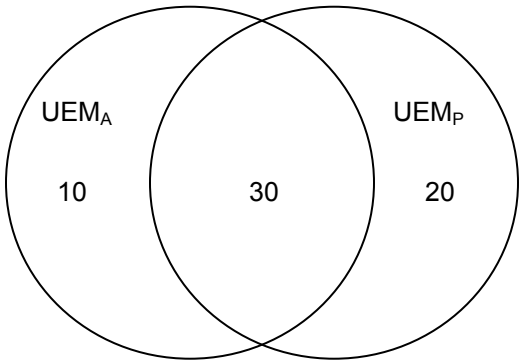

Figure 1. Venn diagram illustrating the intersection between the standard list of real usability problems identified by UEM_A and the list of usability problem identified by UEM_P.

Alternatively many previous UEM comparison studies calculated thoroughness by removing the concept of “real” and simply aggregated the lists of usability problems generated from all UEMs as the total aggregate list of usability problems to which individual UEM performance is compared (Hartson et al., 2001). The “aggregate” thoroughness is calculated as a function of the following:

$$\text{“Aggregate” Thoroughness} = \frac{\text{\# of usability problems identified by a UEM}}{\text{\# of total usability problems identified by all UEMs}} \quad (3)$$

For instance, if 75 total aggregate usability problems were found by all the UEMs and one specific UEM found 50 usability problems, then that UEM would have a thoroughness score of 50/75 = 0.67.

While combining usability problems across all UEMs (i.e., “aggregate thoroughness”) credits each UEM for all the usability problems it identified (e.g., “real” and “non-real”); unfortunately validity scores for each UEM cannot be properly calculated (see below). Calculating a UEM’s thoroughness based on real

usability problems (i.e., "real" thoroughness) provides a better indication of a UEM's ability to detect usability problems that actually impact users and allows researchers to calculate validity using the same data set.

According to previous UEM comparison studies, the heuristic evaluation tends to find more usability problems (i.e., attains higher "aggregate" thoroughness scores) than other UEMs (e.g., cognitive walkthrough and user-testing) when those UEMs are compared head-to-head (Andre et al., 2003). Such conclusions, however, should be cautioned due to experimental design inconsistencies associated with many UEM comparison studies, e.g., failing to control (1) the time allotted for an evaluation, (2) task similarity, or (3) the number of participants contributing to the results of each specific UEM.

Jeffries et al. (1991) found that the heuristic evaluation (0.57) was more thorough ("aggregate" thoroughness) than the cognitive walkthrough (0.15), which in turn was slightly more thorough than user-testing (0.14). However, in the Jeffries et al. study, evaluators were given two weeks to complete the assessment while evaluators using the cognitive walkthrough and user-testing were given a few hours only. Similar conclusions were made by Wang and Caldwell (2002) and Tan and Bishu (2002) and Cuomo and Bowen (1994) despite also allowing unequal evaluation times. When the amount of evaluation time was kept constant, Sears (1997) reported that the heuristic evaluation (0.85) was more thorough than the cognitive walkthrough (0.61). Andre et al. (2003)[6], however, failed to find a difference between the heuristic evaluation (0.18) and cognitive walkthrough (0.20). In order to help control for the effect of evaluation time on UEM performance, in the present study, evaluators will be instructed to complete the evaluation within a given time limit.

Because some UEMs are task-dependent (e.g., cognitive walkthrough, user-testing), it is important that UEM comparison studies use the same tasks for all UEMs. Using the same tasks across UEMs ensures that evaluators will assess the same displays. However, it must be noted that many UEM comparison studies suggest that the heuristic evaluation is more thorough than other UEMs whether they incorporate the same tasks across UEMs (e.g., Doubleday et al., 1997) or fail to utilize the same tasks across UEMs (e.g., Jeffries et al., 1991; Virzi, Sorce, & Herbert, 1993; Wang & Caldwell, 2002). Thus,

[6] The thoroughness conclusions by Cuomo and Bowen (1994), Sears (1997), and Andre et al. (2003) are based on "real" thoroughness calculations. All other thoroughness conclusions are based on "aggregate" thoroughness calculations.

while it has not been specifically shown to influence the relative thoroughness scores for the heuristic evaluation, maintaining task similarity is certainly recommended to ensure a strong empirical design for the present comparison study.

It has been well documented that the more evaluators contributing to the identification of usability issues, the more usability problems are likely to be found (Nielsen, 1994b; Virzi, 1992; Wright & Monk, 1991). Again it must be noted that past research has found the heuristic evaluation to be more thorough that the cognitive walkthrough and user testing both when the number of participants was controlled (Dutt et al., 1994; Sears, 1997) and when fewer participants were involved in the heuristic evaluation than other UEMs (e.g., Jeffries et al., 1991; Doubleday et al., 1997; Virzi et al., 1993; Wang & Caldwell, 2002). However, maintaining an equal number of evaluators will ensure that any differences found between UEMs in future studies cannot be traced to this inconsistency.

Validity

A UEM attains high validity scores if it maximizes identification of real usability problems and minimizes detection of trivial usability problems. The extent to which a particular UEM can identify important or real usability problems (i.e., issues the impact user performance, productivity, or satisfaction) reflects its validity (Hartson et al., 2001). Sears (1997) defined validity as an indication of the number of problems found that are real problems to the total number of problems identified:

$$Validity = \frac{\#\ of\ real\ problems\ found}{total\ \#\ of\ problems\ identified} \quad (4)$$

Validity and thoroughness use the same standard-of-comparison list of usability problems and can be calculated from the same data. In reference to the UEM thoroughness example given above, for the UEM that found 50 usability problems, of which 30 were real usability problems, using equation (3) produces a validity rating of 30/50 = 0.60. UEMs with high validity scores detect large proportions of usability problems that are real. In signal detection theory terms, a UEM's validity refers to its ability to attenuate false alarms and ignore non-usability issues (i.e., correct rejections). UEMs with low validity indices find large numbers of usability problems that are not relevant, not important, or not real. Consequences from a UEM with low validity include the potential masking of usability problems that require attention and

wasting of time and effort evaluating, analyzing, and reporting these non-real usability problems (Hartson et al., 2001).

Similar to computing thoroughness, validity can be explained in terms of the intersection between two usability problem lists. Again, let P represent the 50 total number of usability problems found by a particular UEM (e.g., UEM_P) and A represent the 40 usability problems found by UEM_A (e.g., the usability problems found through user-testing). The intersection between A and P provides the following equation:

$$Validity = \frac{|P \cap A|}{|P|} \quad (5)$$

Thus, the intersection between P and A yields 30 usability problems, which leads to a validity index of 30/50 = 0.60.

If researchers use the aggregation of usability problems found from various UEMs to produce the standard list of usability problems (A), then the validity calculation is flawed. To illustrate this flaw, suppose a UEM comparison study compared UEM_P, UEM_Q, and UEM_R. Let P(X) represent the list of usability problems identified by UEM_P, Q(X) represent the list of usability problems identified by UEM_Q, and R(X) represent the list of usability problems identified by UEM_R. The union of UEM_P, UEM_Q, and UEM_R is represented by the following equation:

$$A(X) = P(X) \cup Q(X) \cup R(X) \quad (6)$$

Figure 2 provides a Venn diagram that illustrates the union three UEMs where UEM_P identified seven usability problems [P(X)], UEM_Q and UEM_R identified nine usability problems each [Q(X) and R(X) respectively]. Five problems were shared by all three methods. One problem was shared between UEM_P and UEM_R; two problems were shared between UEM_R and UEM_Q. The 13 total unique usability problems were found by all three UEMs represent the standard list of usability problems, A(X).

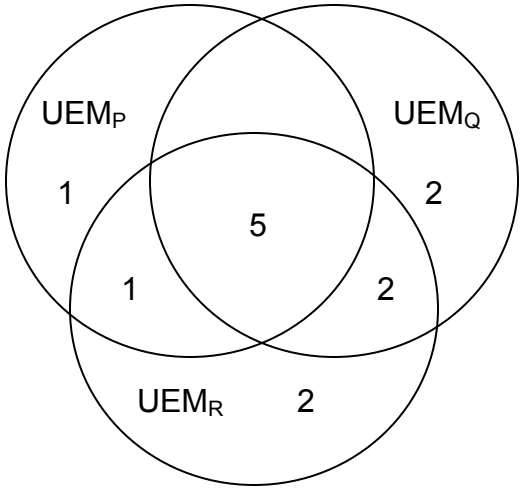

Figure 2. Venn diagram illustrating the union of three UEMs.

Using the following formula, the validity of UEM_P can be calculated:

$$\textit{Validity of } UEM_P = \frac{|P(X) \cap A(X)|}{|PX)|} \quad (7)$$

Now here's the problem with the union of UEMs to generate a standard list and then calculating a particular UEM's validity. A(X) is a union of all three UEMs, which contains P(X) (see equation 6). Thus, P(X) is a subset of A(X). Applying whole numbers, the equation shows the intersection between the seven usability problems found by UEM_P and the 13 total usability problems found by all three methods to be the seven usability problems found by UEM_P. Nothing is removed from P(X) when intersected with A(X). Adopting this approach guarantees that the intersection of a particular UEM usability problem list and the standard usability problem list will be the UEM problem list itself, meaning that all usability problems detected by each UEM are always 100% valid. Equation 7 can be re-written as the following:

$$\textit{Validity of } UEM_P = \frac{|P(X)|}{|P(X)|} = 1.0 \quad (8)$$

While it has been demonstrated that the union of UEMs to generate a standard list of real usability issues confounds validity measures, thoroughness measures are more robust.

Few UEM comparison studies have reported validity measures and those that do have conflicting conclusions about the validity of the UEMs. For instance, Cuomo and Bowen (1992), Sears (1997), and Andre et al. (2003) all report that the cognitive walkthrough is more valid than the heuristic evaluation

when using the usability problems found through user-testing as the real problems. Sears reported a validity score of 0.69 for the heuristic evaluation and 0.91 for the cognitive walkthrough while Andre et al. (2003) reported 0.48 and 0.70 validity scores for the heuristic evaluation and cognitive walkthrough respectively. Conversely, using user-testing data as the standard for real problems also, Desurvire, Kondziela, and Atwood (1992) found the heuristic evaluation to be more valid than the cognitive walkthrough. However, many of these validity conclusions were based on unequal numbers of participants contributing to the usability problems found through user-testing, the heuristic evaluation, and the cognitive walkthrough (e.g., Desurvire et al., 1992; Andre et al., 2003; Cuomo & Bowen, 1992) and on unequal time allotted for the evaluation (e.g., Cuomo & Bowen, 1992). Thus in the present study the number of evaluators contributing to the data and the amount of evaluation time will be controlled across UEMs to attenuate the possibility that these factors may influence validity scores.

Effectiveness

Hartson et al. (2001) argue that neither thoroughness nor validity alone is sufficient for measuring a UEM's performance. For instance, it has been shown that the heuristic evaluation finds many of usability problems. Many of those identified usability problems may be real usability problems (e.g., high thoroughness score) or many may be trivial (e.g., low validity score). The "ideal" UEM would detect a lot of real usability problems (i.e., high thoroughness) in its total list of usability problems (i.e., high validity). Consequently, Hartson et al. developed the effectiveness metric, which is the product of thoroughness and validity scores:

$$\textit{Effectiveness} = \textit{"Real" Thoroughness x Validity} \quad (9)$$

Thus, if a UEM's thoroughness and validity scores are high; its effectiveness score will be high as well. Effectiveness measures the overall merit of a UEM (Andre, 2000). Andre et al. (2003) reported that the cognitive walkthrough (0.15) was more effective than the heuristic evaluation (0.09). Therefore in the present study the effectiveness metric will be used as an overall indication of how the heuristic evaluation, the cognitive walkthrough, and the User-CHAT compare to the "ideal" UEM (i.e., one that has high thoroughness and high validity).

Reliability

The reliability metric is an indication of whether similar results would be obtained under similar conditions (Sears, 1997). One method of assessing reliability is to determine if similar numbers of usability problems are identified by different evaluators or groups of evaluators across sessions within each UEM. That is, reliability can index the degree of consistency of a UEM to find similar numbers of problems across evaluation sessions (Hartson et al., 2001). Sears (1997) proposed the following ratio for calculating a reliability index for a given UEM:

$$Reliability_{temp} = 1 - \frac{Stdev\ (\#\ of\ real\ problems\ found)}{Average\ (\#\ of\ real\ problems\ found)} \quad (10)$$

$$Reliability = \text{Maximum}(0, \text{Reliability}_{temp}) \quad (11)$$

The possibility of negative $\text{Reliability}_{temp}$ values is eliminated through this two-step computation. Unfortunately, this reliability calculation fails to show consistency in the types of usability problems identified by evaluators. However, Sears (1997) contends that a minor modification would allow for reliability indices to be calculated for specific types of usability problems identified by each evaluator.

Although not reported frequently in the literature, Sears (1997) and Hartson et al. (2001) argue that reliability ratings should be an important measure to include in UEM comparison studies. Sears (1997) reported that the heuristic evaluation was more reliable than the cognitive walkthrough.

Time per Usability Problem

Another UEM comparative metric not reported in the HCI literature is the time spent to detect a single usability problem during an evaluation using a particular UEM. Usability practitioners may find it useful to know how much time, on average, was devoted to identifying one usability problem for a given UEM. As shown in the equation below, time per usability problem is simply the ratio of the total time required during the evaluation to the total number of usability problems identified:

$$Time\ per\ Usability\ Problem = \frac{total\ evaluation\ time}{total\ \#\ of\ problems\ identified} \quad (12)$$

Small time scores indicate that, on average, a short amount of time was spent identifying usability problems whereas greater time scores signify that large amounts of time were spent identifying usability problems. That is, UEMs that identify lots of usability problems in a relatively short amount of time may be

more appealing to usability practitioners versus a UEM that takes longer to identify a few usability problems.

While it is nearly impossible to maximize all the comparative metrics simultaneously (Andre, 2000), when deciding which UEM to incorporate into the product design lifecycle, Gray and Salzman (1998) contend that usability practitioners should consider more than one comparative metric. For instance, while considering only a UEM's thoroughness score, usability practitioners will gain a partial picture of the UEM's overall performance. Thus, practitioners should consider all comparative metrics when deciding on a usability technique.

Hypotheses

For this UEM comparative study, five hypotheses were tested.

- *Hypothesis #1 – Thoroughness*
    a. *"Real" Thoroughness (i.e., the number of real" usability problems found by a UEM relative to the total number of real usability problems that exist)* -- Because the User-CHAT is a variant of user-testing and the list of total real usability problems is identified through user-testing, the User-CHAT will identify more real usability problems (i.e., higher "real" thoroughness) than the heuristic evaluation and the heuristic evaluation will identify more real usability problems than the cognitive walkthrough.
    b. *"Aggregate" Thoroughness (i.e., the number of usability problems identified by a UEM relative to the total number of usability problems identified by all UEMs)* – Because of the breadth of its underlying methodology (i.e., focus on all individual display elements), the heuristic evaluation will identify more usability problems (i.e., higher "aggregate" thoroughness) than the cognitive walkthrough and the User-CHAT.
- *Hypothesis #2 – Validity (i.e., the number of real usability problems found by a UEM relative to the total number of usability problems identified by that UEM)* – Because the User-CHAT is a variant of user-testing, the User-CHAT will contain more real usability problems in its list of usability problems (i.e., higher validity) than the cognitive walkthrough, which will be more valid than the heuristic evaluation.
- *Hypothesis #3 – Effectiveness (i.e., the product of "real" thoroughness and validity)* – Because the User-CHAT is a variant of user-testing, the User-CHAT's list of usability problems will contain many real usability problems (i.e., high thoroughness) in the total number of usability problems it identified (i.e., high validity). Thus, the User-CHAT will be more effective than the cognitive walkthrough, which in turn will be more effective than the heuristic evaluation.
- *Hypothesis #4 – Reliability (i.e., the ability to find similar usability problems under similar conditions)* – Because the usability problems found through the User-CHAT are based primarily upon user-performance (i.e., objective) data, the User-CHAT will find more similar usability problems across evaluation sessions (i.e., more reliable) than the usability problems found

through the heuristic evaluation and the cognitive walkthrough. The usability problems found through the heuristic evaluation and the cognitive walkthrough are strongly tied to the evaluator's background and motivation, which could lead to greater variability across evaluation sessions. Greater variability leads to lower reliability scores.

- *Hypothesis #5 – Problem Severity* -- Because the usability problems found through the User-CHAT are problems that actually affect user performance, the User-CHAT will identify a larger proportion of usability problems that are rated severe than the cognitive walkthrough and the heuristic evaluation.

Chapter 2 -- Method

Participants

Eighteen general aviation (GA) pilots and eight human factors specialists volunteered to assist in evaluating an avionics system using one of four UEMs (e.g., user-testing, User-CHAT, heuristic evaluation, or cognitive walkthrough). Several usability researchers (e.g., Nielsen, 1994b; Virzi, 1992; Wright & Monk, 1991) have shown that, for optimal cost effectiveness, five participants are needed for each UEM in order to detect approximately 80% of the usability problems in an interface. Table 3 shows participant configurations of all evaluation sessions for each UEM in the comparative study. Except for the human factors specialist that assisted in the user-testing sessions (see below for explanation), all remaining participants had minimal to no exposure with the avionics system used for evaluation. It is important to note that each UEM had the same total number of people actively contributing to the usability data.

Table 3. Configuration and Characteristics of the Evaluation Sessions for each Usability Technique.

		Pilots		*Human Factors Specialists*		
UEM	*Session*	*Rating*	*Total Flight Hrs*	*Design, Testing, and Evaluation Experience*	*Session Configuration*	*# of People Contributing to the Data*
User-testing	1	IFR	210	2 yrs	Pilot *actively* completes tasks, HF specialist *passively* observes (same HF specialist was used for all user-testing sessions)	6
	2	IFR	1426			
	3	IFR	800			
	4	IFR	4000			
	5	IFR	350			
	6	IFR	500			
			M = 1214.33			
User-CHAT	1	IFR	1670	6 yrs 2 mo	Pilot *actively* completes tasks, HF specialist *passively* observes (same HF specialist was used for all User-CHAT sessions)	6
	2	IFR	750			
	3	IFR	900			
	4	IFR	350			
	5	IFR	2400			
	6	IFR	350			
			M = 1070.00			
Heuristic Evaluation	1	IFR	3000	7 yrs	Pilot and HF specialist *actively* evaluate system together	6
	2	IFR	2500	13 yrs		
	3	IFR	3170	6 yrs		
			M = 2806.66	*8.60 yrs*		
Cognitive Walkthrough	1	IFR	1425	10 yrs	Pilot and HF specialist *actively* evaluate system together	6
	2	VFR	150	8 yrs 6 mo		
	3	VFR	100	5 yrs		
			M = 558.33	*7.83 yrs*		

Nielsen (1994b) suggested that usability assessments are most effective if the evaluators are both human factors specialists and subject matter experts (i.e., domain experts). However, individuals with expertise in both human factors and aviation are scarce. Thus, it was more practical to allow a GA pilot and a human factors specialist to collaborate when assessing the avionics system. The pilot brought aviation domain knowledge while the human factors specialist brought human factors and usability domain knowledge. Working together, the pilot and the human factors specialist provided a suitable arrangement to maximize the identification of usability problems supported by each UEM than if they were assessing the system independently. The benefit of this collaboration was especially apparent in the heuristic evaluation and cognitive walkthrough sessions, as the meaningfulness of the results from those two usability inspection methods is influenced directly by the skillfulness of the evaluators and the extensiveness of their background knowledge (both domain knowledge and usability experience).

All evaluation sessions were comprised of one GA pilot and one human factors specialist. Regardless of the UEM, the GA pilots represented end-users of the avionics system; as such, GA pilots actively contributed to the list of identified usability problems. The same cannot be said for the human factors specialists. In the heuristic evaluation and cognitive walkthrough, both the pilot and the human factors specialist collaborated together and actively contributed to the identification of usability problems. The resulting list of usability problems compiled after the three sessions of either the heuristic evaluation or the cognitive walkthrough were based on the *active* input of six evaluators (three pilots and three human factors specialists). Conversely, for user-testing and User-CHAT sessions, the pilot was the only person who actively contributed to the identification of usability problems, through his/her performance on the benchmark tasks (see below). The two human factors specialists serving in the user-testing and User-CHAT sessions, respectively, were simply passive observer/data collectors; in these sessions, the human factors specialists did not actively identify usability problems. Thus, the human factors specialist assigned to user-testing was the same data recorder/observer for all six user-testing evaluation sessions while the role of the User-CHAT supervisor was performed by the same human factors specialist (different from the user-testing human factors specialist) for all six User-CHAT evaluation sessions. Accordingly, the resulting list of usability problems compiled after the six evaluation sessions of either user-testing or the

User-CHAT were also based on the active input of six evaluators only (e.g., the six pilots). In sum, the lists of identified usability problems for all UEMs were based on the active contribution of six evaluators.

The GA pilots did not have any background or experience in usability assessments. Pilots assigned to the user-testing and the User-CHAT sessions required little training, as their main responsibility was to complete the benchmark tasks. However, since the User-CHAT required both the pilot and human factors specialists to classify usability problems into violations of one or more display design heuristics, prior to the evaluation a copy of the heuristics list associated with the User-CHAT was sent to the pilots participating in the User-CHAT sessions. These pilots were explained the general purpose of the display design heuristics and were asked to read through the heuristics list in preparation for their evaluation sessions. Any questions or concerns were answered prior to conducting the evaluation session.

Conversely, each GA pilot assigned to the heuristic evaluation and cognitive walkthrough sessions were trained with their respective usability inspection method. Training materials for the heuristic evaluation and cognitive walkthrough can be found in Appendix B and Appendix C respectively. Several months prior to participation in the evaluation sessions, these GA pilots completed a 1-hour training session and a 2-hour hands-on practice evaluation on several tasks (e.g., create and modify flight plan, change view range, display and interpret various types of weather, terrain, and traffic information, access navigation frequencies and other airport information, etc.) on a commercially available multi-function display (MFD) in order to gain familiarity with employing their respective UEM. Upon completion of the initial training sessions, pilots were instructed to practice evaluating other interfaces on their own time using either screen shots of example tasks from a computer simulation of a different commercially available MFD (see Appendix D) or using other available interfaces. To help ensure the pilots were applying their respective UEM appropriately, a few weeks prior to assessing the avionics system, each pilot participated in a 1-hour “refresher” training session.

The human factors specialist assigned to the User-CHAT sessions was one of the User-CHAT developers; thus no training on how to use the technique was required. The human factors specialist for the user-testing sessions was employed at the avionics manufacturer, was familiar with the user-testing process, and had conducted numerous user-testing evaluations. However, the human factors specialists

in either the heuristic evaluation and/or cognitive walkthrough sessions were recruited from industry and academia. These human factors specialists all had human factors knowledge and exposure to either the heuristic evaluation or cognitive walkthrough. However, in order to ensure consistency when applying the heuristic evaluation or cognitive walkthrough and to reacquaint (if necessary) them with the usability inspection methods, weeks prior to evaluating the avionics system, each human factors specialist was given the training materials for their assigned technique (the same materials administered to the GA pilots) to review before the session. Any questions or concerns regarding the application of their respective UEM were discussed and resolved prior to the evaluation sessions.

The System & Benchmark Tasks

A state-of-the-art suite of flightdeck interfaces housed in a fixed-based flight simulator provided the context to compare the efficacy of the UEMs. The avionics suite was designed to accommodate both single pilot and pilot / co-pilot configurations, incorporated bezel buttons, dedicated function keys, and knobs as input devices, and could display terrain, traffic, flight plans, weather, engine performance, and other flight relevant information on either a center mounted MFD, two PFDs (primary flight displays) that flanked the MFD, or a Flight Guidance Control Panel (FGCP) positioned above the MFD. The MFD and PFDs each had their own control panel. The MFD control panel was placed directly below the MFD while the PFD control panels were adjacent to the PFDs, located on the left and right side of the MFD. However, some information presented on the PFDs (e.g., data associated with the autopilot, barometric and minimums settings) was manipulated by the FGCP. Both the MFD and PFDs were 10.4 in. flat panel liquid crystal displays that were state-of-the-art in resolution and provided wide viewing capability to support cross-cockpit scanning by either pilot.

Eighteen benchmark tasks, along with their respective gold standard(s) (i.e., most efficient route(s) for task completion), were established in order to evaluate the flightdeck interfaces. The benchmark tasks represented tasks supported by the system that are typically performed by pilots. According to Karat, Campbell, and Fiegel (1992), using predetermined, typical end-user tasks to assess a system is preferred over allowing evaluators to evaluate a system at their own discretion. Additionally, the use of benchmark tasks provided more control to the assessment as the evaluators assessed the same functionality regardless of the usability method employed. In order to compare the effectiveness of the

usability methods under similar conditions, the same tasks were used for all evaluation sessions. All tasks were evaluated independently by two pilots for realism, readability, and understandability. The following 18 tasks were used to explore a variety of functionality associated with the avionics system:

1. Display the Terrain Awareness Warning System and Flight Plan Information
2. Pull up the Fuel Flow Diagram
3. Change the Moving Map's View Range
4. Display Graphical Weather Information
5. Pull up the Electrical System Diagram
6. Display the Oil Temperature and Pressure Gauges
7. Display the Anti-ice System Diagram
8. Overlay Traffic Information on the Horizontal Situation Indicator (HSI)
9. Overlay Weather Information on the HSI
10. Set up Communication Active and Standby Frequencies
11. Tune the Distance Measuring Equipment
12. Set up Navigation Frequencies
13. Set Course, Tune Automatic Direction Finder, and Set Number One Bearing Pointer
14. Set Heading Bug and Altitude Select
15. Turn on the Autopilot, Select Altitude and Heading Hold Modes, and Arm the Approach
16. Calibrate the Barometer and Minimums
17. Tune Transponder and Identify
18. Set Transponder to Transmit Altitude

The first seven tasks involved functionality specific to the MFD and the MFD control panel, while tasks 8-13 and 17-18 involved functionality specific to the PFD and the PFD control panel. Tasks 14-16 focused primarily on functionality associated with the FGCP. Task order was maintained throughout the study.

Additionally, the first nine tasks were focused more on evaluating and interpreting displayed information (i.e., assessing the characteristics of the various display elements) while the last nine tasks were focused more on accessing information (i.e., the input actions required to display the desired

information). That is, while all tasks required evaluators to assess the input actions necessary to complete the task and evaluate and interpret the displayed information, the first nine tasks depended more on display elements and less on the input actions necessary to display the desired information; vise versa for the last nine tasks.

Procedure

Prior to the evaluation sessions, all participants read a description of the study's purpose, signed a non-disclosure agreement[7] and an informed consent sheet and completed demographic questionnaires (see Appendix E). Participants were given a brief orientation session of the simulator and were explained the purpose of the video recorder (used to record all evaluation sessions). Participants were then explained how data collection would proceed (depending upon the UEM to which the pilot and human factors specialist were assigned). See Appendix F for instructions administered for each UEM. In order to ensure equivalent pre-evaluation exposure time, all participants were given approximately 10-15 min to freely explore the system and to examine more specific aspects of the system (e.g., become familiar with system, how input devices interface with the system, how information is grouped, etc.). After free-exploration, the evaluation session commenced. Because many usability evaluations occur typically under some time constraints (Andre, 2000), regardless of the UEM incorporated, participants were instructed to try to assess the system using the 18 benchmark tasks in three hours. The time restriction was designed to be short enough to create some time pressure. While participants were ultimately allotted additional time to complete tasks if necessary, they were not informed that additional time was available or possible during the session. Additionally, all appropriate reference material (e.g., description of UEM, list of heuristics for the heuristic evaluation and User-CHAT, examples of cognitive walkthrough failure/success stories, etc.) were made available during all evaluations.

User-Testing

During user-testing sessions, the human factors specialist was seated behind and out-of-sight from the pilot. The specialist presented sequentially to the pilot the benchmark tasks to complete and then unobtrusively and impartially observed the pilot as he completed each task. As recommended by Prail

[7] At the time of the present study, the avionics suite was proprietary information.

(1991) and Dumas and Redish (1999), human factors specialists for user-testing sessions should have in-depth knowledge of the target system. The specialist recruited for these user-testing sessions was a member of the system's development team and had extensive experience with the avionics system. Additionally, she was a commercial airline pilot prior to working for the manufacturer.

Pilots were asked to "think aloud" as they proceeded through the tasks. Each pilot was explained the purpose of the thinking aloud protocol. Asking the pilot to think aloud allowed the specialist to capture users' thought processes while completing the task, uncovering clues about expectations, misconceptions, and confusion which could help reveals the source of a usability problem.

While the pilot completed the benchmark task, the specialist was free to record, on the user-testing problem score sheet, any observations and/or comments made by the user (see Appendix G). Additionally, the specialist noted specifically if any of the following occurred:

1. Did the pilot give up on a task or ask for help?
2. Did the pilot perform an input error? (e.g., turn the wrong knob or press the wrong line-select key)
3. Did the pilot verbalize confusion and/or difficulty when performing the task?
4. Did the pilot show considerable delay in accomplishing part of the task? (e.g., 10 sec or more)

Upon task completion, the pilot was asked to interpret the displayed information while the specialist compared the pilot's interpretation of the information with the meaning supplied by the avionics manufacturer. After the user-testing session was complete, a debriefing session was conducted by the specialist to elicit comments and feedback about issues specifically regarding the efficacy of the system's interfaces.

User-CHAT

The User-CHAT was divided into two major phases (unstructured exploration and structured exploration) and required two people, the pilot (i.e., user) and the human factors specialist (i.e., supervisor) for each evaluation session. The User-CHAT session began with the unstructured exploration phase in which the pilot was allowed to freely explore the functionality of the system for approximately 10-15 min which was completed during pre-evaluation.

During structured exploration, the pilot completed each benchmark task, verbalizing his/her thought processes while stepping through each task. While the pilot completed each benchmark task, the specialist compared the pilot's performance to the "gold standard" (i.e., the most efficient action sequence for task completion). When pilot performance deviated from the gold standard action sequence, the specialist recorded the first inefficient action (e.g., an action that departs from the gold standard action sequence). Once the first inefficient action is recorded, the specialist then tallied each subsequent inefficient action until the pilot initiated the next required gold standard action towards task completion. It was stressed to the specialist to emphasize the correct identification (i.e., naming) of the first inefficient actions and estimate the number of subsequent inefficient actions. Additionally, during task completion, if the pilot exhibited an "extended amount of head-down time" (e.g., 10 sec or more contemplating the next step) the specialist indicated this excessive time spent with a "T" at the appropriate place on the User-CHAT score sheet (see Appendix H).

When the benchmark task was completed (i.e., the required information was displayed), the pilot interpreted the meaning and/or implication of the displayed information (e.g., symbology, terminology, etc.) while the specialist compared the pilot's interpretation with what was intended by the manufacturer. For each misunderstanding, both the pilot and specialist determined and documented why the meaning and/or implication of the information that was misinterpreted.

Then, for each first inefficient action that was initiated by the pilot, both participants discussed and identified the best answer to two questions:

1. Why was the gold standard action not initiated?
2. Why was the specific inefficient action initiated?

Once the aforementioned questions were answered, the pilot and specialist classified the reason(s) for the inefficient actions as violations of one or more general display design and usability heuristics (see Appendix I for the accompanying heuristics list used during the User-CHAT). Then both assigned a severity rating to the usability problem just discussed. After rating the severity of the usability problem, both participants discussed if memory influenced pilot performance. That is, was the inefficient action initiated because the next action was not clearly visible, descriptive, intuitive, etc. thereby causing

the pilot to rely on memory in order to recall functionality? If so, then the specialist placed an "M" in the Memory column at the appropriate place on the User-CHAT benchmark task score sheet.

Heuristic Evaluation

During the heuristic evaluation sessions, after the human factors specialist and the pilot read through the task[8], they then collaborated and judged whether the display elements (e.g., input devices, symbology, labeling, terminology, etc.) of the avionics system conformed to a list of display design heuristics and usability principles. Both pilot and specialist stepped through the gold standard action sequence for each benchmark task and evaluated the individual display elements along each step in the task action sequence. It was stressed to the participants to avoid simply completing the task and that they were to evaluate the system. If multiple sequences existed for a particular benchmark task (i.e., more than one gold standard was identified for the completion of the task), they were asked to evaluate the displays in all benchmark task sequences. Both identified usability problems and the heuristic(s) that were violated and recorded their findings on the heuristic evaluation data recording sheet (see Appendix J). Participants were asked to be as detailed as possible when describing a usability problem. If a usability problem could not be classified into one of the heuristics provided, they were instructed to create a new heuristic and classify the problem accordingly. Once a usability problem was classified, together the pilot and the specialist discussed and assigned a severity rating (see below) to each usability problem.

Cognitive Walkthrough

For each benchmark task, the human factors specialist and the pilot were instructed to read through the task and work together to sequentially evaluate the steps necessary to complete that task (i.e., the gold standard). Participants were instructed to avoid simply completing the task; rather they were instructed to assess how well the system supports task completion by exposing potential interface errors that may interfere with learning the avionics system. At each step in the action sequence, both discussed the answers to three of the four cognitive walkthrough questions:

[8] Doubleday et al. (1997) also incorporated a task-based approach for the heuristic evaluation in their comparative study.

1. Will the user initiate the correct goal? [However, because the goal was stated explicitly in the task, this question was not applicable. Thus, this question was not used in the present study.]

2. Will the user notice that the correct action is available? (e.g., see a menu option)

3. Will the user associate the correct action with the goal? (e.g., recognize that the menu option correlates with what is trying to be accomplished or instead perceive that all options look “wrong” or inconsistent with the goal)

4. If the correct action is performed, will the system provide feedback indicating that progress is being made toward completing the goal?

For each cognitive walkthrough question, the participants documented either a detailed success story or a detailed failure story. For success stories, they were to be explicit as to why the system does a good job. For failure stories, the pilot and specialist were instructed to not only include the usability problem, but include a description as to why the problem was an issue so designers could understand the usability problem they identified. The pilot and specialist documented their responses on the cognitive walkthrough problem report form (see Appendix K). Once a problem was identified, both the pilot and specialist assigned a severity rating.

Severity Ratings

Participants in the User-CHAT, heuristic evaluation, and cognitive walkthrough evaluation sessions assigned a severity rating (based on how the problem could or did affect pilot performance) immediately after each usability problem was identified. Three severity categories were used to classify usability problems: serious, intermediate, and minor (adapted from Sears, 1997).

Serious usability problems greatly hinder performance and continue to cause problems even after the problem has been experienced (e.g., a symbol is interpreted in the opposite way it was intended). These are usability issues that must be resolved because they are design or operation characteristics that could constitute a safety concern when using the avionics system.

Intermediate usability problems also hinder performance but can be overcome through experience (e.g., a menu option does not make sense before the first encounter). These are usability issues that are of great concern because they may have safety concerns and should be resolved but do not necessarily warrant a serious rating.

Minor usability problems are issues that do not hinder performance per se, but are recommendations on how to improve the system's design (e.g., the buttons are different sizes). These usability problems are not associated with safety concerns.

Summary of Procedure and Data Collection

The following table summarizes the evaluation steps and types of information collected across user-testing, User-CHAT, heuristic evaluation, and cognitive walkthrough evaluation sessions.

Table 4. Summary of the Evaluation Steps and the Information Gathered from each UEM.

Evaluation Steps / Information Collected During the Evaluation Sessions	*User-testing*	*User-CHAT*	*Heuristic Evaluation*	*Cognitive Walkthrough*
Non-Disclosure Agreement, Informed Consent, and Orientation to the Avionics System	√	√	√	√
Complete Demographics Questionnaires	√	√	√	√
10-15 min Free Exploration of MFD	√	√	√	√
Read Benchmark Task Description	√	√	√	√
Pilot and Human Factors Specialist Collaborate to Complete Task	--	--	√	√
Pilot Completes Task Individually, Human Factors Specialist Records Pilot Performance	√	√	--	--
Interpret Displayed Information	√	√	√	√
Identify Usability Problems	√	√	√	√
Diagnose Usability Problems	√	√	--	√
Pilot and Human Factors Specialist Classify Usability Problems into Violations of Heuristics	--	√	√	--
Assign Severity Ratings	√*	√	√	√
Head-down Time Recorded	√	√	--	--
Other Notes, Comments, Concerns, etc. Documented	√	√	√	√
Evaluation Session Video Recorded	√	√	√	√

** Severity ratings were obtained after all the usability problems were identified and transcribed (see below for details).*

Data Synthesis

The video recordings for the user-testing sessions were reviewed and any additional usability problems that were not documented during the evaluation sessions were transcribed. The four user-testing questions and the benchmark tasks' gold standards served as guides to identify additional usability problems. Pilot comments, observations, and concerns that were missed (i.e., not documented)

during the user-testing session were transcribed also. The combination of the usability problems documented during the user-testing sessions and those transcribed from the video recordings comprised the comparison list of usability problems upon which the performance of the other UEMs will be compared.

User-testing is the *de facto* standard often used to produce a list of usability problems (i.e. real usability problems) on which to compare candidate usability techniques because the problem list is generated by representative end-users performing typical tasks and the usability problems identified actually impacted user performance. However, because some usability problems identified from user-testing can be questionable (e.g., a particular issue may impact one user only), Hartson et al. (2001) recommended that representative end-users review the list of usability problems generated through user-testing to ensure that the list, in their opinions, is comprised of issues that actually affect user performance. The six pilots who participated in the user-testing sessions reviewed independently the usability problem list after it was compiled (weeks after their sessions were completed) and assigned a severity rating (using the same severity categories discussed above) to each usability problem found through user-testing. Thus, the quality of the comparison usability list used to represent usability problems that actually impact users was improved by conducting this additional review.

Pilots reviewed each usability problem and marked whether or not they thought the issue would affect pilot performance and should be resolved (i.e., a “yes” response indicated the usability problem should be fixed while “no” indicated the usability problem was not an issue). Responses from all six pilots were used to determine which usability problems should be included in the standard list and which ones should be removed. Specifically, for those usability problems that affected one or two pilots only during the user-testing sessions, responses from the six pilots were used to determine if those usability problems were true usability issues. Appendix A provides further explanation of the process of generating and verifying a standard list of usability problems yielded through user-testing. For disagreements regarding whether or not usability problems were issues, the rubric depicted in Table 5 was used to determine if a usability problem required further evaluation in order to ascertain if the usability problem should be included into the comparison list.

Table 5. Rubric for Determining Whether a Usability Problem Identified through User-testing Required Further Evaluation in order to be Included in the Comparison List of Usability Problems.

Number of Pilots Who Indicated a Problem was an Issue	*Number of Pilots Who Indicated a Problem was not an Issue*	*Usability Problem Further Evaluated?*	*Usability Problem Included in Standard List?*
---	6 Pilots	No	***Did not*** include problem in the list
1 Pilot	5 Pilots	No	***Did not*** include problem in the list
2 Pilots	4 Pilots	Yes	Determined upon further evaluation
3 Pilots	3 Pilots	Yes	Determined upon further evaluation
4 Pilots	2 Pilots	Yes	Determined upon further evaluation
5 Pilots	1 Pilot	No	***Included*** problem in the list
6 Pilots	---	No	***Included*** problem in the list

Note: This rubric was used for instances where the usability problem was identified through the performance of one or two pilots only. Usability problems identified through the performance of three or more pilots were automatically included in the comparison list of usability problems.

Based on this rubric, usability problems where at least two pilots identified an issue as a problem (while the other four did not) were evaluated further. That is, usability problems of questionable importance were further assessed to determine if the usability problem should be included in the comparison list of user-testing usability problems in cases where there was at least one-third disagreement. The decision to include that usability problem in the comparison list was ultimately made by weighing the severity ratings for that usability issue. For instance, the usability problem in question must be rated as either an intermediate or serious issue by at least two pilots in order for the usability problem to be included in the comparison list.

Classifying Usability Problems into Heuristics

The classification of each usability problem into one or more violated heuristics for all UEMs was used to ascertain the types of usability problems (e.g., consistency, top-down processing, etc.) that a particular UEM is likely to detect. Recall that usability problems identified through the User-CHAT and the heuristic evaluation were classified into violations of heuristics during the evaluation sessions. However, the usability problems identified through user-testing and the cognitive walkthrough require classification that is beyond their traditional protocols. In order to obtain an index of the types of usability problems

found by each UEM, two additional human factors professionals[9] recruited from industry, independently classified usability problems found in each of the UEMs as a violation of one or more display design heuristics. The heuristics list that accompanied the User-CHAT was used for classifying usability problems because it contained more detailed and aviation specific heuristics relative to Nielsen's (1994b) display design heuristics list. Both human factors professionals were instructed to classify the usability problems into as many heuristics as deemed appropriate. Once the independent classifications were completed, the classification lists were combined for each UEM.

Determining Unique or Shared Usability Problems

The same two human factors professionals who classified usability problems into heuristics also evaluated and sorted independently the usability problems identified from all the UEMs. After reading through the entire description of all the usability problems, these professionals grouped together usability problems that described similar issues; idiosyncratic usability problems remained ungrouped.[10] Specifically, the usability problems fell into one of three categories:

1. The usability problem was the same as one identified through user-testing,
2. The usability problem was not the same as one identified through user-testing but was the same as one identified by at least one other UEM, or
3. The usability problem was unique to a specific UEM (i.e., not identified through any other UEM or user-testing).

Parsing the usability problems into the aforementioned categories permitted the comparison between usability techniques using contemporary UEM performance metrics. Task sheets from the evaluation and screen shots of the avionics system were available to provide context for each usability problem. After sorting the usability problems independently, the human factors specialists discussed and resolved any grouping differences and produced one master list of sorted usability problems, combining their independent groupings and resolving disagreements on a consensus basis. This final grouping provided the foundation for subsequent analyses.

[9] These human factors professionals had advanced degrees (an M.S. and Ph.D.) and each had over 2 ½ years of applied human factors and usability experience including designing and evaluating system interfaces.

[10] It is important to note that these individuals were blind to the experimental design and research intentions of the comparative study.

Chapter 3 -- Results

According to Stevens (1996), small sample sizes (N < 20) justify use of a more liberal alpha level in order to improve power. Therefore, subsequent inferential statistical analyses were conducted with alpha (α) maintained at .10. Additionally, due to small sample size and the applied nature of the current study, it is important to examine UEM performance in terms of practical significance versus statistical significance. Statistical significance refers to whether an effect exceeds a predetermined alpha level; practical significance refers to whether the statistical significance is deemed useful or not in a particular context. Practical significance depends upon context where researchers' judgments determine ultimately if a significant effect is large enough to be important (Light, Singer, & Willett, 1990). For instance, while an effect may demonstrate statistical significance, the differences may be minute or there is considerable consensus among researchers that the application of the results is not worthwhile or practical.

One effective computational technique to illustrate practical significance is through confidence intervals (Stevens, 1996). As with common inferential analyses, confidence intervals[11] utilize a predetermined alpha level. The alpha level for constructing confidence intervals was set at .05 (i.e., 95% confidence intervals) throughout. While non-overlapping confidence intervals and significant inferential tests suggest that the UEMs were statistically different for a given comparative metric, the practical application of the results still resides with the researchers.

Time Spent per Usability Problem

For each evaluation session within each usability technique, the amount of time taken to evaluate the avionics system, the number of usability problems identified, and the average amount of time spent identifying a usability problem were computed (see Table 6). Evaluation time was the sum of time taken to assess the system. For each task, evaluation time started once the task was read and was stopped once the evaluator(s) indicated verbally that the task was complete. Time indices (sec) for each task were captured from video recordings. The number of usability problems represented the sum of distinct (i.e., unique or non-repeated) usability problems identified by each evaluation session. Average time taken to

[11] The reader is referred to Cumming, Williams, and Fidler (2004) for a discussion of how to appropriately interpret confidence intervals.

document a usability problem was calculated by dividing evaluation time by the number of usability problems identified for each evaluation session.

Table 6. Summary Data for Evaluation Time, Number of Usability Problems Identified, and Average Time to Identify a Usability Problem.

UEM	*Session #*	*Evaluation Time (H:MM:SS)*	*# of Usability Problems Identified*	*Average Time to Identify a Usability Problem (MM:SS)*
User-CHAT	1	2:03:51	46	2:42
	2	1:51:13	41	2:43
	3	1:12:54	28	2:36
	4	1:36:06	34	2:50
	5	1:27:56	35	2:31
	6	1:56:05	36	3:13
	Total	10:08:05	220	16:34
	M	1:41:21	36.67	2:46
	SD	0:19:10	6.19	0:15
Heuristic Evaluation	1	2:24:03	59	2:26
	2	1:38:45	44	2:15
	3	3:35:49[a,c]	116	1:52
	Total	7:38:37	219	6:33
	M	2:32:52	73.00	2:11
	SD	0:59:02	37.99	0:18
Cognitive Walkthrough	1	5:19:56[a]	34	9:25
	2	3:47:58[b]	30	7:36
	3	2:15:19	57	2:22
	Total	11:23:13	121	19:23
	M	3:47:44	40.33	6:28
	SD	1:32:19	14.57	3:39

[a] *Completed the first nine tasks only within three hours.*

[b] *Completed the first 13 tasks only within three hours.*

[c] *Due to a simulator malfunction, tasks 10-12, 17, and 18 were not evaluated. Subsequent results and conclusions were the same when these tasks were removed.*

Note: Analyses conducted on the first nine tasks (adhering to the 3-hour evaluation limit) yielded similar results and conclusions relative to the results and conclusions from analyses conducted on all the tasks.

Data for evaluation time, the number of usability problems identified, and the average time taken to identify a usability problem were subjected to a multivariate analysis of variance (MANOVA) in order to test for differences between the User-CHAT, heuristic evaluation, and cognitive walkthrough. Using Wilks' Lambda (λ) as the test statistic, results from the MANOVA indicated significance between the three

usability methods, Willks' $\lambda = 0.24$, $F(6,14) = 2.45$, $p<.10$. Based on Wilks' λ, subsequent univariate analysis of variance (ANOVAs) were conducted to further examine for differences between the three techniques. Univariate summary tables are presented in Table 7.

A univariate ANOVA conducted on the total time taken to evaluate the system revealed significance between the three methods, $F(2,9) = 5.60$, $p<.05$. On average each session for the User-CHAT, heuristic evaluation, and cognitive walkthrough took approximately 1hr 42min, 2hr 33 min, and 3hr 48 min, respectively to complete the evaluation of the avionics system. Post hoc Tukey's HSD (see Table 8) comparisons revealed that heuristic evaluation sessions were similar in evaluation time when compared to User-CHAT sessions and cognitive walkthrough sessions. Post hoc comparisons showed also that User-CHAT sessions took less time to evaluate the system than cognitive walkthrough sessions. While only one post hoc comparison showed significance, the practicality of these results suggests that the User-CHAT completed the evaluation 50 min faster than the heuristic evaluation, which was in turn more than an hour faster than the cognitive walkthrough.

The mean number of usability problems identified by each session was examined also with a univariate ANOVA. Results showed significance between the three usability methods, $F(2, 9) = 3.60$, $p<.10$. Tukey's HSD post hoc comparisons indicated that cognitive walkthrough sessions ($M = 40.33$) identified similar numbers of usability problems relative to heuristic evaluation sessions ($M = 73.00$) and User-CHAT sessions ($M = 36.67$). Additionally, heuristic evaluation sessions identified more problems than the User-CHAT sessions. However, data for the heuristic evaluation and cognitive walkthrough were each influenced greatly by one session that identified almost twice as many problems relative to the other two.

Finally, univariate ANOVAs revealed significance between the usability methods in terms of the average amount of time spent on identifying each usability problem, $F(2,9) = 5.75$, $p<.05$. Post hoc comparisons showed that the cognitive walkthrough (M = 6 min 28 sec) took longer to identify a usability problem than both the heuristic evaluation (M = 2 min 15 sec) and User-CHAT (M = 2 min 46 sec); the User-CHAT and heuristic evaluation did not differ. The results indicated that cognitive walkthrough sessions took more than twice as much time to identify a usability problem than User-CHAT and heuristic evaluation sessions.

Table 7. ANOVA Summary Table for Evaluation Time, Number of Usability Problems Identified, and Average Time Taken to Identify each Usability Problem.

Source	*Dependent Variable*	*df*	*SS*	*MS*	*F*	*p*	*Power*	ω^2
	Evaluation Time							
Method		2	115733814.70	57866907.38	5.60	<.05	0.71	0.43
Error		9	93054342.17	10339371.35				
Total		11	208788156.90					
	Mean # of Problems Identified							
Method		2	2800.67	1400.33	3.60	<.10	0.51	0.30
Error		9	3502.00	389.11				
Total		11	6302.67					
	Average Time per Usability Problem							
Method		2	125129.24	62564.62	5.75	<.05	0.73	0.44
Error		9	97853.02	10872.56				
Total		11	222982.26					

MANOVA Statistics: Wilks' Lambda (λ) = 0.24, F(6,14) = 2.45, p<.10

Table 8. Tukey's HSD Post Hoc Comparisons for Evaluation Time (format = H:MM:SS), Number of Usability Problems Identified, and Average Time Taken to identify a Usability Problem (format = M:SS).

Dependent Variable	*(I) Usability Method*	*(J) Usability Method*	*Mean Difference (I – J)*
Evaluation Time	User-CHAT	Heuristic Evaluation	-0:51:31
		Cognitive Walkthrough	-2:06:23**
	Heuristic Evaluation	Cognitive Walkthrough	-1:14:52
Mean # of Problems Identified	User-CHAT	Heuristic Evaluation	-36.33*
		Cognitive Walkthrough	-3.67
	Heuristic Evaluation	Cognitive Walkthrough	32.67
Average Time per Usability Problem	User-CHAT	Heuristic Evaluation	0:35
		Cognitive Walkthrough	-4:17**
	Heuristic Evaluation	Cognitive Walkthrough	-3:42**

**p<.10, **p<.05*

Usability Problem Partitions

A total of 78 different usability problems were identified through user-testing. The User-CHAT, heuristic evaluation, and cognitive walkthrough identified a combined total of 329 distinct usability problems. Of these 329 problems, 59 were shared with the 78 problems identified through user-testing.

Nineteen usability problems from user-testing were unaccounted for (missed) by the UEMs. Figure 3 presents a Venn diagram that illustrates how the usability problems partitioned according to problems unique to each usability technique and common to two or more methods. The total number of usability problems identified by each individual usability technique is presented in parentheses[12]. As can be seen in the diagram, the heuristic evaluation (187) identified more distinct usability problems than the User-CHAT (121) and the cognitive walkthrough (100). As shown in Table 9, Chi square (χ^2) tests reported significant differences between all three methods on the total number of usability problems found by each usability method. Further tests showed that the heuristic evaluation identified more usability problems than the User-CHAT, which identified more usability problems than the cognitive walkthrough. In other words, Chi-square results suggest that the heuristic evaluation achieved higher "aggregate" thoroughness than the User-CHAT, which achieved higher "aggregate" thoroughness than the cognitive walkthrough. The relatively small number of overlapping usability problems between the UEMs indicates that each technique identified different types of usability problems. This pattern is in accordance with previous studies (e.g., Fu, Salvendy, & Turley, 2002; Karat et al., 1992; Desurvire et al., 1992; Jeffries et al., 1991). Approximately 68% (106/329) of usability problems identified by all three UEMs were identified in the first nine tasks, which focused primarily on issues associated with display elements.

[12] The usability problem totals for the Venn diagram (Figure 3) and the problem totals presented in Table 6 differ because the diagram accounts for common usability problems across evaluation sessions within a given UEM while the totals presented in table represent the absolute number of problems identified per evaluation session where common problems across sessions are counted each time the usability problem occurs. For instance, if a usability problem was found in three sessions for a given usability technique, in Table 6 that usability problem was counted three times while for the Venn diagram, that problem's frequency was reduced to one for that UEM. Therefore, the problem totals in the Venn diagram will be less than the problem totals presented in Table 6.

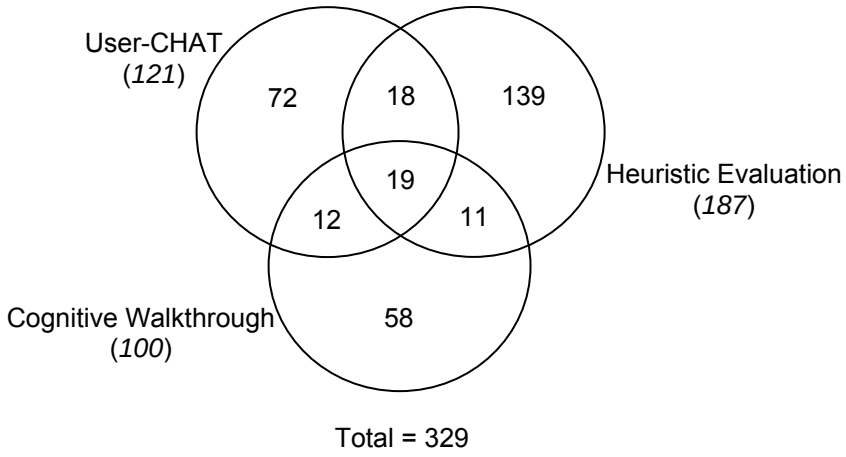

Figure 3. Venn diagram illustrating usability problem partitions for each UEM. The total number of usability problems identified by each method is presented in parentheses.

Table 9. Summary of Chi-Square (χ^2) Tests of Independence for Total Number of Usability Problems Identified by each UEM.

Comparison	χ^2	*df*
User-CHAT (121) vs. Heuristic Evaluation (187) vs. Cognitive Walkthrough (100)	56.56*	2
User-CHAT (121) vs. Heuristic Evaluation (187)	21.13*	1
User-CHAT (121) vs. Cognitive Walkthrough (100)	7.38*	1
Heuristic Evaluation (187) vs. Cognitive Walkthrough (100)	32.95*	1

**p<.05*

Detection Rates

The mean number of problems detected by each evaluation session for each usability method was used to calculate detection rates for their respective usability method. The following formula, $1-(1-p)^n$, where *p* is the probability of detecting a given usability problem (i.e., mean detection rate) and *n* is the number of evaluators, was used to plot detection rates (i.e., curves of diminishing returns) for each usability technique. Mean detection rates were calculated by dividing the mean number of problems for each method by the total number of usability problems identified across all three methods (i.e., the number of usability problems in the system). Many researchers (e.g., Nielsen, 1994b; Virzi, 1992; Wright & Monk, 1991) have demonstrated that $1-(1-p)^n$ is a sufficient equation to approximate of the number of evaluators necessary to attain a specific detection rate level.

Visual inspection of the curves of diminishing returns illustrated in Figure 4 shows that the six evaluators who used the heuristic evaluation detected between 75-80% of the total number of usability problems. The cognitive walkthrough and User-CHAT, however, each detected between 50-55% of the total number of usability problems using six evaluators. Detection rates suggest that the heuristic evaluation had sufficient number of evaluators while more than 10 evaluators would be required by either the cognitive walkthrough or User-CHAT in order to obtain an 80% detection rate.

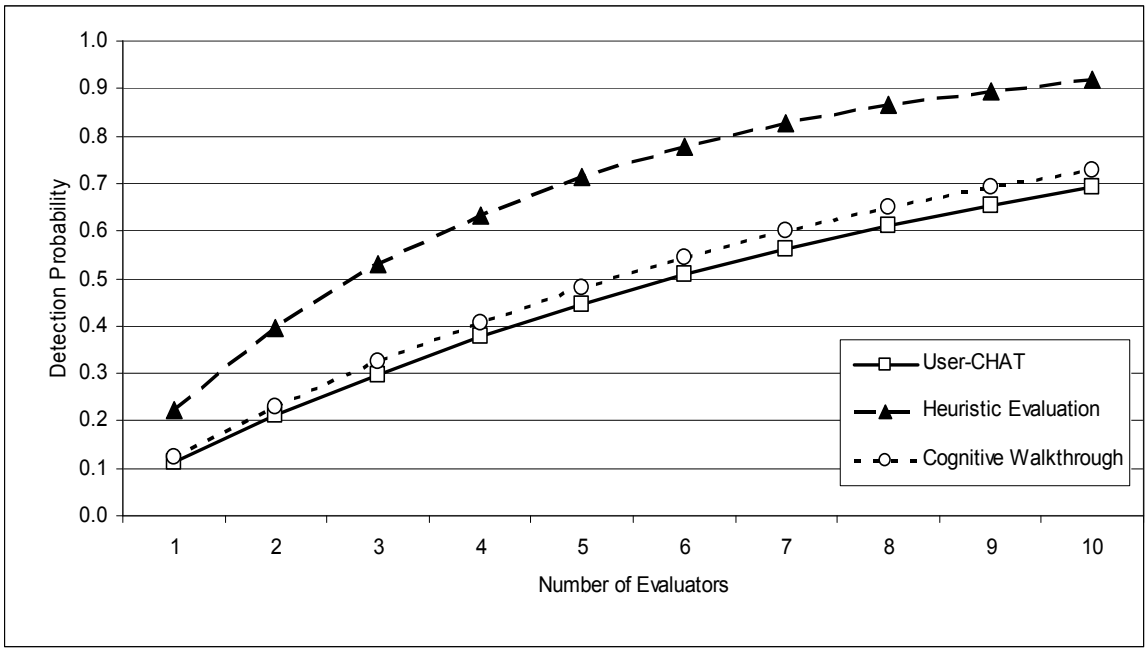

Figure 4. Proportion of the total usability problems identified combined from all UEMs that different numbers of evaluators were likely to detect.

Previous usability research reported that five participants would be sufficient to detect approximately 80% of the usability problems in an interface (Nielsen, 1994b; Virzi, 1992; Wright & Monk, 1991). However the aforementioned researchers used only a single usability method in their studies to establish these guidelines. Accordingly the mean detection rate for any given usability method was based solely on the total number of usability problems it identified. In the present comparison study, however, the total number of usability problems was based on the number of usability problems identified by all three usability methods combined. Thus, the figures presented above reflect detection curves for each method's ability to detect usability problems that were identified by all three methods. Therefore, in order

to allow a more direct comparison with previous research, detection rates for the present study were re-plotted. This time mean detection rates were based only on the respective problem totals for each usability method. That is, mean detection rates for each method were based on the total number of usability problems each method identified and not the total number of usability problems identified by all three. For instance, User-CHAT mean detection rates were based on 121 total usability problems, not 329. Detection rates calculated in this manner may be informative to usability practitioners when deciding on a UEM to incorporate into their usability lab as the number of evaluators required to obtain a certain detection level for a given UEM may sway the usability practitioner's decision.

When detection rates are plotted relative to the total number of usability problems identified by each method, the data tell a different story. As illustrated in Figure 5, regardless of the method applied, six evaluators detected approximately 85-95% of total number of usability problems identified by that particular method. These results reflect the detection rate ranges commonly published throughout much of the HCI literature (e.g., Wright & Monk, 1991; Virzi, 1992; Nielsen, 1994b).

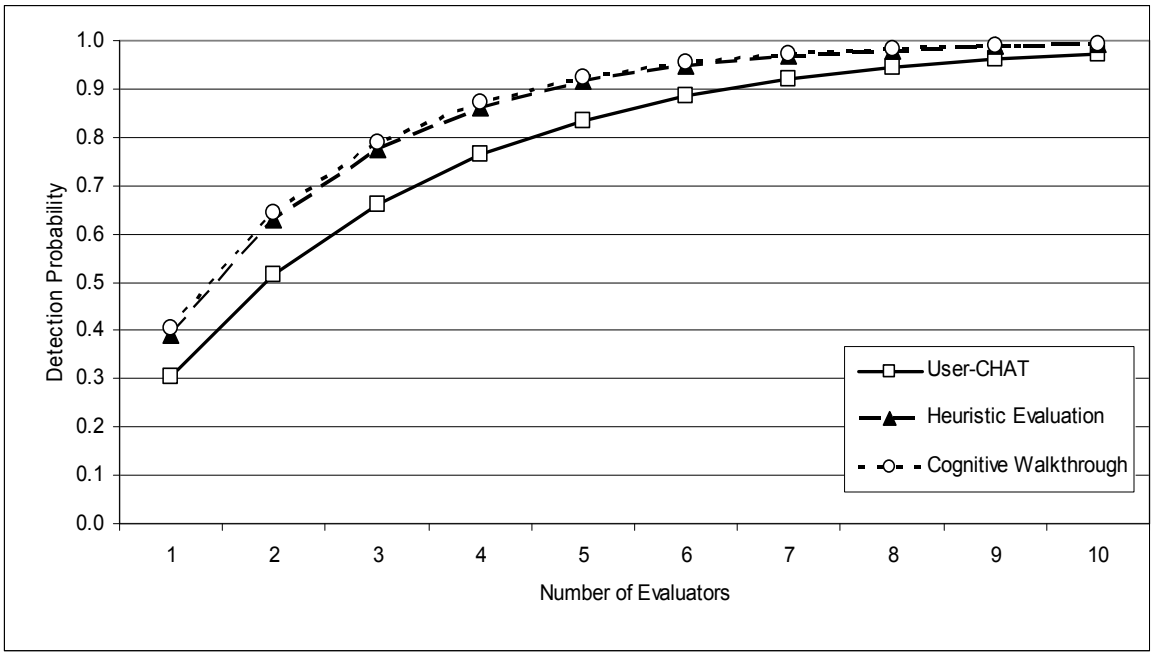

Figure 5. Proportion of the usability problems identified by each UEM only that different numbers of evaluators were likely to detect.

Thoroughness, Validity, and Effectiveness

Thoroughness and validity indices were calculated using the following equations discussed above:

$$\textit{“Real” Thoroughness} = \frac{|P \cap A|}{|A|} \quad (2)$$

$$\textit{Validity} = \frac{|P \cap A|}{|P|} \quad (4)$$

where P represents the set of usability problems detected by one of the usability methods and A represents the set of usability problems identified through user-testing. Thus, A is defined as the set of real usability problems that existed in the target avionics system. Effectiveness is a multiplicative component of thoroughness and validity, defined by the following equation:

$$\textit{Effectiveness} = \textit{“Real” Thoroughness x Validity} \quad (9)$$

UEM performance data are presented from two perspectives: *per evaluation session* and *per usability method*. The following example illustrates the difference. Assume three evaluation sessions utilizing the same UEM detected 10, 15, and 20 real usability problems respectively; assume also that five real usability problems were shared across all three sessions. Per evaluation session reflects the averages for thoroughness, validity, and effectiveness across all three evaluation sessions. While these scores index the approximate performance of a single evaluation session for a given usability method, the performance of the usability technique as a whole is not reflected. For example, in addition to the five shared real usability problems, each session contributed 5, 10, and 15 additional unique real usability problems respectively. Collapsed across all evaluation sessions, the usability method as a whole identified 35 total real usability problems (as opposed to averaging 15 real usability problems per evaluation session). Therefore, per usability method reflects the contribution of each session's shared and unique usability problems to the UEM's overall performance.

Means, standard deviations, and 95% confidence intervals for thoroughness, validity, and effectiveness per evaluation session are shown in Table 10 and illustrated in Figure 6. Comparative measures were subjected to a MANOVA to test for performance differences between the User-CHAT, heuristic evaluation, and cognitive walkthrough. The MANOVA detected significance between the three methods, Willks' $\lambda = 0.06$, $F(6,14) = 7.00$, $p<.05$. Because Willks' λ was significant, subsequent univariate ANOVAs were conducted.

Table 10. Per Evaluation Session Means, Standard Deviations, and 95% Confidence Intervals on Thoroughness, Validity, and Effectiveness when Compared to User-testing Data.

	User-CHAT			*Heuristic Evaluation*			*Cognitive Walkthrough*		
Metric	*M*	*SD*	*95% CI*	*M*	*SD*	*95% CI*	*M*	*SD*	*95% CI*
T	0.24	0.06	0.28<μ<0.19	0.20	0.01	0.22<μ<0.18	0.18	0.06	0.24<μ<0.11
V	0.51	0.10	0.59<μ<0.43	0.25	0.10	0.36<μ<0.14	0.34	0.01	0.35<μ<0.33
E	0.13	0.05	0.17<μ<0.09	0.05	0.02	0.07<μ<0.03	0.06	0.02	0.08<μ<0.04

T = thoroughness, V = validity, E = effectiveness

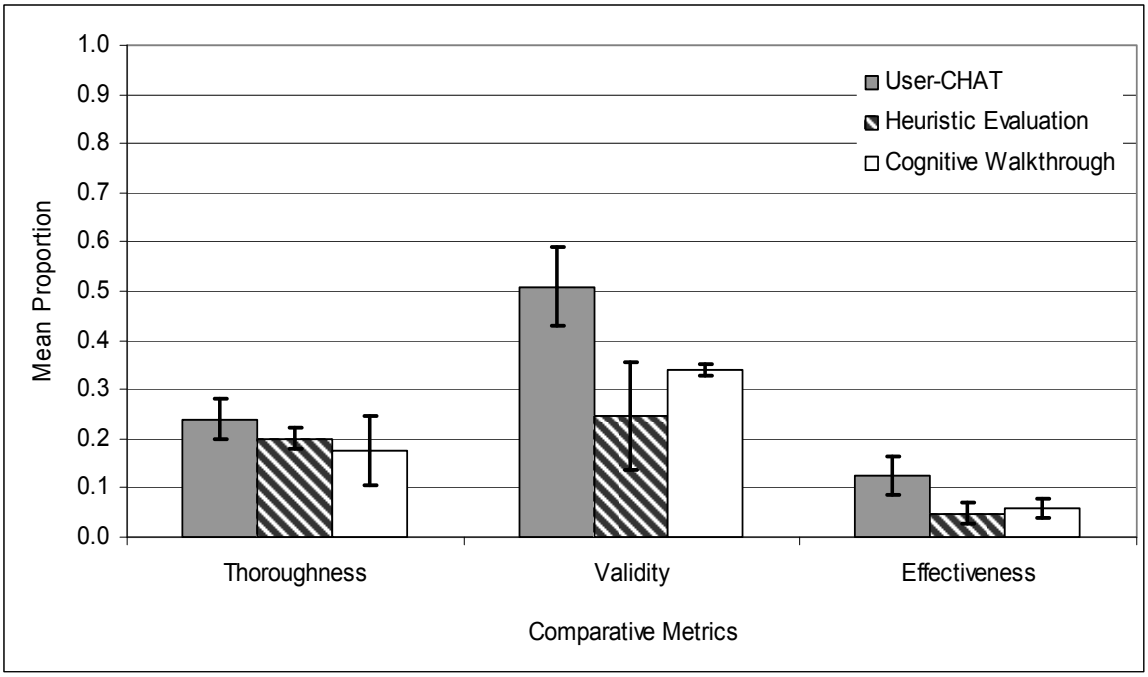

Figure 6. Per evaluation session thoroughness, validity, and effectiveness 95% confidence intervals for each UEM (measured against user-testing data).

Visual examination of the confidence intervals and univariate ANOVA results (see Table 11) demonstrated that all three usability techniques attained similar thoroughness scores (i.e., similar numbers of real problems detected – "hits"), $F(2,9)=1.68$, *ns*. Therefore, when compared to the list of usability problems yielded through user-testing, all three techniques had similar detected similar numbers of real usability problems (between 15-25% of the real problems were detected by each individual evaluation session).

Confidence intervals suggest that the User-CHAT attained higher validity scores than the heuristic evaluation and cognitive walkthrough. Results from the univariate ANOVA support this conclusion, $F(2,9)=10.40$, $p<.05$. Tukey's HSD post hoc comparisons (see Table 12) indicated that the User-CHAT obtained a higher validity score (i.e., detected fewer numbers of trivial or non-real usability problems) than the heuristic evaluation, which had a similar validity scores relative to the cognitive walkthrough. The data imply that the User-CHAT's list of usability problems contained a higher proportion of real usability problems than the heuristic evaluation and cognitive walkthrough. Upon visual examination of the validity scores, approximately 50%, 23%, and 35% of the usability problems in the lists generated by the User-CHAT, heuristic evaluation, and cognitive walkthrough respectively were real usability problems.

Thoroughness and validity indices combine to illustrate a usability method's effectiveness. Upon examination of the confidence intervals, it appears that the User-CHAT was more effective than the other two methods. Univariate ANOVA results found significance between the methods, $F(2,9)= 5.03$, $p<.05$. Post hoc comparisons showed that the User-CHAT had marginally higher effectiveness indices than the heuristic evaluation and cognitive walkthrough, suggesting that it is able to detect more real problems and ignore trivial usability problems relative to the two other methods. The heuristic evaluation and cognitive walkthrough shared similar effectiveness scores.

Table 11. Univariate ANOVA Summary Tables for Thoroughness, Validity, and Effectiveness.

Source	*Dependent Variable*	*df*	*SS*	*MS*	*F*	*p*	*Power*	ω^2
	Thoroughness							
Method		2	.009	.004	1.68	>.10	0.27	0.09
Error		9	.023	.003				
Total		11	.032					
	Validity							
Method		2	.155	.078	10.40	<.05	0.94	0.62
Error		9	.067	.007				
Total		11	.222					
	Effectiveness							
Method		2	.015	.007	5.03	<.05	0.66	0.45
Error		9	.013	.001				
Total		11	.028					

MANOVA Statistics: Wilks' Lambda (λ) = 0.06, F(6,14) = 7.00, p<..05

Table 12. Tukey's HSD Post Comparisons for Thoroughness, Validity, and Effectiveness.

Dependent Variable	*(I) Usability Method*	*(J) Usability Method*	*Mean Difference (I – J)*
Thoroughness	User-CHAT	Heuristic Evaluation	.04
		Cognitive Walkthrough	.06
	Heuristic Evaluation	Cognitive Walkthrough	.03
Validity	User-CHAT	Heuristic Evaluation	.26**
		Cognitive Walkthrough	.17**
	Heuristic Evaluation	Cognitive Walkthrough	-.09
Effectiveness	User-CHAT	Heuristic Evaluation	.07*
		Cognitive Walkthrough	.07*
	Heuristic Evaluation	Cognitive Walkthrough	-.01

**p<.10, **p<.05*

The data presented above illustrate averaged performance across all evaluation sessions with a given usability method; the performance of each UEM as a whole was not presented. Thus, comparative metrics reflecting each usability technique's overall performance (e.g., per usability method) are presented in Table 13 and illustrated in Figure 7. The table presents the total number of real usability problems established through user-testing (78). Also presented are the total usability issues and number of real problems identified by each technique.

The data show that the User-CHAT, as a whole, was more thorough as it identified at least 16 more real usability problems than the heuristic evaluation and at least 21 more than the cognitive walkthrough. The heuristic evaluation detected at least five more real usability problems than the cognitive walkthrough. In terms of hit rates, the User-CHAT detected over 60% of the usability problems identified through user-testing while the heuristic evaluation and cognitive walkthrough detected approximately 41% and 35% respectively.

Similarly the User-CHAT's ability to attenuate detection of trivial usability problems was better than the other two methods. The User-CHAT detected fewer trivial usability problems (e.g., had higher validity) than both the heuristic evaluation and cognitive walkthrough; the cognitive walkthrough attained higher validity indices than the heuristic evaluation. That is, the User-CHAT had a 61% false alarm rate

while the heuristic evaluation and cognitive walkthrough had 83% and 72% false alarm rates respectively. Thus, the data imply that the User-CHAT is better than the other two methods at reducing the number of non-real usability problems identified.

Finally, effectiveness indices illustrate that the User-CHAT was better than both the heuristic evaluation and cognitive walkthrough at detecting more real usability problems and reducing the detection of less important usability problems. The heuristic evaluation and cognitive walkthrough seem to have similar effectiveness scores. Thus, when compared to the list of usability problems generated through user-testing, the User-CHAT found more real problems and fewer trivial problems, thereby outperforming both the heuristic evaluation and cognitive walkthrough. The effectiveness score for the User-CHAT indicated that it was more than twice as effective as the other two methods.

Table 13. Thoroughness, Validity, and Effectiveness Scores based on each UEM's Overall Performance when Measured Against User-testing Data.

		User-CHAT	*Heuristic Evaluation*	*Cognitive Walkthrough*
Total "Real" Problems	78			
# of Problems Identified		121	187	100
# of "Real" Problems Identified		48	32	27
Thoroughness		0.62	0.41	0.35
Validity		0.40	0.17	0.27
Effectiveness		0.24	0.07	0.09

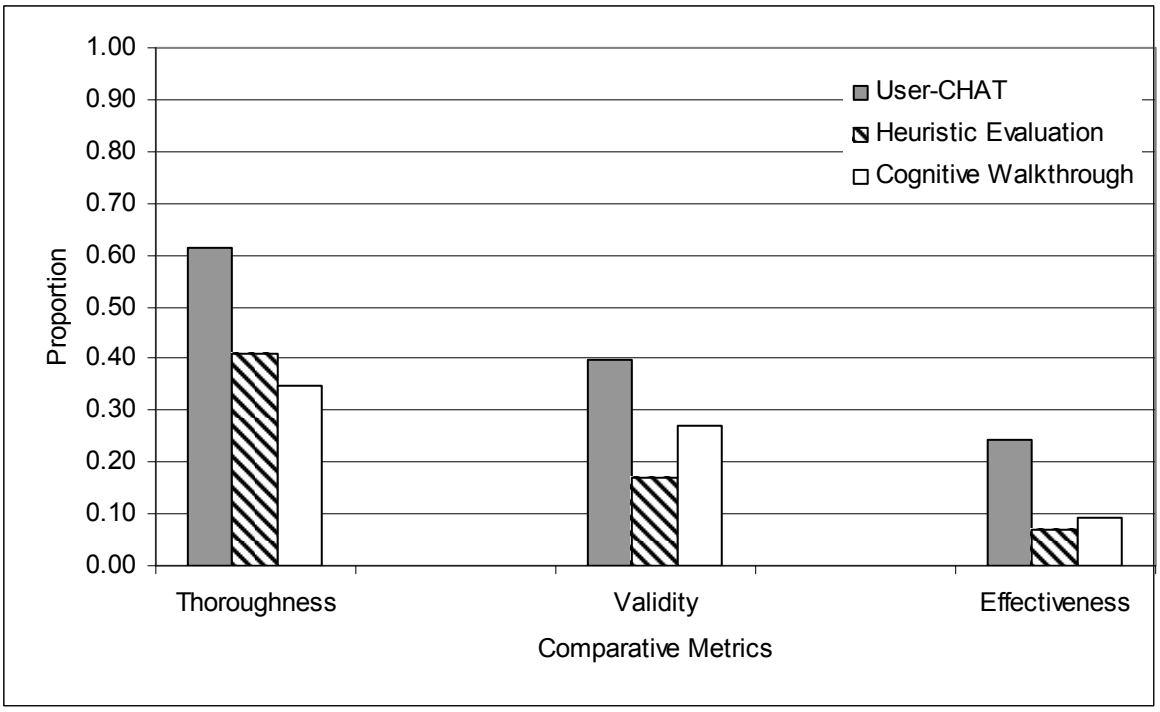

Figure 7. Overall UEM performance for thoroughness, validity, and effectiveness (when measured against user-testing data).

Reliability

The reliability equation postulated by Sears (1997) reflects inter-rater reliability for real usability problems only:

$$Reliability_{temp} = 1 - \frac{Stdev\ (\#\ of\ real\ problems\ found)}{Average\ (\#\ of\ real\ problems\ found)} \quad (10)$$

$$Reliability = \text{Maximum}(0, \text{Reliability}_{temp}) \quad (11)$$

A modification to the equation can reflect reliability scores for the total number of usability problems found, real and non-real. Thus reliability scores based on the total number of usability problems were calculated also using the following adapted equation:

$$Reliability_{temp} = 1 - \frac{Stdev\ (\#\ of\ usability\ problems\ found)}{Average\ (\#\ of\ usability\ problems\ found)} \quad (13)$$

$$Reliability = \text{Maximum}(0, \text{Reliability}_{temp}) \quad (14)$$

The equation above reflects the reliability score for all the usability problems detected for a specific usability technique. By removing the "real" criterion, inter-rater reliability indices for a given usability technique can show how well evaluators applying the same usability technique were able to identify

similar numbers of usability problems. Thus, reliability was calculated using all usability problems and not just real usability problems.

Table 14 presents each UEM's reliability indices for the total number of problems found and the number of real problems found. When reliability is calculated based on the total number of usability problems identified, it appears that the User-CHAT had higher reliability scores than the cognitive walkthrough which had higher reliability scores than the heuristic evaluation. Thus, evaluators applying the User-CHAT had greater similarity in the numbers of usability problems detected between sessions than the heuristic evaluation and cognitive walkthrough. In terms of evaluators detecting similar numbers of real usability problems, the data tell a different story. Generally the heuristic evaluation achieved higher reliability than both the User-CHAT and cognitive walkthrough, which appear to have similar inter-rater reliabilities for the number of real problems identified. Thus, when inter-rater reliability is calculated using the numbers of real problems identified, evaluators applying the heuristic evaluation were more likely to identify similar numbers of real problems, more so than the cognitive walkthrough and User-CHAT.

Table 14. Inter-rater Reliability Indices for Real and Total Number of Usability Problems Identified

	User-CHAT	*Heuristic Evaluation*	*Cognitive Walkthrough*
Total Number of Usability Problems	0.83	0.48	0.64
Real Usability Problems	0.77	0.93	0.65

While the reliability equations above examine inter-rater agreement based on either the number of real or total usability problems found by each evaluation session, the equations were not sensitive enough to determine whether or not the "same" usability problem (e.g., the same unreadable display element or an unlabeled input device) was detected across evaluation sessions for a given UEM. To many usability researchers, measuring inter-rater consistency for identifying a similar number of usability problems is not informative (Andre et al., 1999). Therefore, inter-rater reliability was re-examined for each UEM using percent raw percent agreement scores in order to determine whether or not the same usability problem was detected across evaluation sessions (c.f., Kessner, Wood, Dillon, & West, 2001).

Recall that there were six User-CHAT sessions and three heuristic evaluation and cognitive walkthrough sessions respectively. Thus, percent raw agreement for the User-CHAT would not scale the same as the heuristic evaluation and cognitive walkthrough. The percent agreement for the User-CHAT would be out of six sessions, while the other two methods would be based on percent agreement out of three sessions. Therefore, in order to ensure that all three methods were measured on approximately similar scales, weighted means were used to calculate percent raw agreement.

Using the following equations, all usability problems for each method were weighted by the number of sessions that found the same problem:

$$\textit{Weighted Mean} = \sum_{i=1}^{X} W_i X_i \tag{15}$$

where W_i represents the weight and X_i is the proportion of evaluation sessions for a given UEM that identified the same usability problem. For each weighted mean, weighted standard deviations were calculated using the following equation:

$$\textit{Weighted SD} = \sqrt{\frac{\sum_{i=1}^{N} W_i (X_i - \overline{X}_w)^2}{\frac{(N' - 1)\sum_{i=1}^{N} W_i}{N'}}} \tag{16}$$

where *N'* is the number of weights. Mean percent raw agreement scores and standard deviations are presented in Table 15. The data suggest that all three methods attained similar inter-rater agreement percentages. It appears that, regardless of the UEM applied, evaluators detected similar percentages of the same usability problems across evaluation sessions however none of the percent agreements for all UEMs were above 50%, suggesting that more than half the time evaluators failed to detect the same usability problem.

Table 15. Means and Standard Deviations for Weighted Inter-rater Percent Raw Agreement Scores.

	User-CHAT	*Heuristic Evaluation*	*Cognitive Walkthrough*
M	45.30%	43.83%	47.11%
SD	27.33%	19.90%	22.24%

Severity Ratings

During the User-CHAT, heuristic evaluation, and cognitive walkthrough evaluation sessions, evaluators were asked to assign a severity rating to each usability problem identified. Occasionally, however, the evaluators failed to assign a severity rating to each problem. Table 16 presents the proportions of usability problems that were rated as either serious, intermediate, or minor for each technique. The numbers above the percentages represent the total number of usability problems that were assigned a severity rating. The User-CHAT had slightly higher proportions of usability problems that were rated serious than the heuristic evaluation which in turn had more than the cognitive walkthrough. While overall the heuristic evaluation identified more usability issues rated as serious, this result is tied closely to the total number of usability problems found by the heuristic evaluation. Aside from the cognitive walkthrough, the User-CHAT and heuristic evaluation had higher proportions of minor problems than either intermediate or serious. Throughout, the cognitive walkthrough had a higher proportion of intermediate rated usability problems.

Table 16. Severity Ratings Proportions for Usability Problems Identified by each UEM.

Severity Rating	*User-CHAT*	*Heuristic Evaluation*	*Cognitive Walkthrough*
Total	110	229	115
Serious	23.6%	20.1%	9.6%
Intermediate	35.5%	31.8%	61.7%
Minor	40.9%	48.0%	28.7%

Though evaluators across all UEM sessions utilized the same severity rating categories, each evaluator may have had different conceptions as to what exactly constituted a serious, intermediate, or minor usability problem, thus potentially leading to considerable latitude in the application of the severity rating categories. Therefore, in order to reduce inter-rater variability, two GA instructor pilots were recruited to evaluate and assign a severity to all usability problems identified. Pilots were unaware of the experimental design and the UEM that found each usability problem. Tasks used during the evaluation and screen shots were available in order to provide context for the usability problems. Upon completion of assigning a severity ratings independently, the pilots discussed and resolved rating differences on a

consensus basis. Predominantly, the proportions from above were replicated (see Table 17). Severity rating proportions for problems identified by the User-CHAT, however, demonstrated the smallest proportion of problems rated as minor and the greatest proportion of intermediate-rated problems. The heuristic evaluation had similar proportions of intermediate and minor usability problems.

Table 17. Instructor Pilot Severity Rating Proportions for Usability Problems Identified by each UEM.

Severity Rating	*User-CHAT*	*Heuristic Evaluation*	*Cognitive Walkthrough*
Total	238	245	174
Serious	28.2%	26.2%	8.0%
Intermediate	59.2%	37.6%	55.7%
Minor	12.6%	36.3%	36.2%

Heuristics Classifications

The two human factors specialists that grouped together similar usability problems also classified each usability problem as a violation of one or more display design heuristics. For each usability problem, the human factors specialists indicated the heuristic(s) that were violated using the User-CHAT heuristics list. These heuristics represented general rules of thumb for display design. Frequencies and overall percentages for the heuristics used to classify the usability problems for each UEM are presented in Appendix L. Independent of the UEM, “Descriptive Labeling” and “Design Standards” were among the most frequently used heuristics to classify usability problems. That is, many of the issues with the avionics suite were associated with vague terminology and labeling or a failure to conform to established design guidelines for display elements and/or input devices.

Generating a Comparison List without User-testing

While user-testing has been deemed by many researchers (e.g., Landauer, 1995; Newman, 1998) to be the most appropriate technique to generate a comparison list of usability issues, it was mentioned earlier that some comparative usability studies (e.g., Wang & Caldwell, 2002; Tan & Bishu, 2002) created their own comparison list by simply combining the usability problems found by all of the UEMs in the study. With this procedure, time and resources necessary for conducting user-testing are eliminated because generating the standard list of usability problems does not require effort beyond

conducting the usability methods in the comparison study. However, as demonstrated above when discussing comparative metrics, generating a comparison list without user-testing has a serious drawback in that the validity metric becomes meaningless [viz., all usability problems detected by each UEM are valid (e.g., validity = 100%)].

What some previous comparative studies failed to do was to instantiate an inclusion criterion to determine whether or not a usability problem was real and should be integrated in a comparison list. For instance, the "realness" of a candidate usability problem could be assessed by the frequency that a usability problem was identified across evaluation sessions. If the candidate problem satisfies a predetermined inclusion criterion (e.g., must be detected in two or more evaluation sessions), then the problem could be included into a comparison list. When usability problems in the comparison list are filtered according to some inclusion criterion, the comparative performance metrics can be calculated.

Therefore additional exploratory analyses were conducted to ascertain UEM performance on thoroughness, validity, and effectiveness. The same usability data and problem groupings used to assess UEM performance relative to user-testing data was re-utilized for these analyses. For these analyses, however, the list of real usability problems upon which the methods were compared was not based on data gleaned from user-testing. Rather, the list of real usability problems was generated from usability problems identified from the three usability methods[13]. However, instead of simply including every problem identified by all the methods into one comparison list, various inclusion criteria were implemented to determine whether or not a usability problem was real and should be integrated.

Because the comparison lists for the exploratory analyses was established through the usability problems identified by the UEMs, in order to ensure that the User-CHAT, heuristic evaluation, and cognitive walkthrough are compared fairly, the six User-CHAT evaluation sessions were combined randomly post hoc to create three 2-person "teams". Thus, subsequent analyses examined UEM performance with an equal number of sessions and an equal number of evaluators in each session for all three methods.

[13] A qualification for a usability problem to be considered real is that the problem must impact the end-user when interfacing with the system (Hartson et al., 2001). Because all evaluation sessions in the current study had a pilot participant, it could be argued that the usability problems identified in each session could impact pilot performance and thus have some degree of "realness."

The first inclusion criteria instantiated was that a candidate usability problem must be detected in at least two evaluation sessions regardless of the UEM applied in order to be considered real (e.g., the usability problem must have a frequency of two or greater across evaluation sessions). Fu et al. (2002) reported that a problem must have a replicate (i.e., occurrence of two or greater) in order to be considered an actual problem. For the present study, a candidate usability problem must have a frequency of two or greater across evaluation sessions. This inclusion criterion provided the initial generation of the comparison list of usability problems. Usability problems that satisfied this criterion were included in the comparison list. Then a more conservative criterion was applied – a candidate usability problem must be detected in at least three evaluation sessions regardless of the usability method to be considered real. Usability problems that satisfied this criterion were included in a separate comparison list. Finally, the inclusion criterion was extended to include only usability problems that occurred in at least four evaluation sessions; issues that satisfied this requirement were included in another separate comparison list.

UEM performance data on thoroughness, validity, and effectiveness will be shown for the aforementioned inclusion criterion only. The following reasons explain why results presented below were limited to these aforementioned inclusion criteria:

- Examining UEM performance on at least three inclusion criteria can illustrate trend information; scores for each inclusion criterion can be plotted. These plots can show trends based on the number of evaluation sessions that a usability problem was detected.
- Usability problems identified in at least two evaluation sessions is the first inclusion criteria step above simply combining all usability problems detected by all UEMs.
- The heuristic evaluation and cognitive walkthrough had only three evaluation sessions. The same applied to the User-CHAT after the six sessions were combined post hoc to create three 2-person “teams.” Thus, a usability problem idiosyncratic to either method can have a detection frequency of at most three evaluation sessions.
- An inclusion criterion for four evaluation sessions ensured that the usability problems in the comparison list were detected by at least two different UEMs. Additionally, as shown in the following tables, beyond four evaluation sessions the number of real usability problems dropped off substantially from the total number of usability problems available and performance on some comparative measures reached an asymptote (e.g., validity).

Venn diagrams (see Figures 8, 9, and 10) show the partitioning of real usability problems for each inclusion criterion. Usability problem totals for each UEM are in parentheses. As shown in the figures, 92 usability problems were identified in at least two evaluation sessions, 51 in at least three evaluation

sessions, and 22 in at least four evaluation sessions. Thus, the number of real usability problems (i.e., the number of problems in the comparison list) was reduced considerably when the inclusion criteria shifted to become more conservative. When the inclusion criterion was based on two or three evaluation sessions, the User-CHAT identified a greater total of real usability problems than the heuristic evaluation, which identified more than the cognitive walkthrough. Additionally, more real usability problems were shared between the User-CHAT and heuristic evaluation than between any other two methods. Lastly, while some real usability problems were idiosyncratic to either the User-CHAT (19) or the heuristic evaluation (10) when the inclusion criterion was two evaluation sessions, roughly two-thirds of the total number of real usability problems occurred across two or more UEMs. Meanwhile, when the inclusion criterion was based on three evaluation sessions, nearly 90% (45/51) of the total real problems occurred across two or more UEMs.

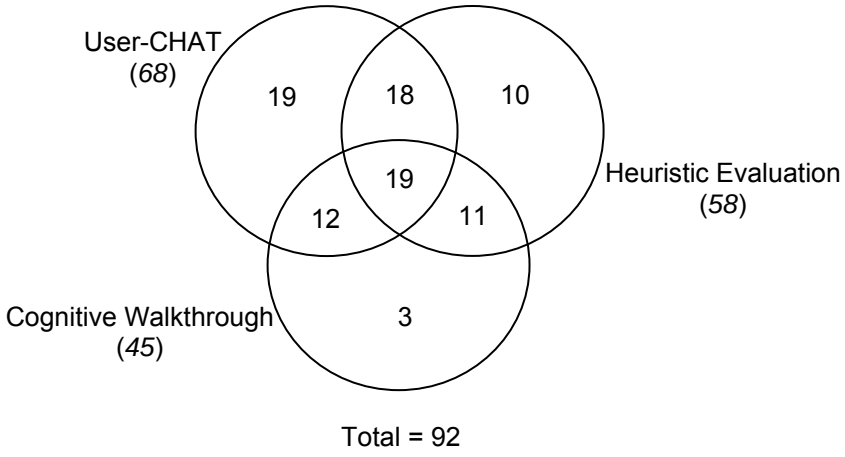

Figure 8. Venn diagram illustrating usability problem partitions for each UEM when the inclusion criterion was based on usability problems that were detected in at least two evaluation sessions.

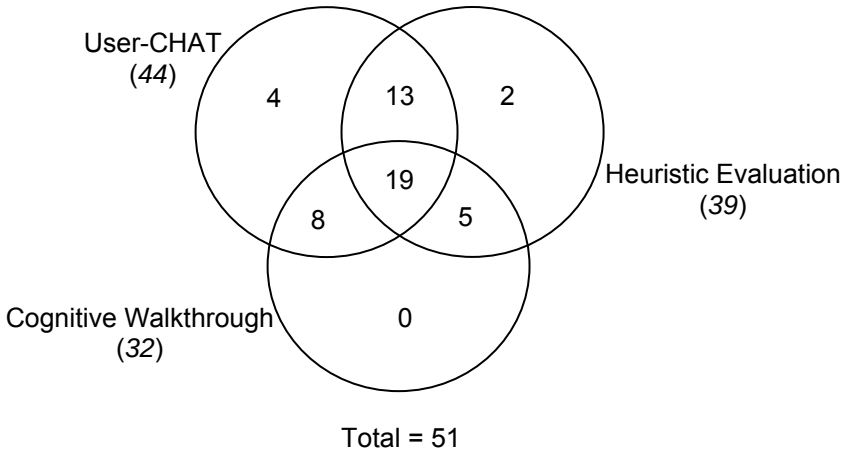

Figure 9. Venn diagram illustrating usability problem partitions for each UEM when the inclusion criterion was based on usability problems that were detected in at least three evaluation sessions.

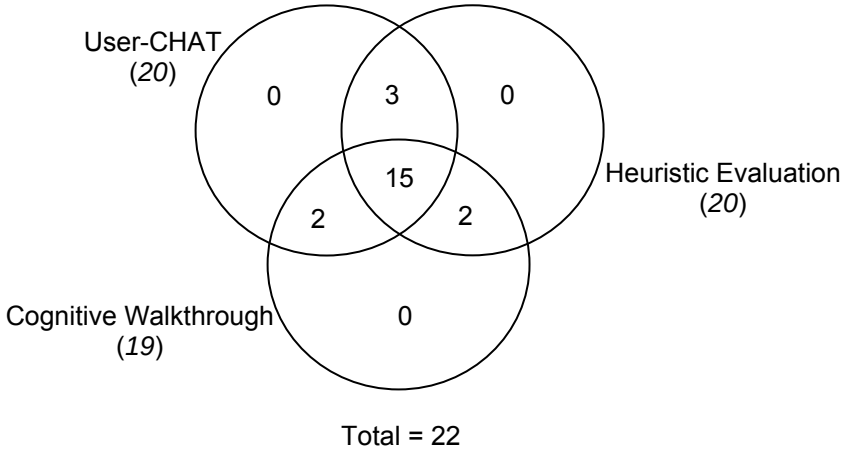

Figure 10. Venn diagram illustrating usability problem partitions for each UEM when the inclusion criterion was based on usability problems that were detected in at least four evaluation sessions.

As with comparing UEM performance to user-testing data, the following results present UEM performance per evaluation session and per usability method. Means, standard deviations, and 95% confidence intervals on thoroughness, validity, and effectiveness per evaluation session are presented in Table 18 and illustrated in Figure 11.

Regardless of the UEM, the plots for each inclusion criterion show that thoroughness tended to increase while validity tended to decrease; effectiveness, for the most part, remained relatively flat across inclusion criterion. The rational for the increased thoroughness scores is fairly straightforward. With a

more conservative inclusion criterion, the number of usability problems included in the comparison list became smaller and the usability problems included in these lists represented prominent usability issues with the avionics suite. That is, the usability problems in these lists were difficult to overlook, therefore the likelihood that those usability problems would be detected across evaluation session increased. For example, a poorly labeled dedicated function key, essential to complete a task, was detected across many evaluation sessions. As the inclusion criterion became more conservative, for all UEMs, more evaluation sessions detected the same usability problem, thus thoroughness scores improved.

Decreasing validity scores, however, can be explained through the validity calculation. For a given UEM, validity reflected the ratio of the number of real problems identified to the total number of usability problems identified. For example, assume that a UEM identified a total of 50 usability problems and for each type of inclusion criterion, 20, 15, and 10 real usability problems were identified. As the inclusion criterion became more conservative, the number of real usability problems identified decreased yet the total number of usability problems remained constant. Because the denominator of the validity ratio was constant, across all UEMs, validity indices decreased as the inclusion criterion shifted to become more conservative.

Finally, effectiveness was the product of thoroughness and validity. Flat effectiveness trends implied that a UEM's thoroughness and validity scores were increasing and decreasing respectively at similar rates. Assume that a UEM's thoroughness scores for each inclusion criterion were .50, .60, and .70 and validity scores were .60, .50, and .40. Effectiveness scores would remain relatively level across all inclusion criteria – thoroughness scores increased and validity scores decreased at similar rates. If thoroughness and validity scores changed at different rates, then the effectiveness trend would be anything but level. The heuristic evaluation and cognitive walkthrough had relatively flat effectiveness trends. The User-CHAT, however, had a decreasing effectiveness trend because validity scores decreased faster than thoroughness scores increased.

In relation to the comparative performance metrics, upon inspection of the means and confidence intervals across inclusion criteria, it appeared that the User-CHAT attained higher thoroughness scores than the heuristic evaluation and cognitive walkthrough. Wilks' λ, univariate ANOVAs and post hoc comparisons (see Appendix M for inferential statistics summary tables and post-hoc comparisons)

supported the notion that the User-CHAT was more thorough (i.e., detected more real usability problems) than the other two methods when the inclusion criteria were either two or three evaluation sessions. The heuristic evaluation and cognitive walkthrough achieved similar thoroughness scores throughout. MANOVA results, however, failed to show significance when the inclusion criteria was four evaluations, Willks' $\lambda = 0.15$, $F(6,8) = 2.16$, $p > .10$, indicating that all three methods shared similar scores for all performance metrics. For the most part, these thoroughness results contradict those above when UEM performance was measured against user-testing data.

In terms of validity, means and confidence intervals showed that the User-CHAT started initially with higher validity scores than the heuristic evaluation. However, as a more conservative inclusion criterion was implemented, both methods merged and attained similar validity scores (i.e., detected similar proportions of trivial usability problems). The User-CHAT and cognitive walkthrough seemed to obtain similar validity scores across inclusion criteria. Inferential statistics supported the conclusions conveyed through the confidence intervals. Recall that when compared to user-testing, the User-CHAT attained higher validity scores than the other two methods. This pattern was not replicated as the User-CHAT and cognitive walkthrough attained similar validity scores across inclusion criteria.

Finally, as evidenced by non-overlapping confidence intervals, the User-CHAT was more effective than both the heuristic evaluation and cognitive walkthrough; the latter two techniques were similar in effectiveness scores across inclusion criteria. These results, supported through inferential analyses in Appendix M, suggest that the User-CHAT is better than the other methods at focusing more on finding real and ignoring non-real usability problems. When compared to user-testing, the User-CHAT was also more effective than either of the other two methods.

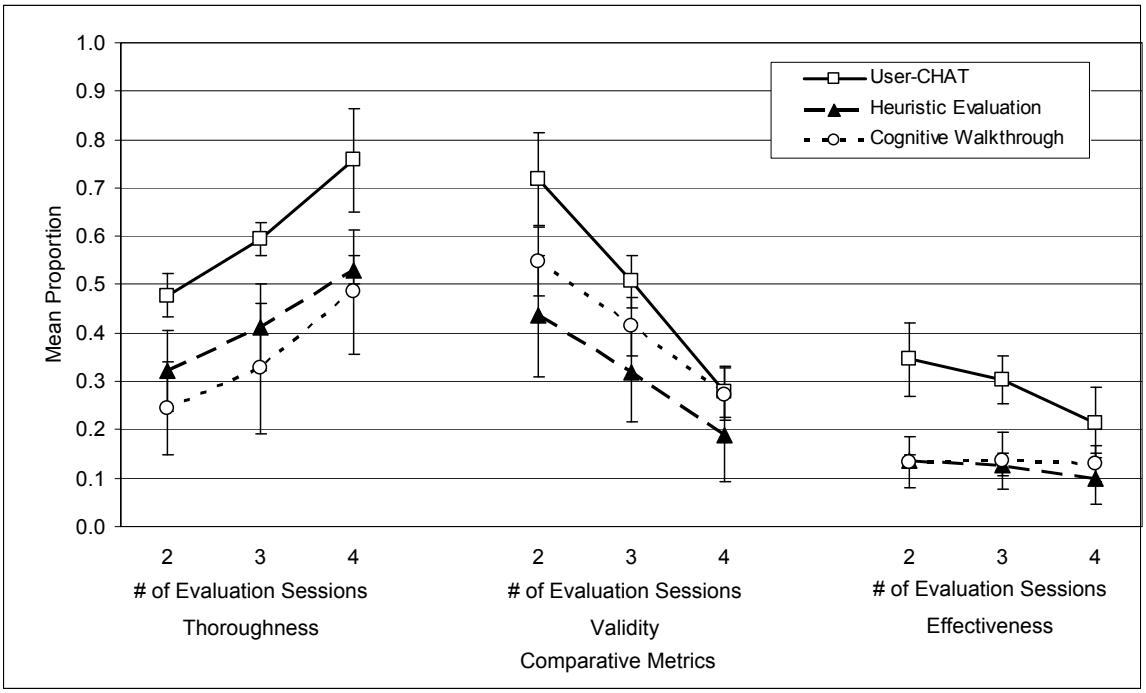

Figure 11. Thoroughness, validity, and effectiveness 95% confidence intervals for per evaluation session based on inclusion criteria of usability problems that were detected in at least 2, 3, or 4 evaluation sessions.

Table 18. Means, Standard Deviations, and 95% Confidence Intervals on Thoroughness, Validity, and Effectiveness Per Evaluation Session when Compared to Various Inclusion Criteria (Usability Problems Detected in at Least 2, 3, or 4 Evaluation Sessions).

# "Real" Problems	Inclusion Criteria	Metric	User-CHAT			Heuristic Evaluation			Cognitive Walkthrough		
			M	SD	95% CI	M	SD	95% CI	M	SD	95% CI
92	At Least 2 Evaluation Sessions	Thoroughness	0.48	0.04	0.52<μ<0.43	0.32	0.07	0.40<μ<0.24	0.24	0.08	0.34<μ<0.15
		Validity	0.72	0.08	0.81<μ<0.62	0.43	0.11	0.56<μ<0.31	0.55	0.06	0.62<μ<0.48
		Effectiveness	0.35	0.07	0.42<μ<0.27	0.13	0.01	0.15<μ<0.12	0.13	0.05	0.19<μ<0.08
51	At Least 3 Evaluation Sessions	Thoroughness	0.59	0.03	0.63<μ<0.56	0.41	0.08	0.50<μ<0.32	0.33	0.12	0.46<μ<0.19
		Validity	0.51	0.05	0.56<μ<0.45	0.32	0.09	0.42<μ<0.22	0.41	0.05	0.47<μ<0.35
		Effectiveness	0.30	0.04	0.35<μ<0.25	0.13	0.02	0.15<μ<0.10	0.14	0.05	0.20<μ<0.08
22	At Least 4 Evaluation sessions	Thoroughness	0.76	0.09	0.86<μ<0.65	0.53	0.03	0.56<μ<0.50	0.48	0.11	0.61<μ<0.36
		Validity	0.28	0.05	0.33<μ<0.23	0.19	0.08	0.28<μ<0.09	0.27	0.05	0.33<μ<0.22
		Effectiveness	0.21	0.06	0.29<μ<0.14	0.10	0.05	0.15<μ<0.05	0.13	0.03	0.17<μ<0.09

Table 19. Thoroughness, Validity, and Effectiveness Scores based on each UEM's Overall Performance when Compared to Various Inclusion Criteria (Usability Problems Detected in at Least 2, 3, or 4 Evaluation Sessions).

		User-CHAT (121)				Heuristic Evaluation (187)				Cognitive Walkthrough (100)			
# Real Problems	Inclusion Criteria	# of Real Problems Identified	T	V	E	# of Real Problems Identified	T	V	E	# of Real Problems Identified	T	V	E
92	At Least 2 Evaluation Sessions	68	0.74	0.56	0.41	58	0.63	0.31	0.20	45	0.49	0.45	0.22
51	At Least 3 Evaluation Sessions	44	0.86	0.36	0.31	39	0.76	0.21	0.16	32	0.61	0.32	0.20
22	At Least 4 Evaluation Sessions	20	0.91	0.17	0.15	20	0.91	0.11	0.10	19	0.86	0.19	0.16

T = thoroughness, V = validity, E = effectiveness

Comparative metrics reflecting each UEM's overall performance (e.g., per usability method) when compared to the various inclusion criteria are presented in Table 19 and illustrated in Figure 12. The numbers in the table adjacent to each UEM represent the total number of usability problems identified by that UEM while the whole numbers below signify the number of real usability problems detected given the inclusion criterion. Thoroughness scores increased and validity scores decreased as the inclusion criterion shifted to become more conservative. With the exception of the cognitive walkthrough, effectiveness scores seemed to decrease with a more conservative inclusion criteria.

In general, thoroughness indices reflected the pattern of results demonstrated from the per evaluation session results above. The User-CHAT appeared to be more thorough than the heuristic evaluation which appeared to be more thorough than cognitive walkthrough (except when the inclusion criterion reached four evaluation sessions). The User-CHAT identified five to ten more problems than the heuristic evaluation and at least 13 more problems than the cognitive walkthrough for the first two inclusion criteria. When the inclusion criterion reached four evaluation sessions, all UEMs identified similar numbers of real usability problems.

Validity scores for each UEM's overall performance illustrated a slightly different pattern. Initially, the UEMs differed in terms of validity scores, where the User-CHAT had higher validity scores than the cognitive walkthrough which had higher validity scores than the heuristic evaluation. However, as the inclusion criterion became more conservative (e.g., three and four evaluation sessions), the User-CHAT and cognitive walkthrough each had similar proportions of trivial usability problems. A pattern replicated from above was that, across inclusion criteria, the heuristic evaluation identified many more trivial usability issues than the User-CHAT.

With regard to the effectiveness results for per evaluation session, results for UEM performance as a whole demonstrated similar patterns. The heuristic evaluation and cognitive walkthrough achieved approximately similar effectiveness for two and three evaluation sessions. The User-CHAT had higher effectiveness scores than either method for the same inclusion criteria, suggesting that it was better at identifying real and attenuating trivial usability problems. When the inclusion criterion was more conservative, effectiveness scores for the User-CHAT and cognitive walkthrough converged, but both were higher than the heuristic evaluation.

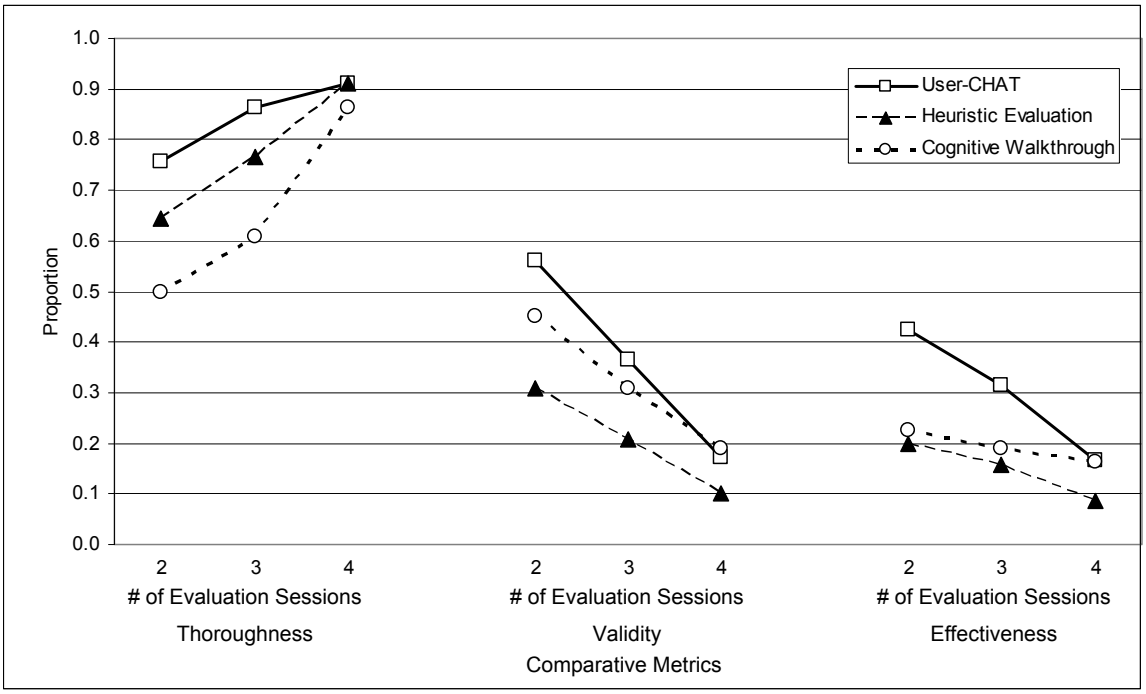

Figure 12. Overall UEM performance for thoroughness, validity, and effectiveness based on inclusion criteria of 2, 3, or 4 evaluation sessions.

Chapter 4 -- Discussion

The present comparative UEM study sought to accomplish two primary objectives: (1) to validate a new hybrid usability technique as an effective technique for detecting important usability problems and ignoring trivial usability problems all while satisfying constraints associated with the FAA certification environment and restrictions common to many other usability laboratories and (2) to directly compare different UEMs using metrics that appropriately assess the overall effectiveness of each UEM.

The User-CHAT is a usability assessment technique that was tailored specifically for use in the FAA certification environment, but can be extended easily to other industries where usability assessments are fundamental to a business's or product's success. The User-CHAT adopts suitable elements from user-testing, the heuristic evaluation, and the cognitive walkthrough. It is intended for easy use like the heuristic evaluation, provides performance-based usability data like user-testing, and provides structure and supplies diagnostic information like the cognitive walkthrough. Additionally, the User-CHAT supports evaluations by individuals with limited human factors backgrounds. However, in order to fully realize the potential of the User-CHAT as an effective usability technique, a comparative usability study was conducted where the User-CHAT's performance was compared directly to the performances of user-testing, the heuristic evaluation, and the cognitive walkthrough.

In order to assess UEM performance, a state-of-the-art suite of avionics displays were evaluated by all usability techniques. The same series of tasks were used for each evaluation session to ensure that all the methods were assessing the same displays. Also the number of evaluators who actively contributed to the usability data was kept constant. After problems found through user-testing were screened and the usability problems from the remaining UEMs were compiled, two human factors professionals grouped together similar usability problems and assigned one or more heuristics to each problem.

Data from the User-CHAT, heuristic evaluation, and cognitive walkthrough were compared on a variety of measures. Some measures included total amount of usability problems detected, average time to identify a problem during the session, inter-rater reliability across evaluation sessions with a given UEM, and the number and proportion of usability problems that were rated as serious. The central purpose of the present study, however, was to compare UEM performance to a standard list of "real"

usability problems. Two competing techniques were used to generate the comparison list: 1) usability problems found through user-testing data, and 2) usability problems identified by the User-CHAT, heuristic evaluation, and cognitive walkthrough that satisfied various inclusion criteria. After the comparison lists were created, each usability method was compared in terms of thoroughness (i.e., ability to detect real usability problems), validly (i.e., ability to ignore non-real usability problems), and effectiveness (i.e., ability to detect a lot of real usability problems while ignoring non-real usability problems).

Hypotheses Tested

Thoroughness

"Real" Thoroughness was defined as the number of real usability problems found by a particular UEM relative to the total number of real usability problems that exist. Because the User-CHAT is a variant of user-testing and the comparison list of real usability problems were identified through user-testing, it was hypothesized that the User-CHAT would identify more real usability problems (i.e., attain higher "real" thoroughness scores) than the heuristic evaluation which will identify more real usability problems than the cognitive walkthrough. Compared to user-testing data, each User-CHAT, heuristic evaluation, and cognitive walkthrough evaluation session achieved similar thoroughness scores – no significant differences existed in the number of real usability problems identified. However, when the overall performance of each UEM was considered (i.e., a holistic perspective), the practical significance of the User-CHAT was revealed as it detected a greater number of real usability problems than the heuristic evaluation which detected more real usability problems than the cognitive walkthrough. Generally this pattern was repeated when the list of real usability problems was generated through various implementations of inclusion criteria – the User-CHAT identified more real usability problems than either of the other two methods, which failed to differ statistically. Additionally, for usability problems detected in four or more evaluation sessions, all three methods achieved similar thoroughness scores. The results across all analyses are somewhat contradictory. Therefore a safe conclusion is that, for the most part, the User-CHAT, heuristic evaluation, and cognitive walkthrough performed similarly with respect to the number of real usability problems identified in the avionics suite.

"Aggregate" Thoroughness reflected the number of usability problems identified by a UEM relative to the total number of usability problems identified by all UEMs. Because of the breadth of its underlying methodology (i.e., focus on all individual display elements), it was hypothesized that the heuristic evaluation would identify more usability problems (i.e., higher "aggregate" thoroughness) than the cognitive walkthrough and the User-CHAT. The Venn diagram and Chi-square analyses demonstrated that the heuristic evaluation detected significantly more usability problems than the User-CHAT, which had a higher aggregate thoroughness score than the cognitive walkthrough. Thus the breadth of the heuristic evaluation was realized as it identified more usability issues than the other two techniques.

Validity

Validity was defined as the number of real usability problems found by a UEM relative to the total number of usability problems identified by that UEM. It was hypothesized that the User-CHAT would contain a greater proportion of real usability problems in its list of usability problems (i.e., higher validity) than the cognitive walkthrough, which would have higher validity scores than the heuristic evaluation. Relative to user-testing data, the User-CHAT achieved a higher validity scores than the cognitive walkthrough and heuristic evaluation (which did not differ), suggesting that the User-CHAT detected fewer trivial issues. Similar performance patterns were observed when the overall contribution of each usability method was considered. Relative to various inclusion criteria, in general, the User-CHAT and cognitive walkthrough attained similar validity scores which were higher than the heuristic evaluation. However, as a more conservative criterion was implemented (e.g., four evaluation sessions), all three techniques converged and had similar validity scores. When the overall contribution of the usability method was considered, similar patterns for validity were observed. Validity reflects how much extra effort is being put forth on usability issues that are not important (Andre, 2000). Essentially the data suggest that the User-CHAT is better at helping evaluators focus on important usability problems and ignore less important issues than the heuristic evaluation and in some cases, the cognitive walkthrough.

Effectiveness

Effectiveness was defined as the product of "real" thoroughness and validity. It was hypothesized that the User-CHAT's list of usability problems would contain many real usability problems (i.e., high

thoroughness) in the total number of usability problems it identified (i.e., high validity). Thus, the User-CHAT would be more effective than the cognitive walkthrough, which in turn will be more effective than the heuristic evaluation. With one exception, the User-CHAT attained higher effectiveness scores on average than the cognitive walkthrough and the heuristic evaluation, which attained similar effectiveness scores; when the inclusion criterion reached four evaluation sessions, all UEMs attained similar effectiveness scores. Overall the results indicate that the User-CHAT detects a lot of usability problems that actually impact users and reduces detection of trivial usability problems, more so than the heuristic evaluation and cognitive walkthrough.

Reliability

Reliability represented the ability to find similar numbers of usability problems under similar conditions. The usability problems found through the User-CHAT were based primarily upon user-performance (i.e., objective) data, whereas the usability problems found through the heuristic evaluation and the cognitive walkthrough were strongly tied to the evaluator's background and motivation, which could lead to greater variability across evaluation sessions. Greater variability leads to lower reliability scores. Thus, it was hypothesized that the User-CHAT would find more similar numbers of usability problems across evaluation sessions (i.e., greater inter-rater reliable) than the heuristic evaluation and the cognitive walkthrough. The User-CHAT had a higher inter-rater reliability when the reliability index was based on the total number of usability problems found. When the reliability index was based on the number of real problems found, the heuristic evaluation achieved the highest score. Neither of these reliability indices was informative as to whether the same usability problem was identified across evaluation sessions for a given UEM. For all UEMs, percent raw agreements on the types of usability problems identified were roughly equivalent.

Problem Severity

The usability problems found through the User-CHAT were problems that actually affected user performance because representative end-users were completing the tasks. Thus it was hypothesized that the User-CHAT would identify a greater proportion of usability problems that received a serious rating than the cognitive walkthrough and the heuristic evaluation. The heuristic evaluation identified more

usability problems that were rated as serious. The User-CHAT, however, had slightly higher proportion of usability problems that were rated serious than the heuristic evaluation, which had a higher proportion of usability problems rated as serious than the cognitive walkthrough. This finding contradicts previous research (e.g., Dutt et al., 1994; Jeffries et al., 1991) that suggests the cognitive walkthrough would find more usability problems that were rated more serious than the heuristic evaluation. With regard to the User-CHAT, the fact that it had a higher proportion of serious-rated usability problems conforms to the research that user-testing finds more usability problems rated as serious than minor (e.g., Wharton et al., 1994; Jeffries et al., 1991).

Summary of UEM Performance

Heuristic Evaluation

As anticipated, due to its broad approach to an interface evaluation (Wharton et al., 1994), the heuristic evaluation detected the most usability problems. These findings support those reported in previous research (e.g., Doubleday et al., 1997; John & Marks, 1997; Tan & Bishu, 2002; Wang & Caldwell, 2002). Coupled with the notion that evaluators spent a relatively shorter amount of time on average identifying each usability problem provides empirical evidence to Nielsen and Molich's (1990) claim that, as a "discount" usability method, the heuristic evaluation can identify a lot of usability problems fairly quickly (see also Virzi et al., 1993). While the heuristic evaluation had the greatest number of usability problems that were rated serious, this outcome was tied closely to the vast number of problems identified. Proportionally, however, the heuristic evaluation had many more usability problems rated as minor compared to serious-rated usability problems; Wharton et al. (1994) reported similar severity proportions. That is, much of the usability problems identified through the heuristic evaluation could be considered low-priority (Jeffries et al., 1994) or cosmetic in nature (Nielsen, 1992; Hornbæk & Frøkjær, 2004). Furthermore, evaluators using the heuristic evaluation showed poor inter-rater consistency and agreement in terms of the total numbers of usability problems identified and the documentation of the same usability problem. However, it achieved a similar inter-rater reliability index reported by Sears (1997) for finding similar numbers of real usability problems.

Compared to user-testing data, the heuristic evaluation identified similar numbers of real usability problems as the other two techniques; its thoroughness score was similar to what Andre et al. (2003)

reported, but considerably smaller than scores reported by Sears (1997). When a holistic perspective was adopted and when usability data from the UEM's themselves generated the comparison list, in general, the heuristic evaluation detected similar numbers of real usability problems as the cognitive walkthrough but less than the User-CHAT. Andre et al. (2003) demonstrated also that the heuristic evaluation and cognitive walkthrough performed similarly in terms of thoroughness while other studies (e.g., Sears, 1997) showed that the heuristic evaluation was more thorough than the cognitive walkthrough. Although the heuristic evaluation showed considerable efficacy for identifying large numbers of usability problems, its advantage was attenuated when compared to data captured from user-testing. The fact that the heuristic method showed decreased thoroughness highlights its propensity to find a large number of false positives or non-real usability problems (Cockton & Woolrych, 2001; Bailey, Allan, & Raiello, 1992).

Because the heuristic evaluation detected a fair amount of specific one-time usability problems, its validity indices suffered. In fact, the pattern of validity scores contradicted previous research. For instance Sears (1997) and Andre et al. (2003) reported higher validity scores for the heuristic evaluation than those reported in the current study and their validity scores for the heuristic evaluation were significantly less than the cognitive walkthrough. In the present study, however, it was demonstrated that, for the most part, the heuristic evaluation and cognitive walkthrough attained similar validity scores. The reason for the low validity scores associated with the heuristic evaluation may be explained by the free-form nature of the method. According to Andre (2000), the heuristic evaluation tends to detect large numbers of false alarms (i.e., trivial usability problems), thereby reducing the usability technique's validity (Sears, 1997).

As a result of decreased validity scores, the effectiveness of the heuristic evaluation suffered as well. Relative to the User-CHAT, it did not optimize the detection of real usability problems and limit the number of non-real usability problems. However the heuristic evaluation was as effective as the cognitive walkthrough. This latter finding contradicts those reported by Andre et al. (2003) who found the heuristic evaluation to be less effective than the cognitive walkthrough.

Despite the heuristic evaluation's effectiveness shortcomings, usability practitioners in industry may still find it attractive because it tends to find many usability problems in a relatively short amount of time. Identifying lots of issues within a system is important when the overall design goal is to *optimize* the

system. For the FAA certification environment, however, interface optimization is not important. Rather, the FAA's intent is to assess whether or not a system is safe for use. That is, the FAA is concerned more with *minima* requirements. Thus with its low effectiveness score, the heuristic evaluation does not appear to be a suitable candidate for implementation into the certification process.

Cognitive Walkthrough

True to its reputation as a time consuming evaluation technique (Lewis et al., 1990; Rowley & Rhoades, 1992; Huart et al., 2004), the cognitive walkthrough required the greatest amount of time to identify each usability problem. Therefore the use of the cognitive walkthrough can be costly (Mack & Nielsen, 1994). The cognitive walkthrough is a task-based usability technique and task-based approaches generally take more time than free-exploration (e.g., heuristic evaluation), but in the process, it exposes the evaluator to issues that are most likely to impact real users (Andre 2000). Additionally, due to its narrow focus on the specific action sequences required to reach a goal state (Jeffries et al., 1994), it was not surprising that the cognitive walkthrough identified the fewest number of usability problems (see also Andre et al., 2003). Many of the usability problems identified through the cognitive walkthrough were concerned with input device labeling or terminology issues. The cognitive walkthrough had a greater proportion of minor-rated usability problems than serious-rated usability problems. Previous research, however, demonstrated that the cognitive walkthrough detects greater (Dutt et al., 1994) or equal proportions of serious and minor-rated problems (Wharton et al, 1994). Inter-rater reliability scores based on the total number of usability problems identified was greater than the heuristic evaluation, but less than the User-CHAT. Thus evaluators applying the cognitive walkthrough were more likely to report similar numbers of usability problem than the heuristic evaluation, but not the User-CHAT. However when based on the number of real problems identified, similar to Sears (1997), the cognitive walkthrough had lower inter-rater reliability than the heuristic evaluation. Despite the structure inherent in the cognitive walkthrough methodology – evaluators are focused on specific characteristics (e.g., input actions) of the system – the percent raw agreement for documenting the same usability problem across cognitive walkthrough evaluation sessions was not much better than the other two methods.

Thoroughness for the cognitive walkthrough was similar to the other methods when compared to user-testing; it detected similar numbers of real usability problems as the heuristic evaluation and User-

CHAT. Andre et al. (2003) failed also to reveal a thoroughness difference between the heuristic evaluation and cognitive walkthrough. The cognitive walkthrough's thoroughness score reported in the current study was three times smaller than the score reported by Sears (1997) but similar to Andre et al. (2003). For the most part, similar thoroughness patterns were observed when the overall contribution of the method was considered. However, when the UEMs generated the comparison list, the cognitive walkthrough failed to find as many real usability problems as the User-CHAT (except when the comparison list of usability problems was generated by those that were documented in four or more evaluation sessions). While the cognitive walkthrough detected similar numbers of real usability problems as the heuristic evaluation (i.e., similar numbers of "hits"), it did not detect as many as the User-CHAT when the inclusion criteria was based on usability problems detected in at least two or three evaluation sessions.

Andre (2000) postulated that the time needed to conduct a comprehensive cognitive walkthrough was time well spent as evaluators are focused on issues that impact end-users. However, when compared to user-testing data, the relatively small validity scores indicated that many of the usability issues documented through the cognitive walkthrough did not impact end-users. In fact, the cognitive walkthrough's validity index was noticeably smaller than those reported by Sears (1997) and Andre et al. (2003). Even though validity improved when compared to the comparison list yielded from the three UEMs, the cognitive walkthrough and heuristic evaluation achieved similar validity scores, implying that both methods spent similar amounts of effort focusing on usability issues that were not important. Conversely, Sears (1997) and Andre et al. (2003) found the cognitive walkthrough to have significantly higher validity scores than the heuristic evaluation.

Throughout all analyses, the cognitive walkthrough and the heuristic evaluation had similar effectiveness scores, which were less than the User-CHAT. Effectiveness scores were nearly three times smaller than those reported by Andre et al. (2003). Thus, like the heuristic evaluation, the cognitive walkthrough did not optimize detection of real usability problems and reduce detection of trivial usability problems.

It is interesting to point out the cognitive walkthrough's performance on the tasks that were predominantly input action oriented (tasks 10-18). On these tasks, the cognitive walkthrough (48)

identified more usability problems than the User-CHAT (24), but fewer than the heuristic evaluation (54). Inter-rater reliability based on the total number of problems identified and percent raw agreement between cognitive walkthrough evaluators were the highest, implying that these evaluators were more likely to document similar numbers and the same usability problems compared to the User-CHAT and heuristic evaluation. Relative to user-testing data, in terms of thoroughness on tasks 10-18, the cognitive walkthrough, on average, identified more real usability problems than either the User-CHAT or heuristic evaluation, but attained a similar effectiveness score as the User-CHAT. Finally, when compared to a comparison list generated by the methods themselves, the cognitive walkthrough and User-CHAT attained similar thoroughness and effectiveness scores for usability problems that were found in either two or three evaluation sessions (all three UEMs performed similarly when the comparison list was compromised of problems that occurred in four or more evaluation sessions). The rationale for this improved performance on these tasks only was that these tasks required evaluators to focus more on evaluating the input actions necessary to reach the desired information and less on evaluating the display elements. These tasks played into the strength of the cognitive walkthrough – focus on the input actions necessary to reach the desired information (Cuomo & Bowen, 1994).

In sum, the cognitive walkthrough is a time-consuming evaluation process that does not typically detect many usability problems. While the cognitive walkthrough tends to limit the number of non-real usability problems in its list of issues, its relatively low effectiveness scores and the amount of time spent to detect each usability problem suggest that the cognitive walkthrough also may not be appropriate for the FAA to incorporate into its certification process, unless the candidate system contains many input actions and few display elements to evaluate. The same rationale applies to industry usability practitioners. Cognitive walkthroughs appear to discourage exploration, thereby limiting the evaluator's ability to find usability problems not related directly to the tasks being performed (Andre, 2000); thus optimizing the interface becomes more challenging as the cognitive walkthrough tends to overlook global consistency issues within the system (Jeffries et al., 1994).

User-CHAT

Nielsen and Molich (1990) defined "discount" usability methods as techniques that are inexpensive, easy to use, and identify usability problems quickly (i.e., fast). Research by Uhlarik et al.

(2003, 2004) demonstrated that the User-CHAT can support comprehensive evaluations in under four hours, can be utilized successfully by evaluators with limited human factors / usability background, and does not require much training in order to assess complex systems effectively. While the User-CHAT does not share the breadth that the heuristic evaluation has, it identified a respectable number of usability problems and spent a similar amount of time as the heuristic evaluation to identify each usability problem (i.e., fast). Most of Nielsen and Molich's criteria have been met, thus it can be argued that the User-CHAT may also be classified as a "discount" usability method. The main cost associated with the User-CHAT is recruiting end-users and staff resources required to conduct the evaluation sessions, costs associated typically with user-testing (Scerbo, 1995), but the costs can be attenuated if usability data is captured in real time (Brooks, 1994). Usability practitioners can make the User-CHAT as cheap or as expensive as their budget allows.

The User-CHAT had a higher proportion of serious-rated usability problems relative to the other two methods. Moreover, the smallest proportion of the usability problems found through the User-CHAT were rated as minor (e.g., simple design recommendations), which conforms to the findings reported by Wharton et al. (1994) and Jeffries et al. (1994) that user-testing protocols identify more usability problems rated as serious and avoid low-priority problems. When based on the total number of usability problems identified, the User-CHAT had the highest inter-rater reliability score; this result was not replicated when based only on the number of real usability problems identified. When inter-rater consistency was calculated for identifying the same usability problem across evaluation sessions, even though the User-CHAT was based on six evaluation sessions, it had similar percent raw agreement between evaluators as the heuristic evaluation and cognitive walkthrough.

It was anticipated that the User-CHAT would identify more real usability problems because the technique's basic foundation reflects a methodology associated with user-testing. As mentioned earlier, when compared to user-testing, User-CHAT evaluation sessions identified similar numbers of real usability problems as heuristic evaluation and cognitive walkthrough evaluation sessions. However when a holistic view was adopted, the User-CHAT identified more real usability problems found through user-testing than the other methods; each evaluation session added to the total number of real problems

identified by the User-CHAT. Similar thoroughness performance patterns were observed when the comparison list was generated by the three UEMs.

Usability methods with a lot of real usability problems in their problem list (i.e., high validity) attenuate speculation as to which problems actually impact end-uses (Jeffries & Desurvire, 1992). Over 50% of the usability problems in the User-CHAT's problem list were ones that were found through user-testing (i.e., usability problems that impact user performance) and its validity score was higher than both the heuristic evaluation and cognitive walkthrough. When usability practitioners sift through the list of usability problems yielded from a User-CHAT session, more than half of the problems identified will be important and impact end-users. The User-CHAT's pattern of validity scores carried over when the overall performance of each method was considered. However, when compared to the comparison list created from various inclusion criteria, the User-CHAT attained similar validity scores as the cognitive walkthrough, but better than the heuristic evaluation. All three UEMs had similar validity scores when the inclusion criteria reached four evaluation sessions.

While conclusions regarding thoroughness and validity may be inconsistent across analyses, conclusions regarding effectiveness are maintained (with one exception). Whether performance was compared to user-testing data or comparison lists generated through various inclusion criteria, the User-CHAT achieved higher effectiveness scores than both the heuristic evaluation and cognitive walkthrough. Even on tasks that were primarily input action oriented, the User-CHAT was as effective as the cognitive walkthrough. Effectiveness indices imply that the User-CHAT is a more efficient usability tool than the heuristic evaluation and cognitive walkthrough because the time spent evaluating a system is focused on identifying usability problems that actually impact users while identification of trivial usability problems is attenuated. That is, in the present study, the User-CHAT best approximated the definition of an "ideal" UEM in which many real usability problems were detected (i.e., thoroughness) in the total list of usability problems identified (i.e., validity).

When all performance metrics are considered, the User-CHAT appears to be most suitable for implementation into the FAA certification process and for industry usability labs. As a discount usability method, the User-CHAT identified lots of usability problems quickly and can be used by evaluators with varying levels of human factors / usability knowledge. The User-CHAT's effectiveness scores indicate that

it is better at helping evaluators focus on “real” usability issues and ignore trivial problems. The User-CHAT’s overall merit suggests that it can be used efficiently by either FAA certification personnel to evaluate whether a candidate system satisfies minima requirements or by usability practitioners to ensure that a product will be usable by the target end users. Additionally, the User-CHAT was equally effective or more effective on tasks that were focused on accessing information or tasks that were focused more on evaluating and interpreting the displayed information. That is, because the User-CHAT combines properties of the heuristic evaluation and the cognitive walkthrough, it was able to perform well on tasks that demanded attention on evaluating display elements and on tasks that demanded attention on evaluating the input actions necessary to reach the desired information. Finally, the User-CHAT provides objective performance data to substantiate the usability problems identified. Therefore, the User-CHAT satisfies constraints associated with FAA certification environment and usability laboratories common to industry and could be implemented easily into these usability assessment arenas.

The following table summaries the relative performances of the User-CHAT, the heuristic evaluation, and the cognitive walkthrough on the comparative metrics.

Table 20. Comparative Metrics Summary for the Heuristic Evaluation, the Cognitive Walkthrough, and the User-CHAT.

Metric	*Heuristic Evaluation*	*Cognitive Walkthrough*	*User-CHAT*
Time per Usability Problem	▪ Quickest	▪ Slowest	▪ Quickest
"Aggregate" Thoroughness	▪ Identified the greatest number of problems.	▪ Identified the fewest number of problems.	▪ "In the middle"
"Real" Thoroughness	▪ User-testing Data ▪ Similar numbers of "real" problems identified by each evaluation session. ▪ Holistic perspective: User-CHAT > Heuristic Evaluation > Cognitive Walkthrough. ▪ Inclusion Criteria ▪ 2 and 3 evaluation sessions: User-CHAT > Heuristic Evaluation > Cognitive Walkthrough.		
Validity	▪ User-testing Data ▪ User-CHAT > Heuristic Evaluation and Cognitive Walkthrough. ▪ Inclusion Criteria ▪ 2 and 3 evaluation sessions: User-CHAT and Cognitive Walkthrough > Heuristic Evaluation.		
Effectiveness	▪ Similar to Cognitive Walkthrough	▪ Similar to Heuristic Evaluation	▪ More Effective
Reliability	▪ Total Number of Problems Identified = User-CHAT ▪ Total Number of "Real" Problems Identified = Heuristic Evaluation ▪ Similar agreement (e.g., same problem identified) across UEMs.		
Problem Severity (%)	▪ Minor > Serious	▪ Minor > Serious	▪ Serious > Minor

Future Research Directions

Future research regarding the User-CHAT could propagate into many different areas. For instance, much of the research concerning current User-CHAT development has been with avionics systems. Now that the User-CHAT has been compared to other well-established UEMs, it would be interesting to see how well the User-CHAT can evaluate non-aviation systems. That is, an interesting line of research would be to investigate the performance of the User-CHAT when it is applied to websites, mobile devices, software programs, construction equipment, etc. This would answer the question for usability practitioners regarding the ubiquity of the User-CHAT.

Another interesting research extension would be to utilize a different supervisor and user for each User-CHAT evaluation session. The way the User-CHAT is intended, the supervisor plays an active role in identifying usability problems. In the present study, however, the supervisor's role was limited due to lack of available human factors specialists. That is, while the User-CHAT was designed for the supervisor to actively assist in evaluating the avionics system, due to experimental design considerations, the supervisor was not allowed to actively participate in detecting usability problems. Thus future research should examine how well the User-CHAT performs with a different user and supervisor for each evaluation session, so that each supervisor may be a full participant in the evaluation process.

Also, as mentioned earlier, the list of general display design heuristics that accompany the User-CHAT originated from a variety of human factors sources, e.g., Nielsen's (1994b) display design heuristics, Wickens' et al., (1999) 13 principles of display design, etc. While this list of heuristics describes many characteristics of usable interfaces, the list could be improved. Relative to aviation, the list could contain additional heuristics that speak to usable avionics systems. For instance, aviation-specific heuristics that should be included are those that are more associated with ergonomic design because most operations with avionics systems occur during flight and sometimes under turbulent conditions. There may be other heuristics in the human factors literature that may be valuable additions to the current User-CHAT list of heuristics.

Lastly, the utility and effectiveness of the User-CHAT as a tool to aid evaluators in identifying and diagnosing usability problems has been examined using evaluators with various human factors and aviation backgrounds. For instance, past User-CHAT evaluations have used evaluators with human factors experience but no pilot experience; other User-CHAT evaluations have used pilots without any human factors backgrounds. However, the User-CHAT has yet to be evaluated using personnel with ideal backgrounds for assessing candidate avionics systems. Therefore, the next logical extension of the User-CHAT validation process would be to investigate its efficacy in an FAA certification environment with evaluators who have expertise in both human factors and aviation. The implications of this line of research are two-fold. First, insight can be gained into how well double experts perform when using the User-CHAT to evaluate a system in an FAA certification environment. That is, it can be determined how well the User-CHAT supports FAA certification personnel in effectively and efficiently conducting an

evaluation on a candidate system in an FAA certification environment. Any shortcomings found (based on the ability of the evaluators to accurately identify usability problems or from verbal feedback offered after the evaluation) will be incorporated in order to improve the User-CHAT. Second, previous User-CHAT evaluations have used evaluators with various combinations of human factors and aviation backgrounds. To date, however, the ability of the User-CHAT to support evaluators with expertise in both human factors and aviation in identifying usability problems has not been assessed. From this research extension, it can be determined how well domain expertise in both human factors and aviation influences the evaluator's ability to accurately document usability problems using the User-CHAT. Feedback from these evaluators, both quantitative and qualitative, can be integrated to improve the User-CHAT as a marketable usability assessment technique.

Limitations

The primary limitation of the present study is small sample size. While the number of evaluators in the current study satisfied the bounds recommended by several researchers (e.g., Nielsen & Molich, 1990; Virzi, 1990) and was in the number range regularly reported in the HCI literature (Wharton et al., 1992; Dutt et al., 1994), instead of six evaluators per usability method who actively contributed to the usability data, it would have been more ideal to have approximately ten evaluators per usability technique. However, as is the case with most usability evaluation sessions, the present study was conceptualized to meet specific time, personnel, and resource constraints common to many usability evaluations (Dutt et al., 1994). Therefore, while the results of the present study may not have sufficient sample size to make generalizations, examination of the data for practical significance yields important information about the performance of each UEM.

Another possible limitation is the notion that the User-CHAT is a task-based usability technique. While focusing on common user-tasks attenuates the incidence of identifying trivial or non-real usability problems, this approach limits the evaluator's ability to explore all aspects of the candidate system (Sears, 1997). Thus the extent of the User-CHAT is dependent upon the comprehensiveness of the tasks and the functionality the tasks exercised (Doubleday et al., 1997). However, according to Sears (1997), an appropriate combination of task-based and free exploration can enable evaluators to detect severe problems while reducing false positive usability problems. Therefore, a modification to the User-CHAT

that should be researched would be to allow evaluators to extend beyond the tasks and assess unexplored areas of the system, thereby increasing the comprehensiveness of the User-CHAT.

Finally, another limitation of the current study was the relative heterogeneous sample of pilot and human factors specialists across evaluation sessions. Other comparative studies (e.g., Dutt et al., 1994; Jeffries et al., 1991) suffered from heterogeneous samples across UEMs as well. For the present study it would have been ideal to have pilots and human factors with similar levels of experience. Experience levels varied across UEMs, especially for the cognitive walkthrough. Mack and Nielsen (1994) reported that experienced evaluators tend to identify more usability problems. However, the cognitive walkthrough session that had a GA pilot with the least amount of total flight hours and a human factors specialist with the least amount of display design, testing, and evaluating experience identified a greater number of usability problems than the other cognitive walkthrough sessions. Additionally, the GA pilots in the heuristic evaluation sessions had more flight experience than the other pilots. This added experience may have contributed to the heuristic evaluation's overall larger number of usability problems detected relative to the other usability methods.

Chapter 5 -- Conclusion

This study sought to evaluate a new hybrid usability method in order to determine its performance relative to other well-established usability techniques and to provide additional UEM performance data for contemporary comparative metrics. The User-CHAT combines aspects of user-testing, the heuristic evaluation, and the cognitive walkthrough. User-testing focuses on obtaining performance data, heuristic evaluations focus on ease of use, and cognitive walkthroughs on structure and problem diagnosis. The present study examined differences between UEMs (e.g., the User-CHAT, the heuristic evaluation, and the cognitive walkthrough) on various performance measures. User-testing as well as various combinations of usability problems identified from each UEM generated a baseline set of "real" usability problems. Results on the performance measures showed that the User-CHAT achieved higher effectiveness scores than the heuristic evaluation and cognitive walkthrough, suggesting that the User-CHAT identifies lots of usability problems that impact users (i.e., higher thoroughness) and does not spend a lot of time and effort on issues that are not important (i.e., higher validity). Additionally, the User-CHAT had higher proportions of usability problems that were rated as serious relative to the other two usability techniques.

The User-CHAT's overall performance suggests that it is appropriate for implementation into the FAA certification process. Usability data generated during User-CHAT evaluation sessions was obtained in real time (which is required during the FAA certification process). Results from the comparative metrics confirm that the data captured in-real time in the User-CHAT sessions was more similar to the data captured from user-testing after hours were spent reviewing the video tapes. The User-CHAT can provide reliable "user-testing type" usability data in real time than either the heuristic evaluation or cognitive walkthrough.

Manufacturers and usability practitioners may find the User-CHAT appealing because the User-CHAT can offer "more bang for the buck." The cost in acquiring usability data is reduced because the data is captured in real time. Thus staff resources necessary to review and transcribe video recordings is reduced. Most of the budget can be applied to recruiting representative end-users to participate in the evaluation. Since the User-CHAT is a modification of user-testing, it seems cost-effective to spend the money on a technique that utilizes end-users in its evaluation process in order to ensure that the product

will be usable by the targeted population. Desurvire et al. (1992) contends that the best method for acquiring data on a product's level of usability is to conduct an evaluation that involves representative end-users. That is, the User-CHAT can help companies maximize returns on investments if implemented into their product design lifecycle.

References

Andre, T.S. (2000). *Determining the effectiveness of the usability problem inspector: A theory-based model and tool for finding usability problems.* Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg.

Andre, T.S., Hartson, H.R., & Williges, R.C. (2003). Determining the effectiveness of the usability problem inspector: A theory-based model and tool for finding usability problems. *Human Factors, 45(3),* 455-482.

Andre, T.S., Williges, R.C., & Hartson, H.R. (1999). The effectiveness of usability evaluation methods: Determining the appropriate criteria. In *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting* (pp. 1090–1094). Santa Monica, CA: Human Factors and Ergonomics Society.

Bailey, R.W., Allan, R.W., & Raiello, P. (1992). Usability testing vs. heuristic evaluation: A head-to-head comparison. In *Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting* (pp. 409-413). Santa Monica, CA: Human Factors and Ergonomics Society.

Bastien, J.M., & Scapin, D.L. (1995). Evaluating a user interface with ergonomic criteria. *International Journal of Human-Computer Interaction, 7*(2), 105-121.

Brooks, P. (1994). Adding value to usability testing. In J. Nielsen & R.L. Mack (Eds.), *Usability inspection methods* (pp. 255-272). New York: John Wiley & Sons.

Clamann, M., & Kaber, D.B. (2004). Applicability of usability evaluation techniques to aviation system. *International Journal of Aviation Psychology, 14*(4), 395-420.

Cockton, G., & Woolrych, A. (2001). Understanding inspection methods: Lessons from an assessment of heuristic evaluation. In A. Blandford, J. Vanderdonckt, & P.D. Gray (Eds.), *People and Computers: Vol 15. Joint Proceedings of HCI 2001 and IHM 2001* (pp. 171-192). Berlin: Springer-Verlag.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3*(4), 299-311.

Cuomo, D.L., & Bowen, C.D. (1994). Understanding usability issues addressed by three user-system interface evaluation techniques. *Interacting with Computers, 6*, 86-108.

Desurvire, H.W. (1994). Faster, cheaper!! Are usability inspection methods as effective as empirical testing? In J. Nielsen & R.L. Mack (Eds.), *Usability inspection methods* (pp. 173-202). New York: John Wiley & Sons.

Desurvire, H.W., Kondziela, J.M., & Atwood, M.E. (1992). What is gained and lost when using evaluation methods other than empirical testing. In *HCI '92 Conference Proceedings on People and Computers VII* (pp. 89-102). New York: Cambridge University Press.

Doubleday, A., Ryan, M., Springett, M., & Sutcliffe, A. (1997). A comparison of usability techniques for evaluating design. In *Designing Interactive Systems (DIS '97) Conference Proceedings* (pp. 101-110). New York: Association for Computing Machinery.

Dumas, J.S., & Redish, J.C. (1999). *A practical guide to usability testing: Revised edition.* Portland: Intellect Books.

Dutt, A., Johnson, H., & Johnson, P. (1994). Evaluating evaluation methods. In G. Cockton, S.W. Draper, & G.R.S. Weir (Eds.), *People and computers IX* (pp. 109-121). Cambridge, UK: Cambridge University Press.

Fu, L., Salvendy, G., & Turley, L. (2002). Effectiveness of user testing and heuristic evaluation as a function of performance classification. *Behaviour & Information Technology, 21*(2), 137-143.

Gray, W.D., & Salzman, M.C. (1998). Damaged Merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction, 13*, 203-261.

Hartson, H.R., Andre, T.S., & Williges, R.C. (2001). Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction, 13*, 373-410.

Holleran, P.A. (1991). A methodological note on pitfalls in usability testing. *Behaviour & Information Technology, 10(5),* 345-347.

Hornbæk, K., & Frøkjær, E. (2004). Usability inspection by metaphors of human thinking compared to heuristic evaluation. *International Journal of Human-Computer Interaction, 17*(3), 357-374.

Huart, J., Kolski, C., & Sagar, M. (2004). Evaluation of multimedia applications using inspection methods: The cognitive walkthrough case. *Interacting with Computers, 16*, 183-215.

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 *CHI '91 Conference Proceedings* (pp. 119-124). New York: Association for Computing Machinery.

Jeffries, R.J., & Desurvire, H.W. (1992). Usability testing vs. heuristic evaluation: Was there a contest? *ACM SIGCHI Bulletin, 24*(4), 39-41.

John, B.E., & Marks, S.J. (1997). Tracking the effectiveness of usability evaluation methods. *Behaviour & Information Technology, 16*(4/5), 188-202.

Karat, C.M., Campbell, R., & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. In *CHI '92 Conference Proceedings* (pp. 397-404). New York: Association for Computing Machinery.

Kessner, M., Wood, J., Dillon, R.F., & West, R.L. (2001). On the reliability of usability testing. In *CHI '01 Conference Proceedings Extended Abstracts* (pp. 97-98). Washington, DC: Association for Computing Machinery.

Landauer, T.K. (1995). *The trouble with computers: Usefulness, usability, and productivity*. Cambridge, MA: The MIT Press.

Lewis, C. (1997). Cognitive walkthroughs. In M.G. Helander, T.K. Landauer, & P.V. Prabhu (Eds.), *Handbook of human-computer interaction* (pp. 717-732). Amsterdam: Elsevier Science.

Lewis, J.R., Polson, P., Wharton, C., & Rieman, J. (1990). Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *CHI '90 Conference Proceedings* (pp. 235-242). New York: Association for Computing Machinery.

Light, R., Singer, J., & Willett, J., (1990). *By design.* Cambridge, MA: Harvard University Press.

Lund, A.M. (1998). The need for a standardized set of usability metrics. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 688-691). Santa Monica, CA: Human Factors and Ergonomics Society.

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.

Mayhew, D.J. (1999). *The usability engineering lifecycle: A practitioner's handbook for user interface design.* San Francisco: Morgan Kaufmann.

Molich, R., & Nielsen, J. (1990). Improving a human-computer dialogue. *Communications of the ACM, 33*(3), 338-348.

Newman, W.M. (1998). On simulation, measurement, and piecewise usability evaluation. *Human-Computer Interaction, 13*, 316-323.

Nielsen, J. (1992). Finding usability problems through heuristic evaluation. In *CHI '92 Conference Proceedings* (pp. 373-380). New York: Association for Computing Machinery.

Nielsen, J. (1994a). Enhancing the explanatory power of usability heuristics. In *CHI '94 Conference Proceedings* (pp. 152-158). New York: Association for Computing Machinery.

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

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *CHI '90 Conference Proceedings* (pp. 249-256). New York: Association for Computing Machinery.

Polson, P., Lewis, C., Rieman, J., & Wharton, C. (1992). Cognitive walkthroughs: A method for theory-based evaluation of user interfaces. *International Journal of Man-Machine Studies, 36*, 741-773.

Prail, A. (1991). Suggestions on collecting observational data. *Common Ground, 1*(2), 3-4.

Rosenbaum, S. (1989). Usability evaluations vs. usability testing: When and why? *IEEE Transactions on Professional Communications, 32*(4), 210-216.

Rowley, D.E., & Rhoades, D.G. (1992). The cognitive jogthrough: A fast-paced user interface evaluation procedure. In *CHI '92 Conference Proceedings* (pp. 389-395). New York: Association for Computing Machinery.

Rubin, J. (1994). *Handbook of usability testing*. New York: John Wiley & Sons.

Scerbo, M.W. (1995). Usability testing. In J. Weimer (Ed.), *Research techniques in human engineering* (pp. 72-111). Englewood Cliffs, NJ: Prentice Hall.

Sears, A. (1997). Heuristic walkthroughs: Finding problems without the noise. *International Journal of Human-Computer Interaction, 9*(3), 213-234.

Shneiderman, B. (1998). *Designing the user interface: Strategies for effective human-computer interaction* (3rd ed.). Reading, MA: Addison-Wesley.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Tan, W., & Bishu, R.R. (2002). Which is a better method of web evaluation? A comparison of user testing and heuristic evaluation. *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 1256-1260). Santa Monica, CA: Human Factors and Ergonomics Society.

Uhlarik, J., Elgin, P.D., & Raddatz, K.R. (2004). *Weather avionics certification assessment tool: Summary of the 2004 modifications to and evaluations of the certification assessment protocol (CAP).* Report for the U.S. Department of Transportation - Federal Aviation Administration [Contract DTFA-02-02-R-03491].

Uhlarik, J., Raddatz, K.R., & Elgin, P.D. (2002). *Cockpit displays of digital weather and operational information: Identification of human factors issues; Part I: Usability assessment of on-board weather graphic displays.* Report for the U.S. Department of Transportation - Federal Aviation Administration [Contract DTFA-02-02-R-03491].

Uhlarik, J., Raddatz, K.R., & Elgin, P.D. (2003). *Graphical weather avionics certification assessment protocol and checklist: Development, evaluation and validation.* Report for the U.S. Department of Transportation - Federal Aviation Administration [Contract DTFA-02-02-R-03491].

Virzi, R.A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors, 34*(2), 457-468.

Virzi, R.A., Sorce, J., & Herbert, L.B. (1993). A comparison of three usability evaluation methods: Heuristic, think-aloud, and performance testing. In *Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting* (pp. 309-313). Santa Monica, CA: Human Factors and Ergonomics Society.

Wang, E., & Caldwell, B. (2002). An empirical study of usability testing: Heuristic evaluation vs. user-testing. *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 774-778). Santa Monica, CA: Human Factors and Ergonomics Society.

Wharton, C., Bradford, J., Jeffries, R., & Franzke, M. (1992). Applying cognitive walkthroughs to more complex user interfaces: Experiences, issues, and recommendations. In *CHI '92 Conference Proceedings* (pp. 381-388). New York: Association for Computing Machinery.

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.

Wright, P.C., & Monk, A.F. (1991). A cost-effective evaluation method for use by designers. *International Journal of Man-Machine Studies, 35*, 891-912.

Appendix A: Determining "Real" Usability Problems

In order to maximize time and personnel resources, the FAA certification process and industry usability practitioners would benefit greatly if they had a tool that identified usability issues that actually impacted user performance (i.e., the HCI literature deems these as real usability problems) and potentially compromised safety while ignoring trivial issues.

According to Hartson et al. (2001), "a usability problem (e.g., found a UEM) is real if it is a predictor of a problem that users will encounter in real work-context usage and that will have an impact on usability (user performance, productivity, or satisfaction or all three)" (p. 383). Emphasizing a usability problem's impact on representative end-users is important because many UEM comparison studies involve inspection methods and a reoccurring issue among inspection methods is that evaluators identify usability problems that do not necessarily affect end-users (Mack & Nielsen, 1994).

There are three practical means to ascertain which usability problems are real: (1) ask representative end-users to evaluate the "realness" of each candidate usability problem (Hartson et al., 2001), (2) ask usability specialists judge the "realness" of each candidate usability problem (Hartson et al., 2001), or (3) compare identified usability problems to a standard list known as the "real" usability problems (Andre, 2000).

"Because usability is ultimately determined by the end user, not an expert evaluator, realness of problems needs to be established by the user" (Hartson et al., 2001, p. 387). With this approach, a few representative end-users are asked to evaluate a list of usability problems to determine which ones they believe will significantly impact their performance. The end-users assign a severity rating to each usability problem identified. High-severity usability problems are issues users judge to greatly affect performance and are more important to fix while low-severity problems have minimal impact and can be remedied after the high-severity problems are resolved.

The realness of usability problems can be obtained also by asking a few usability specialists judge each candidate usability problem and determine if the problem is real or not (e.g., will the problem significantly impact the user or not). Those usability problems deemed real can be summated into a comparison list of usability issues while trivial problems are discarded (Andre, 2000). This technique,

however, shares the same criticism associated with usability inspection methods; a usability problem's "realness" is ascertained primarily through expert opinion.

An alternative to realness judgments by users or specialists is to compare a candidate usability problem to a comparison list of real usability problems. This comparison list of real usability problems serves as a yardstick upon which candidate usability problems are evaluated. If a comparison list existed, the realness of a usability problem found by a UEM could be determined by simply contrasting the candidate usability problem to those in the comparison list (Andre, 2000). If a problem found by the UEM is represented in the comparison list, the problem is considered real. If a candidate problem is not contained in the comparison list, then the problem is deemed trivial.

Producing a Standard List of Real Usability Problems

Some methods to generate a standard-of-comparison list of usability problems include: (1) seeding a system with known usability problems, (2) generating a list through user-testing, and (3) combining lists produced from multiple UEMs.

In some UEM comparison studies, researchers plant known usability problems into a system (e.g., Sears, 1997). With this approach, researchers know of all the existing usability problems in a system (assuming no other real problem exists). A UEM's performance is measured on its ability to detect the seeded usability problems. This seeding technique, however, is not practical in terms of time and resources and thus will not be used in this study.

According to Landauer (1995) and Newman (1998), user-testing is the *de facto* standard used often to generate a comparison list of real usability problems because the problem list is generated by representative end-users performing typical tasks and the usability problems identified actually impact user performance. User-testing, if conducted properly, is a UEM that can produce a high-quality list of real usability problems. However, some usability problems yielded from user-testing can be of questionable realness (e.g., a particular problem may impact only one user). Therefore, Hartson et al. (2001) recommend that usability specialists and end-users review the list of usability problems found through user-testing to further evaluate a problem's realness, thus improving the quality of the comparison list to represent real usability problems.

A comparison list of real usability problems also can be generated by combining the lists of usability problems found by each UEM (Sears, 1997). In order to incorporate this approach, each candidate usability problem must be determined if it is real or not. Non-real problems are discarded. If a usability problem is real, then it is compared to a list that was a combination of usability problems found by all UEMs in the comparison study. If the candidate usability problem is represented in this aggregate list, the problem is discarded; if not, the usability problem is added to the aggregate list (Andre, 2000). The advantage of this technique is that it does not require effort beyond conducting the UEMs in the comparison study. This approach, however, has a serious drawback in that calculating the validity metric is flawed (Hartson et al., 2001).

Appendix B: Heuristic Evaluation Training Materials

Heuristic Evaluation Training Procedure

1. Handout the training package.
    a. Heuristic evaluation description
    b. Heuristics list
    c. Score sheets
    d. Severity rating descriptions
    e. Display screen shots

2. Discuss the heuristics list
    a. Paraphrase the heuristic's description (if necessary), not the heuristics' titles
    b. Think of specific examples for each heuristic

3. Discuss the heuristic evaluation score sheet

4. Discuss the severity ratings

5. Walkthrough and discuss the tasks
    a. If an problem can not be classified, document the problem and create a new heuristic

6. Ask to review periodically the heuristic evaluation procedure and heuristics list using the display screen shots (or other displays if desired)

Heuristic Evaluation Description

Heuristic evaluation is a usability evaluation method for identifying usability problems in an interface. Heuristic evaluation involves having a small set of evaluators examine an interface and judge its compliance with a set of recognized usability principles (the “heuristics”). A heuristic is a guideline/general principle/rule of thumb that can guide and/or critique the design of a display. These heuristics describe common properties of usable interfaces.

During the evaluation session, the evaluator goes through the interface several times and inspects the various display elements and compares them to the heuristics list. In addition to the heuristics list, the evaluator is allowed to consider any additional usability principles or results that come to mind that may be relevant for any specific dialogue element. If a usability problem cannot be classified, evaluators have the freedom to create a new heuristic. In many cases, evaluators are given scenarios to help understand tasks that users accomplish with the system.

The output from the heuristic evaluation is a list of usability problems in the interface with references to those usability principles that were violated. It is not sufficient for evaluators to simply say that they do not like something; they should explain whey they do not like it with reference to one or more heuristics. The evaluator should try to be as specific as possible and should list each usability problem separately. For instance, if there are 3 things wrong with a certain display element, all 3 should be listed with reference to the various usability principles that explain why each particular aspect of the display element is a usability problem. There are 2 main reasons to note each problem separately:

(1) There is a risk of repeating some problematic aspect of a display element, even if it were to be completely replaced with a new design, unless one is aware of all its problems.

(2) It may not be possible to fix all usability problems in a display element or to replace it with a new design, but it could still be possible to fix some of the problems in they are all known.

Heuristic evaluation does not provide a systematic way to generate fixes to the usability problems or a way to assess the probable quality of any redesign. However, because heuristic evaluation aims at explaining each observed usability problem with reference to established usability principles, it will often be fairly easy to generate a revised design according to the guidelines provided by the violated principle for good interactive systems. Also, some usability problems have fairly obvious fixes as soon as they have been identified.

LIST OF HEURISTICS

1. **Visibility of System Status** – the system should always keep users informed about what is going on, through appropriate feedback within reasonable time.

---

2. **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.

---

3. **User Control and Freedom** – users often choose system functions by mistake and will need a clearly marked 'emergency exit' to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.

---

4. **Consistency and Standards** – users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.

---

5. **Error Prevention** – even better than good error messages is a careful design which prevents a problem from occurring in the first place.

---

6. **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.

---

7. **Flexibility and Efficiency of Use** – accelerators – unseen by the novice user – may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

---

8. **Aesthetic and Minimalist Design** – dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.

---

9. **Help Users Recognize, Diagnose, and Recover from Errors** – error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.

---

10. **Help and Documentation** – even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

Severity Ratings

1. *Serious* – usability problems that hinder performance and continue to cause problems even after the problem has been experienced (e.g., the system crashes when a specific action is initiated). These are usability problems that must be resolved because they are design or operation characteristics that could constitute a safety concern when using the system.

2. *Intermediate* – problems also hinder performance but can be overcome through experience (e.g., a menu option does not make sense before the first encounter). These are usability issues that are of great concern because they may have safety concerns and should be resolved but do not necessarily warrant a serious rating.

3. *Minor* – problems are ones that do not hinder performance per se, but are recommendations on how to improve the system's design (e.g., the buttons are different sizes). These usability problems are not associated with safety concerns.

Steps to Complete the Heuristic Evaluation

1. 10 minutes to freely explore the system (do not record any problems at this time).
2. Read the task description.
3. Step through the task using the action sequence provided on the score sheet. Evaluate each display in the action sequence.
4. Record any usability problems and classify. If a usability problem cannot be classified, create a new heuristic and classify the usability problem accordingly.
5. Assign a severity rating to each usability problem identified.
6. Repeat steps 3, 4, & 5 until the all the displays in the action sequence have been evaluated.
7. Upon task completion, record any other notes, comments, concerns, etc. regarding the task.

Appendix C: Cognitive Walkthrough Training Materials

Cognitive Walkthrough Training Procedure

1. Handout the training package
    a. Cognitive Walkthrough description
    b. Score sheets
    c. Severity rating descriptions
    d. Display screen shots

2. Discuss the cognitive walkthrough questions
    e. Paraphrase the questions (if necessary)
    f. Think of specific examples for each questions
    g. Discuss “success” stories

3. Discuss the cognitive walkthrough score sheet

4. Discuss the severity ratings

5. Walkthrough and discuss the tasks

6. Ask to review periodically the cognitive walkthrough procedure and questions using the display screen shots (or other displays if desired)

Cognitive Walkthrough Description

The Cognitive Walkthrough is a usability evaluation method that focuses specifically on evaluating an interface for ease of learning, particularly by exploration. The narrow focus on ease of learning is motivated by the observation that many users prefer to learn a new system through exploration rather than through a training course or reading the accompanying user's manual. That is, users tend to incorporate a "label-following" strategy when interacting with a novel system.

The cognitive walkthrough focuses specifically on user's cognitive problem-solving processes while performing a specific task. Problem solving processes are used to discover correct actions and that the learning mechanisms store representations of correct actions with the users' current goals and tasks contexts. Correct actions are based on their perceived similarity or relevance to the user's current goals. Essentially, the cognitive walkthrough attempts to simulate the user's problem-solving processes where a goal is formulated, an action is selected, and the goal is modified based upon the consequences of the action.

The CE+ theory of exploratory learning provides the underlying theory. Basically, this theory is a 4-step problem-solving process:

1. Start with a rough description of the task they want to accomplish.
2. Explore the interface and select actions they think will accomplish the task.
3. Observe the interface reactions to see if their actions had the desired effect.
4. Determine what action to take next.

Evaluators examine each step necessary to successfully complete a task, attempting to uncover mismatches between users and designers mental models, such as word choices for menu titles and button labels. Additionally, evaluators investigate the system's feedback quality for each action.

During a cognitive walkthrough, evaluators inspect an interface in the context of specified user tasks by adopting the role of the targeted end user. During the walkthrough, evaluators consider each action necessary to successfully accomplish the task. That is, the cognitive walkthrough analyzes the correct action sequence (e.g., most efficient route), asking if the correct sequence will actually be followed by users, suggesting whether users will follow the correct path or not. Thus, the cognitive walkthrough critiques an action sequence that is provided.

For each step in the action sequence, the evaluators consider four primary questions intended to stimulate a story about a typical user's interaction with the system. In particular, the evaluators ask the following four questions:

1. ***Will the users try to achieve the right effect?*** That is, will the user try to do the right thing? For instance, if the task is to save a new document, then the first thing the user must do open the word processing program. Will the user know that he/she should be opening the word processing program?

_____
_____

2. ***Will the user notice that the correct action is available?*** That is, if the user has the right goal, will he/she know what the correct action is? If the action is to select from a visible menu, then there is no problem. But if the word processing icon is buried, then the user may never think of it.

_____
_____

3. ***Will the user associate the correct action with the effect trying to be achieved?*** That is, will the user know if the correct action will achieve the desired effect? If there is a menu option that says, "word processor," things should go smoothly; not so if the menu option says otherwise.

_____
_____

4. ***If the correct action is performed, will the user see that progress is being made toward solution of the task?*** That is, if the correct action is taken, will the user see that things are going OK? If after selecting the word processing program and the system provides a dialog that states, "word processor opening," great. The worst case is no feedback.

_____
_____

While answering each the aforementioned questions, the evaluator(s) documents any usability problems encountered and reasons for those problems. Because of the systematic manner with which the cognitive walkthrough assesses a system, cognitive walkthroughs can thoroughly diagnose usability problems in a microscopic level and provide potential resolutions. That is, the focus of a cognitive walkthrough is spotting problems in an interface, but it also provides an explanation of those problems; those explanations can be useful in generating potential fixes. The output from a cognitive walkthrough is either a success story, constructed by the evaluators, that explains why a user would choose a particular action or a failure story to indicate why a user would not choose the action. If the system is effective, there should be a high degree of mapping between the users' intentions and the interface; the appropriate action should be readily visible.

Some common features of success stories include:

- **Users know what effect to achieve**
  - Because it is part of the original task
  - Because the user has experience using the system
  - Because the system tells the user to do it
- **Users know what actions are available**
  - By experience
  - By seeing some button
  - By seeing a representation of an action (like a menu option)
- **Users know what action is appropriate for the effect they are trying to achieve**
  - By experience
  - Because the interface provides a prompt or label that connects the action to what the user is trying to do
  - Because all other actions look wrong
- **Users know things are going OK after an action**
  - By experience
  - By recognizing a connection between a system response and what the user is trying to do

The following are some general considerations for fixing problems:

- **Will the user be trying to achieve the right effect:**
  - The action might be eliminated
  - A prompt might be provided to tell the user which action must be performed
  - Some other part of the task might be changed so the user will understand the need for the action, perhaps because it is now consistent with another part of the action sequence
- **Will the user know that the correct action is available:**
  - Assign the action to a more obvious control
  - Assign the action to a hidden but more easily discoverable control (e.g., submenu)
- **Will the user know that the correct action will achieve the desired effect:**
  - Provide labels and descriptions for actions that will include words that users are likely to use in describing their tasks
- **If the correct action is taken, will the user see that things are going OK:**
  - Feedback that indicated what happened is better than feedback that just indicated something happened
  - Feedback will be more effective if terms (or graphics) are used that relate to the user's description for the task

Severity Ratings

1. *Serious* – usability issues that hinder performance and continue to cause problems even after the problem has been experienced (e.g., the system crashes when a specific action is initiated). These are usability issues that must be resolved because they are design or operation characteristics that could constitute a safety concern when using the system.

2. *Intermediate* – problems also hinder performance but can be overcome through experience (e.g., a menu option does not make sense before the first encounter). These are usability issues that are of great concern because they may have safety concerns and should be resolved but do not necessarily warrant a serious rating.

3. *Minor* – problems are ones that do not hinder performance per se, but are recommendations on how to improve the system's design (e.g., the buttons are different sizes). These usability problems are not associated with safety concerns.

Steps to Complete the Cognitive Walkthrough

1. 10 minutes to freely explore the system (do not record any problems at this time).
2. Read the task description.
3. Step through the task using the action sequence provided on the score sheet. Evaluate option in the action sequence and the other options on the screen by answering the 4 cognitive walkthrough questions.
4. Record any usability problems (e.g., mismatches between the design and the user's mental model).
5. Assign a severity rating to each usability problem identified.
6. Repeat steps 3, 4, & 5 until the all steps in the action sequence have been evaluated.
7. Upon task completion, record any other notes, comments, concerns, etc. regarding the task.

Appendix D: Example Task Screen Shots for Training Materials

The following tasks and screen shots were captured from the Garmin® GNS 500 Series Simulator. These tasks were administered to the GA pilots after the initial training session on their respective usability method, either the heuristic evaluation or the cognitive walkthrough. The task steps represent correct task action sequences.

Task #1: **Enter a direct to the Wheeler Downtown Airport (KMKC) in Kansas City.**
Start Page: **1st NAV Window**

Step #1: **Press DIRECT TO**

Step #2: **Turn SMALL LOWER RIGHT KNOB (LEFT or RIGHT) to "K"**

Step #3: **Turn LARGE LOWER RIGHT KNOB to the RIGHT to move cursor**

Step #4: **Turn SMALL LOWER RIGHT KNOB (LEFT or RIGHT) to "M"**

Step #5: **Turn LARGE LOWER RIGHT KNOB to the RIGHT to move cursor**

Step #6: **Turn SMALL LOWER RIGHT KNOB (LEFT or RIGHT) to "K"**

Step #7: **Turn LARGE LOWER KNOB to the RIGHT to move cursor**

Step #8: **Turn SMALL LOWER RIGHT KNOB (LEFT or RIGHT) to "C"**

Step #9: **Press ENT**

Step #10: **Press ENT ("ACTIVATE?" is highlighted)**

End Window

Task #2: **On the 2nd NAV window, change the map orientation.**
Start Page: **1st NAV Window**

Step #1: **Turn SMALL LOWER RIGHT KNOB to 2nd NAV window**

Step #2: **Press MENU**

Step #3: **Turn SMALL or LARGE LOWER RIGHT KNOB to highlight "SETUP MAP?"**

Step #4: **Press ENT when "SETUP MAP?" is highlighted**

Step #5: **Turn LARGE LOWER RIGHT KNOB to "ORIENTATION"**

Step #6: **Turn SMALL LOWER RIGHT KNOB to display pop-up window**

Step #7: **Turn SMALL LOWER RIGHT KNOB to move cursor**

Step #8: **Press ENT (when "NORTH-UP" or "TRACK-UP" is highlighted)**

Step #9: **Press CLR to return to the NAV window**

End Window


GARMIN
GNS 530
COM
118.975
119.350
VLOC
108.95
111.30
LOC
ENR
150%-2
GPS
MSG
NAV
THS
HGR
MRB
FDK
EMI
PPM
MXE
DQO
KBWI
KDOV
KIAD
KADW
LDN
CSN
BRV
GVE
KNHK
SBY
SWL
KHAL
RNG
MENU
CLR
ENT
DEFAULT NAV
COM/VLOC
PUSH C/V
CDI
OBS
MSG
FPL
VNAV
PROC
GPS
PUSH CRSR

Task #3: **On the 2nd NAV window, change the 3rd data field to MSA.**
Start Page: **1st NAV Window**

Step #1: **Turn SMALL LOWER RIGHT KNOB to 2nd NAV window**

Step #2: **Press MENU**

Step #3: **Turn SMALL or LARGE LOWER RIGHT KNOB to highlight "CHANGE FIELDS?"**

Step #4: **Press ENT when "CHANGE FIELDS?" is highlighted**

Step #5: **Turn LARGE LOWER KNOB to 3rd DATA FIELD**

Step #6: **Turn SMALL LOWER RIGHT KNOB to display pop-up window**

Step #7: **Turn SMALL or LARGE LOWER RIGHT KNOB to highlight "MSA"**

Step #8: **Press ENT when "MSA" is highlighted**

Step #9: **Press CRSR to remove the cursor**

End Window


GARMIN
GNS 530
COM
118.975
119.350
VLOC
108.95
111.30
LOC
ENR
HPT
KMKC
DTK
287°M
MSA
3800ft
DIS
820nm
GS
0.0kt
THS
HGR
MRB
FDK
EMI
PPM
KBWI
KIAD
KADW
LDN
CSN
BRV
KNHK
GVE
150%-2
GPS
MSG
NAV
RNG
MENU
CLR
ENT
DEFAULT NAV
COM/VLOC
PUSH C/V
CDI
OBS
MSG
FPL
VNAV
PROC
GPS
PUSH CRSR

Task #4: **Find the nearest VOR.**
Start Page: **1st NAV Window**

Step #1: **Turn LARGE LOWER RIGHT KNOB to the RIGHT until "NRST" is displayed**

Step #2: **Turn SMALL LOWER RIGHT KNOB until "NEAREST VOR" is displayed**

End Window

Task #5: **Find the runway configuration for the Manhattan Regional Airport (KMHK).**
Start Page: **1st NAV Window**

Step #1: **Turn LARGE LOWER RIGHT KNOB to "WPT"**

Step #2: **Turn SMALL LOWER RIGHT KNOB to 2nd WPT window**

Step #3: **Press CRSR to display the cursor**

Step #4: **Turn SMALL LOWER RIGHT KNOB to "K"**

Step #5: **Turn LARGE LOWER RIGHT KNOB to the RIGHT to move cursor**

Step #6: **Turn SMALL LOWER KNOB to "M"**

Step #7: **Turn LARGE LOWER RIGHT KNOB to the RIGHT to move cursor**

Step #8: **Turn SMALL LOWER RIGHT KNOB to "H"**

Step #9: **Turn LARGE LOWER RIGHT KNOB to the RIGHT to move cursor**

Step #10: **Turn SMALL LOWER RIGHT KNOB to "K"**

Step #11: **Press ENT**

Step #12: **Press CRSR to remove cursor**

End Window


GARMIN
GNS 530
COM 118.975
119.350
VLOC 108.95
111.30
LOC
ENR
APT KMHK Public
RUNWAY 03-21
INFO
7000' x 150'
Hard Surface
Pilot Cntrl 118.550
MHK
KMHK
13
21
31
03
ALET
1.5% over200m
GPS MSG WPT
RNG
MENU
CLR
ENT
DEFAULT NAV
COM/VLOC
PUSH C/V
CDI
OBS
MSG
FPL
VNAV
PROC
GPS
PUSH CRSR

Appendix E: Demographics Questionnaires

Pilot Demographic Information

Participant #: _____

Sex: **M** **F**

Age: _____

Total number of hours flown: _____

Total number of hours flown within the last 12 months: _____

How would you characterize the majority of your flight time (should equal 100%)?

- Pleasure: _____
- Business: _____
- Training: _____
- Other: _____

What would you estimate to be the average length of your flights? _____

What Certificates / Ratings do you hold? Certificates_____ Ratings_____

During the past 6 months, what percentage of your flying has been on an IFR flight plan? _____%

Do you have experience with MFDs? **Yes** **No**

- If yes, please describe your experience (what type of displays / what type of weather):
  _____
  _____

In what type of aircraft do you spend the majority of your flight time? (List as many as necessary)

_____

_____

Type(s) of aircraft you have flown most frequently in past 6 months_____

Do you have any vision problems that cause difficulties perceiving color? If so, please describe.

_____

_____

Did you have any problem interpreting the colors on the FIS display pages? If so, please describe.

_____

_____

Do you consider yourself adept at decoding the text products - METAR, TAF or PIREP?

Yes No

If no, which text reports do you have difficulty decoding?

_____

Which best describes your approach to learning a new MFD before using it in flight? (Circle one)

1. Attend vendor sponsored tutorial on how to use the MFD
2. Read the complete user's manual from cover to cover
3. Read the user's manual for specific tasks only (e.g., entering a flight plan)
4. Read the quick reference manual only
5. Simply playing around with the MFD, no reading of any supporting documentation

Human Factors Specialists Demographic Information

Participant #:_____

Age: _____

Sex: M F

Academic Level: Bachelors_____ Masters_____ PhD_____

Major: _____

Current Job Title_____

How long have you been at your current position? Years_____ Months_____

Specialty?

Testing/Evaluation_____ Management_____ Design_____

Research_____ Teaching/Academic_____ Other_____

How much experience do you have in the specialty selected above?

Years_____ Months_____

How often do you perform the following activities?

	Never	About once per year	Few times per year	Few times per month	Few times per week	Daily
Usability testing/observation						
Interface design						
Using the cognitive walkthrough						
Using the heuristic evaluation						

Appendix F: Instructions for Each Usability Evaluation Method

Instructions for User-Testing

Prior to evaluating the avionics system

1. Hand out and sign ***Informed Consent Sheets*** and ***Non-Disclosure Agreements***.
2. Explain the purpose of the evaluation and the video recorder.
3. Go through a brief orientation session of the avionics system.
4. Explain the role of the **User**.

    - **User** (the Pilot) will complete a series of tasks.
        - While the Pilot completes the tasks, he/she will be asked to "think out loud"

        "During this evaluation, we are interested in what you say to yourself as you perform some tasks that we give you. In order to do this, we will ask you to think aloud as you work on the tasks. What we mean by think aloud is that we want you to say out loud everything that you say to yourself silently. Just act as if you are alone in the room speaking to yourself. If you are silent for any length of time, we will remind you to keep thinking aloud. Do you understand what we want you to do?"

During the evaluation session

1. Allow 10 minutes of free exploration on the system.
2. Begin the evaluation by giving the **Pilot** the task description found on the ***Task Sheets***, ask him/her to read it out loud, and ask if he/she has any questions about the task before he/she attempts to complete it. Be sure the **Pilot** understands that he/she needs to answer all the questions in the task.
3. Ask the **Pilot** to complete the task and think out loud while completing the task.
4. While the pilot is completing the task, the human factors specialist (sitting outside the pilot's field of view) is free to record on the ***User-testing Score Sheet*** any observations and/or comments made by the user. Additionally, the human factors specialist will be asked to specifically note if any of the following occur:
    a. Did the pilot give up on a task or ask for help?
    b. Did the pilot perform an input error? (e.g., press the wrong button or line-select key)
    c. Did the pilot verbalize confusion and/or difficulty when performing the task?
    d. Did the pilot show considerable delay (in which they just looked at the display and did not initiate an action) in accomplishing part of the task? (e.g., 10 sec or more)
5. Repeat 2, 3, and 4 until all tasks are completed.
6. After the user-testing session is complete, a debriefing session will be conducted by the human factors specialist to elicit comments and feedback about issues specifically regarding the efficacy of the avionics system.

Things for the Pilot to Keep in Mind

- Remember, we're evaluating the avionics system, not you.
- Take your time – this is not a race. You should have plenty of time to get through all the tasks.
- Think out loud – we're trying to understand the thought processes you go through as you complete the task. The only way we can understand this is if you verbally tell us!

Instructions for User-CHAT Sessions

Prior to the Evaluation Session

1. Hand out and sign ***Informed Consent Sheets*** and ***Non-Disclosure Agreements***.
2. Explain the purpose of the evaluation and the video recorder.
3. Go through a brief orientation session of the avionics system.
4. Explain the role of the **User** and **Supervisor**.

    - **User** (the Pilot) will complete a series of tasks.
        - While the Pilot completes the tasks, he/she will be asked to "think out loud" "During this evaluation, we are interested in what you are thinking as you perform some tasks that we give you. In order to do this, we will ask you to think aloud as you work on the tasks. What we mean by think aloud is that we want you to say out loud everything that you say to yourself silently. Just act as if you are alone in the room speaking to yourself. The reason we would like you to think out loud is because not only are we interested in how you complete a task, but we are also interested in your thoughts and processing that happen while you are completing each task. If you are silent for any length of time, we will remind you to keep thinking aloud. Do you understand what we want you to do?"

    - **Supervisor** will record user's performance with each benchmark task

5. Describe the ***Task Sheets,*** the ***User-CHAT Score Sheet***, the ***User-CHAT List of Heuristics***, and the ***Severity Ratings***.

During the Evaluation

1. Allow 10 minutes of free exploration on the system.
2. Begin the evaluation by giving the **Pilot** the task description found on the ***Task Sheets***, ask him/her to read it out loud, and ask if he/she has any questions about the task before he/she attempts to complete it. Be sure the **Pilot** understands that he/she need to answer all the questions in the task.
3. Ask the **Pilot** to complete the task and think out loud while completing the task.
4. **Supervisor** will record user performance relative to the gold standard (documenting First Inefficient Actions, Subsequent Inefficient Actions, Head-down Time, etc.) on the ***User-CHAT Score Sheet***.
5. Upon task completion, both the **Pilot** and **Supervisor** will identify and diagnose usability problems found during task completion by answering the two diagnostic questions found on the ***User-CHAT Score Sheet***.
    a. Classify the usability problem into one or more violated heuristics from the ***User-CHAT List of Heuristics***.
    b. Assign a ***Severity Rating*** for each identified usability problem.
    c. Discuss whether or not the first inefficient action was initiated because the next action was not clearly visible, descriptive, intuitive, etc. thereby causing the pilot to rely on memory in order to recall functionality.

Things for the Pilot to Keep in Mind

- Remember, we're evaluating the system, not you.
- Take your time – this is not a race. You should have plenty of time to get through all the tasks.
- Think out loud – we're trying to understand the thought processes you go through as you complete the task. The only way we can understand this is if you verbally tell us what you are thinking and why you did what you did!

Instructions for Heuristic Evaluation Sessions

Prior to the evaluating the avionics system

1. Hand out and sign ***Informed Consent Sheets*** and ***Non-Disclosure Agreements***.
2. Explain the purpose of the evaluation and the video recorder.
3. Go through a brief orientation session of the avionics system.
4. Describe the ***Heuristic Evaluation Task Sheets***, ***Heuristic Evaluation Score Sheets***, and the ***Severity Ratings***.

Steps

1. 10 minutes to freely explore the system.
2. Read the task description on the ***Heuristic Evaluation Task Sheets***.
3. Step through the task using the action sequence provided on the ***Heuristic Evaluation Task Sheet*** and evaluate all display elements in the action sequence.
4. Record any usability problems on the ***Heuristic Evaluation Score Sheet*** and classify the usability problem into one or more heuristics that were violated. Be as detailed as possible in your description of the problem. If a usability problem cannot be classified, create a new heuristic and classify the usability problem accordingly.
5. Assign a severity rating after a usability problem has been identified and classified.
6. Repeat steps 3, 4, and 5 until all display elements (symbology, labeling, terminology, etc.) associated with the action sequence for the task have been evaluated.
7. Upon task completion, record any other notes, comments, concerns, etc. on the ***Heuristic Evaluation Score Sheet*** that you deem necessary.

Things for the Evaluators to Keep in Mind

- Take your time – this is not a race. You should have plenty of time to get through all the tasks.
- Work together and communicate with each other – each of you brings a unique perspective to the evaluation, so capitalize on that.
- Your opinion is valid – so feel free to record any usability problems you want. If you have any suggestions (no matter how small) that you think will help design a better system, write it down.
- Avoid simply completing the task – we asked for your help in evaluating the avionics system and not to see how well you can complete the task. Besides, you have the right answer in front of you, so your ability to complete the task is not of importance!!

Instructions for Cognitive Walkthrough

Prior to evaluating the avionics system

1. Hand out and sign ***Informed Consent Sheets*** and ***Non-Disclosure Agreements***.
2. Explain the purpose of the evaluation and the video recorder.
3. Go through a brief orientation session of the avionics system.
4. Describe the ***Task Sheets***, the ***Cognitive Walkthrough Score Sheets***, and the ***Severity Ratings***.

Steps

1. 10 minutes to freely explore the system.
2. Read the task description on the ***Task Sheet***.
3. Step through the task using the action sequence provided on the ***Cognitive Walkthrough Score Sheet***. Evaluate each option in the action sequence as well as every other option (either in menus or on the screen) by answering the three questions on the ***Cognitive Walkthrough Score Sheet***.
4. For each question, document either a detailed success story or a detailed failure story on the ***Cognitive Walkthrough Score Sheet***. Refer to the ***Cognitive Walkthrough Success Stories*** and the ***Cognitive Walkthrough Failure Stories*** sheets for examples. You may document multiple stories for each cognitive walkthrough question.
5. Assign a severity rating for each usability problem / issue identified in the failure story(s).
6. Repeat steps 3, 4, and 5 until the all steps in the action sequence have been evaluated.
7. Upon task completion, record any other notes, comments, concerns, etc. regarding the task.

Things for Evaluators to Keep in Mind

- Take your time – this is not a race. You should have plenty of time to get through all the tasks.
- Work together and communicate with each other – each of you brings a unique perspective to the evaluation, so capitalize on that.
- Your opinion is valid – so feel free to record any usability problems you want. If you have any suggestions (no matter how small) that you think will help the manufacturer design a better system, write them down.
- Avoid simply completing the task – we asked for your help in evaluating the avionics system and not just to see how well you can complete the task. Besides, you have the right answer in front of you, so how fast you can complete the task is not important!!
- Be detailed with your stories – for success stories, be explicit as to why the system does a good job (designers like positive feedback too!). For failure stories, be sure to include not only the problem, but a description as to why the problem is a problem so designers can fully understand the issue you identified.

Appendix G: User-testing Score Sheet

USER-TESTING SCORE SHEET
Task:
Participant #:
Things to Consider:
Did the pilot give up on a task and ask for help?
Did the pilot perform an error that required recovery to proceed?
Did the pilot verbalize confusion and/or difficulty when performing the task?
Did the pilot show a delay in accomplishing a part of the task?
Observation, Comments, & Notes
Task Completed Successfully?
Information Interpreted Correctly?

Appendix H: User-CHAT Score Sheet

User-CHAT PROBLEM SCORE SHEET

Task:

Gold Standard	***1st Inefficient Action***	***# of Subsequent Inefficient Actions***	***T***	***Notes:***	***Why was the inefficient action initiated?***	***Why wasn't the gold standard initiated?***	***Severity of Problem***	***M***
GS Action #1								
GS Action #2								
GS Action #3								
GS Action #4								
GS Action #5								

Task Successfully Completed?

Information Interpreted Correctly?

Notes:

GS = "Gold Standard"

Appendix I: User-chat Display Design Heuristics

Heuristic	*Description*
Avoid Absolute Judgments AAJ	Systems should <u>not</u> require users to judge the severity/magnitude level of a symbol (*e.g., weather*) based solely on just one of it's characteristics like color, size, or loudness when that characteristic has more than **5-7 possible levels**. For example, suppose a symbol can achieve one of 10 different colors when it is displayed, with each color supplying the symbol with a different meaning. Research has shown that users have a difficult time associating a given color with its specific meaning when more then 5-7 different colors are possible. The same holds true for different sizes or loudness of tones (i.e., people have a hard time distinguishing the meanings of more than 5-7 tones). _____ _____ _____
Color Population Stereotypes CPS	Colors used to depict levels of severity should follow **population stereotypes**, **norms**, or **standards** (*e.g., red represents the most severe, yellow/amber represents moderate severity, and green represents the least severe*). _____ _____ _____

Consistent use of Design & Labeling CD	Display elements (*e.g., labels, terminology, symbology, icons, etc*.) should be used in a **consistent manner** throughout the system in terms of their:
	• **Location or position on the display**
	  o e.g., currency depictions should be placed in a consistent location on the display; BACK or EXIT button should always appear in the same place regardless of display mode.
	• **Formatting characteristics** (including color coding & color usage, shape coding, size coding, texture coding, etc)
	  o e.g., if red is used to denote severe conditions, it should be used consistently throughout the system and not used for anything other than showing the most severe conditions.
	• **Meaning**
	  o e.g., an EXIT term should always result in exiting the user from something every time the term appears in the display; the symbol used to denote airports should not vary in different display modes.
	• **Menu organization –** the basic organization of the menu structure should not change with display modes.
	_____
	_____
	_____

Design Standards DS	Display elements (*e.g., labels, terminology, symbology, icons, etc*.) should be formatted, used, and arranged according to **established standards**. • When possible, **weather symbology** (e.g., cold/warm fronts) should conform to meteorological standards. o E.g., warm fronts should be depicted in red; cold fronts should be depicted in blue. • All NAVAID symbology should conform to ICAO symbols. • All abbreviations should conform to ICAO standards. • If a system is designed to operate on a standard platform (e.g., Windows), then functionality should conform to that standard. o E.g., drop-down menus, cursor controls, etc. _____ _____ _____
Descriptive Labeling DL	There should be **sufficient and specific description** in the label of a function in order to explicitly **describe what will happen** (e.g., the outcome and/or the information that will be displayed) if the associated function or input device is activated. In other words, system design should avoid the use of terminology and labeling that is vague. _____ _____ _____
No Misleading Labeling ML	The description provided by the labeling and terminology should not **mislead the user** into thinking the option performs one function when in fact it performs a completely different option. • E.g., *the term "EXIT" should never function as an "ENTER" option.* _____ _____ _____
Clear & Visible Labeling CVL	**Labels** should be **clearly and visibly presented** at all appropriate times. All important functions (objects, actions, menu options, labels, functions, etc.) should be labeled clearly and understandably **visible at all appropriate times**. *Users should not be responsible for remembering important information.* **Hidden functions should be avoided**. _____ _____ _____
Map Orientation MO	Map orientation should always be explicitly and clearly **indicated at all times**. Users must never question what orientation is being depicted (*e.g., north-up, heading-up, track-up, or desired track-up*). _____ _____ _____

Information Need **IN**	**All information necessary** for making a decision should be **available, easily accessible,** and **easily understood**. • *Graphical depictions of weather information should be accompanied by text versions of that same weather phenomenon, with the text version supplying more detailed information.* o E.g., When wind at an airport is displayed graphically as a red arrow, the pilot should have some way of also accessing actual wind speed information (e.g., 30 knots) when needed. _____ _____ _____
Information Currency **IC**	There should be a continuous, easily visible, easily understood, and valid **indication of the currency** of the displayed information (e.g., time stamp), especially weather information. That is, there should always be an indication of how old the presented information is. _____ _____ _____
Information Grouping **IG**	Based upon **logical expectations** and **relevant past experiences** of the users, different types of information that **share conceptual similarities** should be **grouped together** in the display. • E.g., most pilots expect NEXRAD depiction and METARs to be grouped together under a weather menu labeled “WX” because both products are conceptually similar – as they are both weather products. _____ _____ _____
Intuitive Symbology **IS**	The meaning of the symbology should be **intuitive or easily understood.** In other words, the meaning of the symbology must be **clear** and **not require excessive thought** to interpret. _____ _____ _____
Visible & Distinct Symbology **CVS**	The depiction of the symbology should be **easily visible** and **easily distinguishable or distinct from other symbology**. That is, symbology representing one type of information should look distinctively different from symbology representing a different type of information - no two types of symbols should be confused with each other. • E.g., the symbol for lightning strikes should be easily distinguishable from the symbol for traffic, especially if lightning and traffic are able to be overlaid. _____ _____ _____

Legend **L**	A **legend** should be available to provide further information about the meaning of color coding, shape coding, texture coding, or size coding, etc. The legend should be **accessible by 1 key press or input action**. _____ _____ _____
Match between System and Real World **MSRW**	Display elements should **look like** and **move like** the environmental variables they represent. • E.g., convective weather should move in a spatial pattern and direction consistent with its real world path. _____ _____ _____
Menu Accessibility **MA**	If menus aren't always present on the screen, they should be **easily accessible** (within one key press). _____ _____ _____
Menu Removal (if necessary) **MR**	Menus should not occlude important display information. If a menu is temporarily superimposed on a display (as in drop-down menus or pop-up menus), users should be able to **remove the menu with minimal key presses**. A superimposed menu should also automatically **"time-out**" after a short duration. _____ _____ _____
Minimizing Information Access Cost **MIAC**	The system functions and menu structure should be organized such that two or more **function**s that are **frequently accessed together should be able to be accessed by 1 input action or key press**, so that the cost of traveling between these functions is small. However, these two or more functions need NOT be conceptually similar – just merely two or more functions that usually need to be accessed or performed together. • *E.g., often pilots need to change the range of view for a weather display; thus, the system should allow the pilot to change the view range while simultaneously viewing the weather display.* . _____ _____ _____
Number of Menu Options **NMO**	The number of options per menu should range from **4 to 13**, depending upon the amount of available display real-estate. _____ _____ _____

Display Proximity for Mental Integration **DPMI**	If two or more sources of **information** are **related to the same task** and must be **mentally integrated** in order to complete the task, then these sources of information should have **close display proximity –** in other words these two information sources should be **presented very close to each other.** Close display proximity can be accomplished by: • *Placing* the two information sources *side-by side* on the display or *superimposing* them; o e.g., NEXRAD depiction overlaid with stormscope data. • Presenting both information sources in the *same color*; • *Linking* the two information sources with *lines*; • *Configuring* the information sources in a *spatial pattern* that results in an *emergent feature.* These two information sources that need to be mentally integrated may be conceptually *similar* (e.g., overlaying NEXRAD and stormscope weather data) or conceptually *different* (e.g., overlaying NEXRAD with traffic information). _____ _____ _____
Reduce Mental Workload **RMW**	Steps should be taken in order to **reduce the user's mental workload** when interacting with the system. The user should not need to perform any **unnecessary mental calculations** when using the display. • *E.g., pilots should not be required to calculate the currency of weather information by subtracting the time the weather information was generated from the current time* _____ _____ _____
Redundant Coding of Information **RCI**	When the same message is presented more than once, it will be more likely to be interpreted correctly. This will be particularly true if the **same message is presented in multiple formats** (*e.g., auditory and text).* Thus, conditions that may degrade one form (*e.g. noise degrading an auditory message*), may not degrade another (*e.g. text*). • Color should not be the sole means of obtaining information about the severity/magnitude of a variable. **Color should be used along with another dimension** (*e.g., shape, size, etc*.) in order to display meaning. • *E.g., Pilots should be able to access weather information in both graphical and text format.* _____ _____ _____
Frequently Used Information **FUI**	Important information and/or frequently used information should be **readily accessible** and not buried under many layers in the menu structure. Frequently or repeatedly performed tasks should be shortened by "hot keys" or "short-cut keys." _____ _____ _____

User Expectations & Past Experience **EPE**	Systems should use concepts, ideas, metaphors, menu organization, terminology, etc., that are well known to users, thereby **capitalizing on user expectations**. User expectations are based on past experiences and/or logical expectations. If it is absolutely necessary that a display element (*e.g., menu option, label, function, etc.*) contradict these expectations, then it is even more important that the corresponding labels be explicitly descriptive of its "unusual" outcome. _____ _____ _____
Unnecessary Information **UI**	Task irrelevant and/or rarely needed information should not be constantly visible. Systems should support a means of systematically **decluttering** and/or removing information. _____ _____ _____
Undo/Exit Functions **UEF**	User should be allowed to move freely in the system and should be able to **undo actions** and **exit** from undesired screens (*e.g., the system should support UNDO, REDO, and EXIT functions*). _____ _____ _____
Alternative Routes **AR**	The system should support **alternative routes** for accessing the same information. _____ _____ _____
Visibility of System Status **VSS**	The system should keep pilots informed about the **status of the system** through timely **feedback** (*e.g., an hourglass can be used to show that the system has acknowledged the user input and is processing information*). _____ _____ _____
Trial and Error **T&E**	Actions classified as trial and error results when **the user does not know** exactly what the correct input action is and consequently begins to **systematically search** for the correct action. The search must follow some strategy (*e.g., I am going to press every option and see what it does*). _____ _____ _____

Appendix J: Heuristic Evaluation Score Sheet

HEURISTIC EVALUATION PROBLEM RECORD FORM					
Task:					
Name of Display Window	***Description of Problematic Display Element***	***How was the Problem Found (general observation, performing an action, etc?***	***Heuristic(s) Violated***	***Severity of Problem***	***Notes***

Appendix K: Cognitive Walkthrough Score Sheet

COGNITIVE WALKTHROUGH SCORE SHEET				
Task:				
Action Sequence	***Cognitive Walkthrough Questions***	***Success Story***	***Failure Story***	***Severity of Problem***
Action #1	Will the user notice the correct action is available?			
	Will the user know that the correct action will achieve the desired effect?			
	If the correct action is performed, will the user see that things are going OK?			
Action #2	Will the user notice the correct action is available?			
	Will the user know that the correct action will achieve the desired effect?			
	If the correct action is performed, will the user see that things are going OK?			
Action #3	Will the user notice the correct action is available?			
	Will the user know that the correct action will achieve the desired effect?			
	If the correct action is performed, will the user see that things are going OK?			
Action #4	Will the user notice the correct action is available?			
	Will the user know that the correct action will achieve the desired effect?			
	If the correct action is performed, will the user see that things are going OK?			
Action #5	Will the user notice the correct action is available?			
	Will the user know that the correct action will achieve the desired effect?			
	If the correct action is performed, will the user see that things are going OK?			
Notes:				

Appendix L: Usability Problems classified into Heuristics

The following tables show the break-down of the display design heuristics used to classify the usability problems identified by each usability technique. Heuristics that were not used (i.e., frequency equal zero) are excluded. The totals for the frequency column are greater than the total number of usability problems detected for each UEM because more than one display design heuristic could be used to classify each usability problem.

Table L 1. Numbers and Percentages of Heuristics used to Classify Usability Problems Identified through the User-CHAT.

Rank	Heuristic	Frequency	Percentage of Classifications
1	Descriptive Labeling	134	22.30%
2	Design Standards	72	11.98%
3	Information Need	57	9.48%
4	User Expectations and Past Experience	49	8.15%
4	Clear and Visible Labeling	49	8.15%
6	Color Population Stereotypes	42	6.99%
7	Intuitive Symbology	36	5.99%
8	Display Proximity for Mental Integration	24	3.99%
9	Legend	22	3.66%
10	Visible and Distinct Symbology	18	3.00%
11	No Misleading Labeling	17	2.83%
12	Information Grouping	16	2.66%
13	Consistent use of Design and Labeling	13	2.16%
14	Reduce Mental Workload	11	1.83%
15	Minimizing Information Access Cost	9	1.50%
16	Unnecessary Information	8	1.33%
17	Redundant Coding of Information	7	1.16%
18	Match between System and Real World	5	0.83%
19	Information Currency	4	0.67%
19	Alternative Routes	4	0.67%
21	Map Orientation	2	0.33%
21	Frequently Used Information	2	0.33%
	Total	601	

Table L 2. Numbers and Percentages of Heuristics used to Classify Usability Problems Identified through the Heuristic Evaluation.

Rank	Heuristic	Frequency	Percentage of Classifications
1	Descriptive Labeling	97	21.65%
2	Design Standards	86	19.20%
3	Clear and Visible Labeling	46	10.27%
4	Color Population Stereotypes	34	7.59%
5	Information Need	30	6.70%
6	Unnecessary Information	28	6.25%
7	User Expectations and Past Experience	23	5.13%
8	Intuitive Symbology	18	4.02%
9	Consistent use of Design and Labeling	14	3.13%
10	Information Grouping	12	2.68%
11	Visible and Distinct Symbology	11	2.46%
12	Legend	10	2.23%
13	Display Proximity for Mental Integration	9	2.01%
14	No Misleading Labeling	7	1.56%
14	Reduce Mental Workload	7	1.56%
16	Minimizing Information Access Cost	5	1.12%
17	Frequently Used Information	2	0.45%
17	Undo/Exit Functions	2	0.45%
17	Visibility of System Status	2	0.45%
20	Map Orientation	1	0.22%
20	Match between System and Real World	1	0.22%
20	Menu Accessibility	1	0.22%
20	Redundant Coding of Information	1	0.22%
20	Alternative Routes	1	0.22%
	Total	448	

Table L 3. Numbers and Percentages of Heuristics used to Classify Usability Problems Identified through the Cognitive Walkthrough.

Rank	Heuristic	Frequency	Percentage of Classifications
1	Descriptive Labeling	135	39.13%
2	Design Standards	53	15.36%
3	Display Proximity for Mental Integration	40	11.59%
4	Clear and Visible Labeling	19	5.51%
5	Intuitive Symbology	13	3.77%
6	User Expectations and Past Experience	13	3.77%
7	Visible and Distinct Symbology	12	3.48%
8	No Misleading Labeling	9	2.61%
9	Information Need	8	2.32%
10	Color Population Stereotypes	7	2.03%
10	Information Grouping	7	2.03%
10	Visibility of System Status	7	2.03%
13	Menu Accessibility	6	1.74%
14	Consistent use of Design and Labeling	4	1.16%
15	Alternative Routes	3	0.87%
16	Match between System and Real World	2	0.58%
16	Minimizing Information Access Cost	2	0.58%
16	Reduce Mental Workload	2	0.58%
19	Number of Menu Access Cost	1	0.29%
19	Redundant Coding of Information	1	0.29%
19	Unnecessary Information	1	0.29%
	Total	345	

Appendix M: Summary Inferential Statistics

The following tables present summary statistics and post-hoc comparisons for thoroughness, validity, and effectiveness for usability problems that were identified in a least two evaluation sessions and in at least three evaluation sessions. Because Wilks' Lambda failed to find significance between the UEMs when comparison list was comprised of usability problems identified in a least four evaluation sessions, the corresponding summary and inferential statistics are not presented.

Usability Problems Identified in at least Two Evaluation Sessions

Table M 1. Univariate ANOVA Summary Tables when Inclusion Criterion was Two Evaluation Sessions

Source	*Dependent Variable*	*df*	*SS*	*MS*	*F*	*p*	*Power*	ω^2
	Thoroughness							
Method		2	.084	.042	9.51	<.05	0.85	0.66
Error		6	.027	.004				
Total		8	.111					
	Validity							
Method		2	.125	.062	8.05	<.05	0.79	0.61
Error		6	.046	.008				
Total		8	.171					
	Effectiveness							
Method		2	.091	.046	19.51	<.05	0.99	0.81
Error		6	.014	.002				
Total		8	.105					

MANOVA Statistics: Wilks' Lambda (λ) = 0.08, F(6, 8) = 3.42, p<.10

Table M 2. Tukey's HSD Post Comparisons for Thoroughness, Validity, and Effectiveness when Inclusion Criterion was Two Evaluation Sessions

Dependent Variable	*(I) Usability Method*	*(J) Usability Method*	*Mean Difference (I – J)*
Thoroughness	User-CHAT	Heuristic Evaluation	.15*
		Cognitive Walkthrough	.23**
	Heuristic Evaluation	Cognitive Walkthrough	.08
Validity	User-CHAT	Heuristic Evaluation	.29**
		Cognitive Walkthrough	.17
	Heuristic Evaluation	Cognitive Walkthrough	-.12
Effectiveness	User-CHAT	Heuristic Evaluation	.21**
		Cognitive Walkthrough	.21**
	Heuristic Evaluation	Cognitive Walkthrough	.00

**p<.10, **p<.05*

Usability Problems Identified in at least Three Evaluation Sessions

Table M 3. Univariate ANOVA Summary Tables when Inclusion Criterion was Three Evaluation Sessions

Source	*Dependent Variable*	*df*	*SS*	*MS*	*F*	*p*	*Power*	ω^2
	Thoroughness							
Method		2	.115	.057	8.28	<.05	0.80	0.62
Error		6	.042	.007				
Total		8	.156					
	Validity							
Method		2	.050	.025	5.82	<.05	0.65	0.53
Error		6	.026	.004				
Total		8	.076					
	Effectiveness							
Method		2	.056	.028	16.91	<.05	0.98	0.78
Error		6	.010	.002				
Total		8	.065					

MANOVA Statistics: Wilks' Lambda (λ) = 0.09, F(6, 8) = 3.17, p<.10

Table M 4. Tukey's HSD Post Comparisons for Thoroughness, Validity, and Effectiveness when Inclusion Criterion was Three Evaluation Sessions

Dependent Variable	*(I) Usability Method*	*(J) Usability Method*	*Mean Difference (I – J)*
Thoroughness	User-CHAT	Heuristic Evaluation	.19*
		Cognitive Walkthrough	.27**
	Heuristic Evaluation	Cognitive Walkthrough	.08
Validity	User-CHAT	Heuristic Evaluation	.18**
		Cognitive Walkthrough	.09
	Heuristic Evaluation	Cognitive Walkthrough	-.09
Effectiveness	User-CHAT	Heuristic Evaluation	.17**
		Cognitive Walkthrough	.16**
	Heuristic Evaluation	Cognitive Walkthrough	-.01

*p<.10, **p<.05