The User-CHAT as a Tool for Implementation into the Current FAA Certification Process: A Report on Qualitative Feedback from Certification Personnel

Report for the U.S. Department of Transportation - Federal Aviation Administration [~~Contract DTFA-02-02-R-03491~~]

05-6-001

John Uhlarik, Peter D. Elgin, & Kimberly R. Raddatz
Kansas State Aviation Research (KStAR) Lab
Department of Psychology
Kansas State University
Manhattan, Kansas

March 8, 2007

Executive Summary

For the past several years, researchers at the Kansas State Aviation Research (KStAR) lab at Kansas State University (KSU) have been developing and evaluating the User-Centered Hybrid Assessment Tool (*User-CHAT*) as a candidate certification protocol that maximizes efficient and comprehensive diagnosis of interface problems identified from a behaviorally-based perspective while minimizing time and resource limitations typically associated with the FAA certification environment. The *User-CHAT* is a unique usability method that extracts suitable components from several well-established usability methods (e.g., Heuristic Evaluation, Cognitive Walkthrough, and User-testing) to satisfy FAA certification environment constraints.

To date, 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. The User-CHAT has yet to be evaluated by FAA personnel with ideal backgrounds for assessing candidate avionics systems. Thus the present study investigated the User-CHAT's efficacy in an FAA certification environment with evaluators who have expertise in both human factors and aviation. One ACO and one AEG official used the User-CHAT to evaluate a commercially available MFD. The primary focus of the evaluation was to expose certification personnel to the User-CHAT in order to obtain qualitative feedback regarding the User-CHAT's strengths and weaknesses as a potential tool for implementation into the current certification process. Weaknesses in the terminology, user's guide, and score sheet identified by the ACO and AEG officials were resolved in the latest iteration of the User-CHAT.

Table of Contents

Introduction

For the past several years, researchers at the Kansas State Aviation Research (KStAR) lab at Kansas State University (KSU) have been developing and evaluating the User-Centered Hybrid Assessment Tool (*User-CHAT*) as a candidate certification protocol that maximizes efficient and comprehensive diagnosis of interface problems. The behaviorally-based protocol minimizes time and resource limitations typically associated with the FAA certification environment. The *User-CHAT* is a unique hybrid usability method that extracts suitable components from several well-established usability methods (e.g., Heuristic Evaluation, Cognitive Walkthrough, and User-testing) to satisfy constraints of the FAA certification environment. In order to validate the *User-CHAT* as an efficient, effective, and useful certification aid, the following assertions about the *User-CHAT* have been tested and confirmed through empirical evaluation:

1. The *User-CHAT* supports an evaluation within 3-4 hours.
2. The *User-CHAT* supports an evaluation by evaluators with little or no formal human factors knowledge or training.
3. The *User-CHAT* does not require extensive training of evaluators.
4. The *User-CHAT* (requiring the identification of gold standards) can be applied to the evaluation of complex systems supporting multiple gold standards for benchmark task completion.
5. Relative to the parent usability methods from which it was derived:
    a. The *User-CHAT* better facilitates evaluators in detecting important and ignoring non-important usability problems in a candidate interface.
    b. The *User-CHAT* helps evaluators identify a greater proportion of problems rated as serious (i.e., usability problems that could compromise safe flight and thus must be resolved).

Throughout its development, the utility of the User-CHAT 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 evaluations have used pilots without any human factors backgrounds. However, the User-CHAT's effectiveness has yet to be evaluated using FAA Certification personnel with human factors and aviation backgrounds (i.e., double experts). Therefore, the next logical extension of the User-CHAT validation process is to investigate its efficacy when used by 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. Any shortcomings found (based on from verbal feedback obtained after the evaluation session) will be incorporated to improve the User-CHAT. Second, the ability of the User-CHAT to support evaluators with expertise in both human factors and aviation can help determined how domain expertise in both human factors and aviation influences the evaluator's ability to accurately document usability problems. Feedback from these evaluators, both quantitative and qualitative, will be integrated to improve the User-CHAT as a marketable usability assessment technique.

Note about the rest of the report

In previous versions of the User-CHAT, the two participants' roles were referred to as the "user" who completes the tasks and the "supervisor" who records user performance and leads the evaluation process. However, upon the suggestion of the FAA personnel who participated in this most recent evaluation, the role of the supervisor will, from this point forward, be referred to as the "observer." More explanation about this terminology change will be given in later sections of this paper.

Method

Participants

Two people (one from the Wichita Aircraft Certification Office – ACO, the other from the Wichita Aircraft Evaluation Group – AEG) volunteered to participate in a mock evaluation of a candidate avionics system. The ACO official had a Ph.D., a pilot's license with a commercial / instrument rating, approximately 500 hrs of total flight time, and has spent more than 12 years with FAA Certification / Flight Safety.

System

A commercially available MFD (multi-function display) capable of supporting weather (text and graphic), traffic, terrain, communications and flight plan information was used. The MFD used dedicated function keys, soft keys with corresponding labels, and knobs as input devices, and was the same MFD used in previous User-CHAT evaluations. The ACO official had some previous experience with the MFD while the AEG official had no experience with the MFD.

Benchmark Tasks

Eight benchmark tasks (representative of the primary functions supported by the MFD) were completed during the evaluation session. User performance and qualitative feedback from both participants were obtained for each of the following tasks:

1. Create a flight plan (from KSLN to KOSH) by inserting the waypoints KSLN, LN, DSM, KCID, RACNY, and KOSH. Then return to the VFR map.
2. Delete only the fourth waypoint (KCID) from the flight plan and return to the VFR map.
3. Circumnavigate to the right of your current course (to avoid a storm) by inserting one waypoint using cursor controller.
4. Access and view tower and ground frequencies at the Wheeler Downtown Airport in Kansas City (KMKC), using the Joystick. Then return to the VFR map.
5. Narrow the number of closest airports that appear in the list by only showing airports with runway length of 5000ft or more. Identify closest airport with a runway of approximately 7000ft, find its elevation, and then return to the VFR map.
6. On the IFR map, change the 2nd data field from the top to depict the minimum safe altitude (MSA).
7. Select and deselect traffic overlay on the VFR map.
8. Display a text METAR report for Washington D.C.'s Reagan International Airport (KDCA). Find the age of that text METAR report.

Review of User-CHAT Protocol

The User-CHAT is comprised of two major phases (*unstructured exploration* and *structured exploration*) and requires at least two people, a User and an Observer. The evaluation session begins with the *unstructured exploration* phase in which the User is allowed to freely explore the functionality of the system for approximately 10-15 min and practice thinking out loud while the Observer practices recording user performance.

During *structured exploration*, the User completes each benchmark task, verbalizing his/her thought processes while stepping through the task while the Observer compares the User's performance to the "gold standard," or most efficient action sequence for task completion. When user performance deviates from the gold standard action sequence, the Observer records the first inefficient action (i.e., the first user action that departed from the gold standard action sequence) and tallies each subsequent inefficient action until the User initiated the next required gold standard action towards task completion. User-CHAT instructions emphasize the importance of the Observer correctly identifying (i.e., naming) the first inefficient action and merely estimating the number of subsequent inefficient actions. In addition, if the User exhibits an "extended amount of head-down time" (e.g., 10 sec or more contemplating the next step) during task completion, the Observer records this instance of excessive head-down time with a "T" at the appropriate place on the score sheet.

When a benchmark task is complete (i.e., the target information is displayed), the User interprets the meaning and/or implication of the displayed information (e.g., symbology, terminology, etc.) while the Observer compares this interpretation with what the manufacturer intended. For each misunderstanding, both the User and Observer determine and document the reasons for the misinterpretation. Next, for each first inefficient action initiated by the User and recorded by the Observer, both the User and Observer discuss and identify the best answer to two questions:

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

The User and Observer then classify the reason(s) for the inefficient actions as violations of one or more general display design and usability heuristics. Then a severity rating is assigned to each usability problem using the following:

- *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.

- *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.

- *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.

Once a severity rating has been assigned, the User and Observer discuss if memory influenced user 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? If so, the Observer places an "M" in the Memory column at the appropriate place on the score sheet.

Procedure for the Evaluation Session

For the current evaluation of the User-CHAT, the AEG official assumed the role of the User and the ACO official assumed the role of the Observer. Prior to the evaluation session, the User-CHAT's list of heuristics (see Appendix A) and User Guide were sent to the ACO for review. On the day of the evaluation, these participants were also given a brief 30 min overview of the User-CHAT's protocol.

The User and Observer completed *unstructured exploration* (10 min), *structured exploration* (2 hrs), and a debriefing session (30 min) in approximately 3 hrs. During unstructured exploration, the User freely interacted with the MFD and practiced thinking aloud, while the Observer practiced documenting user performance on the benchmark task score sheet. During structured exploration, the User completed the benchmark tasks and the Observer documented user performance relative to the task's gold standard(s), recorded notes, head-down time, etc.

It should be noted that, although the User-CHAT requires that all first inefficient actions be classified as violations of one or more heuristics, the User and Observer did not perform this heuristic classification during the current User-CHAT evaluation. The heuristic classification was tabled primarily because of the participants' lack of familiarity with the heuristics list and the fact that the AEG official (the User) was a last minute replacement and did not have a chance to look at the User-CHAT's User Guide or the list of heuristics. However, the ACO official (the Observer) did make note of first inefficient actions that were violations of specific FAA regulations, which provides an approximation of how the heuristics classification is intended to function.

After the structured exploration, the AEG and ACO officials spent 30 min discussing the strengths and weaknesses of the User-CHAT and offered suggestions for how to make the User-CHAT more valuable to the current certification process. The results presented in this report focus primarily on the qualitative feedback received during the unstructured and structured exploration and the debriefing period.

Results and Discussion

The results of previous User-CHAT evaluation studies provided quantitative data that speak to the ability of the Observer to record behavioral and subjective usability data during an evaluation session. Because of prior research and the time constraints of the participants in this current evaluation, less emphasis was placed on obtaining quantitative data and more emphasis was placed on obtaining qualitative feedback, especially feedback regarding the practical issues associated with using the User-CHAT within the certification environment.

Qualitative feedback was parsed according to the specific areas of the User-CHAT protocol to which it was most relevant, beginning with a general summary of strengths and weaknesses identified during the evaluation and debriefing sessions.

Overall User-CHAT Strengths and Weaknesses

The ACO professional felt that the User-CHAT can be of great value to the current certification process. Specifically, he remarked that the User-CHAT was "one of the best things he's seen come across his desk in a long time." Table 1 presents the strengths and weaknesses identified by the ACO:

Table 1: Overall Strengths and Weaknesses Identified by the ACO Official

Strengths	***Weaknesses***
▪ Emphasis of the User-CHAT for obtaining and basing the evaluation of the candidate avionics system on behavioral data.	▪ Use of the term "supervisor" to describe the FAA evaluator is problematic especially when used within the context of an avionics certification; "observer" does not carry any connotation of hierarchy or judgment.
▪ Facilitation of dialog between the pilot (User) and the FAA evaluator (Observer).	▪ Score sheets printed on legal sized paper are awkward.
▪ Structured approach to data collection and the ability to capture a lot of good data, especially issues that might be missed if not for the structured approach of the User-CHAT.	▪ Lack of space to record notes and subsequent inefficient actions on the score sheet. This ACO official prefers to document the actual subsequent inefficient action rather than simply tallying the number of subsequent inefficient actions.
▪ Ability of the User-CHAT to capture the effect of head-down time and the need for reliance on memory to overcome the issue in the future.	▪ Gold standards for benchmark tasks are broken down in to components that are too specific for the FAA's evaluation purposes (e.g., not necessary to have the task of entering in a waypoint divided into actions such as entering in each specific letter, controlling the cursor position, etc.).
▪ Emphasis placed on assessing the "label-following" strategy of novice users which parallels FAA approach to evaluating a system; enabling "power" uses is less emphasized.	▪ Unlikely that manufacturer will provide an exhaustive list of gold standards for each benchmark task; more reasonable to ask manufacturer to provide one "primary" and one "secondary" gold standard for each task (where appropriate).

The following sections present qualitative feedback regarding specifics of the User-CHAT protocol and accompanying materials. The ACO and AEG officials offered the following suggestions to improve specific aspects of the User-CHAT.

User-CHAT User Guide

- User Think Aloud. More emphasis should be placed on the importance of the User (pilot) thinking out loud while interacting with the system. Also, more description should be included on how the Observer should instruct/remind the User to think out loud during the evaluation.
- Observer Authority. More emphasis should be placed on counseling the Observer that he/she has the authority and the obligation to pause the User during the evaluation to ask for clarification or to catch up on data recording. The cost to disrupting User momentum during the task completion sequence is not as great as the cost of missing something if the Observer is unable to record important information.

Benchmark Tasks & Gold Standards

- Nature of Benchmark Tasks. Suggested that for each of the avionics system's major functions there should be a corresponding benchmark task. Also, both officials felt that the evaluation of the candidate system should focus more on benchmark tasks that must be performed in the air rather than tasks that are more set-up in nature and can be performed on the ground (e.g., flight planning). The ACO official suggested that benchmark tasks should reflect "operation critical" tasks (e.g., finding nearest airport in an emergency).
- Multiple Gold Standards. The ACO official acknowledged that it would be time-consuming for manufacturers to provide all possible gold standard task paths for each benchmark task. Instead, a more reasonable requirement is for manufacturers to provide one "primary" and one "secondary" gold standard. Limiting the number of gold standards per benchmark task would simplify the data recording procedure, as the ACO official found it difficult to identify the correct gold standard the User was following for a task with more than two gold standards.
- Emphasis on Novices. The ACO official was less concerned with evaluating gold standards designed to support experienced or "power" users (e.g., use of hidden or unlabeled short-cuts). The FAA's primary goal during an evaluation is to see if the "menu driven" gold standard allows the pilot to access information without much difficulty (e.g., ensure the interface supports a label-following strategy). The FAA acknowledges the importance of designing for the power user but is more concerned with evaluating how well the system supports the untrained user to reach the target information without too many serious problems.
- Additional Logical Task Completion Sequences. If the number of gold standards assessed is restricted to two (a primary and alternate gold standard), there is a definite possibility that the user will find a logical task completion path that does not encompass either gold standard. In this case, the User and Observer need to discuss why and how the User found another gold standard (e.g., was the User being creative and innovative because he/she *wanted* to be or because he/she *had* to be?) In other words, the User and Observer will evaluate whether the User had to explore other task paths because the display did not support the task being performed through the gold standard path through intuitive labeling, menu structure, etc.

Score Sheets

- Provide more space to support the ability to take general notes on user performance.

- Provide more space to specifically document:
  1. Specific subsequent inefficient actions,
  2. The answer to a 3rd cognitive walkthrough question that asks how the user recovered from an inefficient action, if recovery occurred (see below).

- Alter the format of the score sheet so that it may be printed on letter-sized instead of legal sized paper enabling the score sheets to fit on a standard clipboard.

- Provide all information the Observer needs for recording task completion on one page (i.e., the gold standard, 1st inefficient actions, subsequent inefficient actions, time component, notes field) and provide all of the diagnosis materials on a second page (i.e., cognitive walkthrough questions, memory component, severity ratings, heuristics classification).

- Provide separate score sheets for each gold standard with a top page that provides a selection rule for the subsequent score sheets based on the user's first action. For example,
  - *"If the User selects WX, then go to packet 1."*
  - *"If the User selects MENU, then go to packet 2."*

- When the User is required to input some piece of information (e.g., a waypoint identifier), it is not necessary to break up that input into its component actions on the score sheet (e.g., "rotate knob," "scroll to M," "cursor to next field," "rotate knob," "scroll to H," etc.).

- The ACO official (Observer) found it more efficient and informative to connect gold standard actions with an arc where the User made a correct transition between gold standard actions.

Cognitive Walkthrough Questions

- Inefficient Action Recovery. The ACO official appreciated the data that came from asking the two cognitive walkthrough questions ("*Why was the inefficient action chosen?*" and "*Why wasn't the efficient action chosen?*"). However, he suggested adding a third question that evaluated how the User was able to recover from an inefficient action.
  - *"What about the interface helped you recover from the inefficient action?"*

Severity Ratings

- Evaluation of Usability Problem Consequence. The decision to assign a severity rating to an inefficient action should take into account how easily recoverable it is and what type of consequence that usability problem has for flight safety. For example, the consequence of the pilot entering the wrong waypoint during preflight planning is vastly different than the consequence of the pilot's inability to quickly locate the "direct to" option during a flight emergency). In other words, severity ratings should indicate the impact of the usability problem and whether or not a usability problem was easy to recover from (which may be classified as *Minor*), difficult to recover from (which may be classified as *Intermediate*), or nearly impossible to recover from (which may be classified as *Serious*).

Implications

As a result of the feedback obtained during the present evaluation, many changes have been made to the User-CHAT's User's Guide (including Severity Ratings) and Score Sheets. Templates for the Score Sheets can be found in Appendix B. The changes for each are discussed in further detail below.

User-Centered Hybrid Assessment Tool (*User-CHAT*)

— Quick Reference —

Pre-Evaluation Assumptions

- Manufacturer has provided the **Benchmark Tasks** for the evaluation.
    - Benchmark tasks *represent the capabilities and functions of the system* (e.g., displaying weather, traffic, navigational aids, formatting range, overlaying different types of information, etc.). More emphasis should be placed on "operation critical" tasks (i.e., tasks that are performed in flight).

- Manufacturer has defined and provided a primary and a secondary "**Gold Standard**" for each benchmark task (where appropriate).
    - A gold standard represents *the least number of input actions required to successfully complete the benchmark task.* Non-gold standard actions could be performed between gold standard actions, which would result in inefficient performance.
    - The primary gold standard is the task path that the manufacturer anticipates will be used by the majority of the targeted users. The secondary gold standard is an alternate task path that the manufacturer anticipates some users will follow.
    - The gold standard task path does not take into account short-cuts or hidden functions designed for power users. Rather, gold standards should be established with the novice device user in mind, one who must adopt a label-following strategy to complete tasks.

- Manufacturers do not need to include all steps in the gold standard for tasks that have repetitive actions (e.g., entering a waypoint identifier).

- Manufacturer has supplied correct interpretations for all symbology and other displayed information.

Evaluator Roles

The User-CHAT requires at least 2 evaluators, one "User" and one "Observer". If a 3rd person is available, he/she fills the role of "Monitor":

- The **"User"**
    - Completes a series of benchmark tasks.
    - Assumes the role of a pilot operating the candidate system in the cockpit of an aircraft under multiple flight conditions.
    - Verbalizes all thought processes and actions when completing the benchmark tasks.
        - *NOTE: Task completion speed is not of primary interest.*
    - Interprets all displayed information relevant to completion of the benchmark task.

- The **"Observer"**
    - Records and compares user performance to the gold standard(s) on **Benchmark Tasks Score Sheet** (see below).
    - Documents the **1st Inefficient Action** (i.e., input action initiated instead of the gold standard action) when user performance deviates from the gold standard.
    - Tallies or records each subsequent inefficient action until User initiates the next required gold standard action required to complete the task.
    - Marks a "T" for instances when User was "thinking" about the next step toward task completion – no action was taken in 10 sec.
    - Offers hints as to the next correct input action when User is sufficiently frustrated.

  - o Documents User's interpretation of the displayed information at the end of the score sheet.
  - o Record any personal observations and/or other dialogue offered by User.

- The "**Monitor**"
  - o Also takes notes on User performance, focusing more on User comments about and reactions to the system.
  - o Maintain a dialog with representatives from the manufacturer who may be observing the session (answer questions, relay information, etc.)

User-CHAT Procedure

Phase 1 – Unstructured Exploration

- User freely explores the candidate avionics system for 10-15 min.
- User practices "thinking aloud."
- Observer records any initial personal impressions or comments about the system made by User during unstructured exploration.

Phase 2 – Structured Exploration

- User reads the benchmark task aloud and then verbalizes his/her thought process while completing the task.
  - o ***Note:*** it is extremely important that the User verbalizes all thoughts and actions so that the Observer may record them on the score sheet for data analysis purposes. The Observer has the responsibility and the obligation to pause the User's performance to ask for clarification or to catch up with data recording. Because of the importance of the Observer being able to correctly record performance, the loss of task completion momentum that occurs due to these interruptions is considered a justifiable consequence.

During task completion, the Observer:

- Compares User's performance to gold standard and tallies or records first inefficient action and subsequent inefficient actions.
  - o Identification of specific gold standard: If the benchmark task has both a primary and secondary gold standard, the Observer records the first action the User initiated and answers the "if-then" statement at the beginning of the score sheet packet to determine which gold standard action sequence the User appears to be following.
  - o If the first action does not match a gold standard: If the first inefficient action does not match a gold standard, the Observer will record user performance in its entirety on the first page of the packet and then, after task completion, discuss with the User why the alternate task completion was chosen over either one of the gold standard task paths.
- Places a "T" each time User spent thinking about the next input action to perform.
- Records any notes regarding User performance.

Upon completion of each benchmark task, both User and Observer:

- Review documented User performance, identify where performance deviated from gold standard, and discuss best answer to three questions:
  - o *Why wasn't the correct input action (i.e., gold standard) selected?*
  - o *Why was the incorrect input action selected instead of the gold standard action?*
  - o *What led to the recovery from the inefficient action (if, in fact, a recovery was made)?*
- Identify instances where displayed information was misinterpreted; then discuss and document *why* the information was misinterpreted.
- Assign a **Severity Rating.**
  - o When assigning a severity rating to each usability problem identified, the User and Observer should also account for how easily recoverable it is and what type of consequence that usability problem has for flight safety. For example, the consequence of the pilot entering the wrong

waypoint during preflight planning is vastly different than the consequence of the pilot's inability to quickly locate the "direct to" option during a flight emergency). In other words, severity ratings should indicate the context or impact of the usability problem and whether or not the usability problem was easy, difficult, or nearly impossible to recover from.

  - Serious = issues that greatly hinder performance, continue to cause problems after initial experience, and have a high potential to impact flight safety. These issues must be resolved because they are almost impossible for the pilot to recover from and thereby constitute a safety concern.
  - Intermediate = issues that also hinder performance but can be overcome with experience. It is difficult for the pilot to recover from these usability problems thus they may have safety concerns but do not warrant a serious rating.
  - Minor = issue that do not hinder performance and do not compromise safety but are recommendations for system improvement. Pilots can either easily recover from these usability problems or the inability to recover has no consequence to flight safety.

- Discuss if **Memory** influenced performance (i.e., 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). Place an "M" in the Memory column if so.
- Using a list of "**Heuristics**" (general usability issues) and their definitions, for each usability problem identified, discuss and record the display design heuristics that were violated.

(Decision Tree for Benchmark Tasks with Multiple Gold Standards)

Task Description

Task #_____: Access and view tower and ground frequencies for the Wheeler Downtown Airport in Kansas City (KMKC) using the Joystick only. Then return to the VFR map.

First User Action: FPL

Documenting user's first action and decision tree to decide which gold standard score sheet to follow.

- If Action = **Joystick**, then go to Gold Standard Packet #1
- If Action = **FPL**, then go to Gold Standard Packet #2
- If Action <> Joystick or FPL, then record below:

[illegible]	[illegible]

Score sheet if neither the primary or secondary gold standard was followed.

Why was the alternative task path chosen instead of either gold standard task path?

(Gold Standard Packet #1 – Page 1)

First User Action = Joystick

Gold Standard (#1)	First Inefficient Action	Subsequent Inefficient Action(s)	[illegible]	Notes
Joystick				
MORE INFO				
NEXT				
MAP				
RESET STICK				

Placeholders for documenting first inefficient actions

More space to accommodate Observers who want to record every subsequent inefficient action.

More space for notes.

Placeholder for recording head-down time

(Gold Standard Packet #2 – Page 1)
First User Action = FPL

Gold Standard (#2)	[illegible] Inefficient Action	Subsequent Inefficient Actions	[illegible]	Notes
FPL				
USE STICK	Moved JOYSTICK	IIIIII	T T	There should be a better indication that USE STICK must be pressed before the JOYSTICK can be used to move the cursor in the FLP map view.
JOYSTICK				
MORE INFO				
NEXT	MORE INFO	USE STICK, MAP, JOYSTICK	T	
MAP				
RESET STICK				

The secondary gold standard was used

Notes documented

Subsequent Inefficient Actions – Tallied or Specific Actions Indicated

Indications that gold standard actions have been followed

"Head-down time" indicated

(Gold Standard Packet #2 – Page 2)

After task is complete….

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
FPL						
USE STICK	Did not realize that USE STICK must be pressed before moving JOYSTICK	USE STICK is not intuitive		M	S	DL
JOYSTICK						
MORE INFO						
NEXT	MORE INFO should display more information	NEXT is not descriptive enough	Realized that after pressing MORE INFO that more information was not being displayed and knew they were in the wrong place.		I	DL
MAP						
RESET STICK						

Descriptions for why *efficient* action was not selected and why *inefficient* action was initiated

Memory would be needed to complete task in the future

Severity Rating and Heuristics Violated

Data Synthesis & Analysis

The following are high-level suggestion about how to synthesis and analysis usability data gathered from multiple User-CHAT evaluation sessions.

- *If multiple evaluation sessions*:
    - For each Gold Standard step in the benchmark task, aggregate usability data into one master list of:
        - First inefficient actions
        - Subsequent inefficient actions
        - Head-down time
        - Explanations for why Gold Standard action was not chosen and why inefficient action was chosen instead
        - Severity ratings
        - Importance of memory to complete task in future
        - Display heuristics that were violated
        - Other notes, comments, observations, etc.
        - User and Observer should review master list of usability issues to determine on a consensus basis which ones are similar across evaluation sessions and which ones are unique.

Testimonials

The following are specific comments made by the Wichita ACO during his interaction with the User-CHAT tool.

- "I think it has a lot of value to it. It's one of the best things I've seen come out"
- Likes that the User-CHAT helps identify where in the action sequence the errors occur.
- Thought the tool works well and likes the structure in the tool.
- Thought that the tool caught and tracked errors very well
- Sometimes the "observer" needs to stop the pilot to find out what is going on even though the "observer" may not want to interrupt the task but he/she may miss things if they don't.
    - This is important for the "observer" to make sure he/she understands what the pilot is doing/thinking.
- Thinks it is reasonable to ask the manufacturer to provide the gold standards for each benchmark task
    - There should be at least one task for every dedicated function key.
        - The manufacturer should provide tasks that exercise the functionality of their system.
    - Suggests that for each task, the manufacturer should provide the most likely gold standard and at least one alternative gold standard.
    - To go beyond two gold standards for each task may be asking too much.
- Said that the FAA is more concerned with the novice user than the power user.
    - Primary goal to see if the "menu driven" gold standard (labeling following strategy) gets the pilot to the information without a lot of hiccups.
    - FAA acknowledges designing for the power user but the FAA wants to see how the untrained (not very familiar with the system) can use the system and work through the system to get to the targeted information without too many problems.

Appendix A: 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** ○ 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) ○ 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** ○ 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. ○ 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. ○ 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 **functions** 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 <u>NOT</u> 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 B: Templates for Benchmark Task Score Sheets

(Decision Tree for Benchmark Tasks with Multiple Gold Standards)

Task #_____: "*Description of the task.*"

First User Action: ______________________________

- If Action = **"Gold Standard #1, Action #1"**, then go to *Gold Standard Packet #1*
- If Action = **"Gold Standard #2, Action #2"**, then go to *Gold Standard Packet #2*
- If Action <> **"Gold Standard #1 or #2"**, then record below:

Actions & Head-down Time	[illegible]

Why was the alternative task path chosen instead of either gold standard task path?

(Gold Standard Packet #1 – Page 1)

First User Action = “Gold Standard #1, Action #1)

Gold Standard #1	Same Indicator Y/N?	Subsequent Indicator Actions	[illegible]	Notes
Action #1				
Action #2				
Action #3				
Action #4				
Action #5				

(Gold Standard Packet #2 – Page 1)

First User Action = Gold Standard #2, Action #1

Gold Standard [illegible]	[illegible]	[illegible] Actions	[illegible]	Notes
Action #1				
Action #2				
Action #3				
Action #4				
Action #5				
Action #6				
Action #7				

(Gold Standard Packet #2 – Page 2)

After task is complete….

-- Similar score sheet can be used for Gold Standard Packet #1 --

Gold Standard #2	Why did you choose that action?	Why might you change the Gold Standard [illegible]	How might you discover [illegible]	[illegible]	Severity	[illegible]
Action #1						
Action #2						
Action #3						
Action #4						
Action #5						
Action #6						
Action #7						