Multiple Expert Evaluations of a PC Computer-Based Aviation Flight Training Device

Jeffrey S. Forrest
School of Computer and Information Sciences
Nova Southeastern University

ABSTRACT

The usability of a personal computer based aviation-training device (PCATD) was investigated by conducting multiple expert evaluations. One group of experts performed a heuristic evaluation of the PCATD system. Experts in a second group evaluated the PCATD by conducting a cognitive-walkthrough analysis. An ethnographic analysis was also carried out by directly observing and interviewing the participating experts during the evaluations. Experts evaluated the usability of the PCATD as applied to various practical test standards used for instrument flight training. Strong consensus by the experts in both groups indicated that the PCATD was usable for fundamental flight training as required by the Federal Aviation Administration's Instrument Rating curriculum. Issues concerning various PCATD simulation fidelities and related inconsistencies in interface design were discovered. These issues caused concern over using the PCATD for training that could be applied to actual flight time.

STATEMENT OF THE PROBLEM

Aviation flight training and pilot certification within the US is administered by the US Department of Transportation's Federal Aviation Administration (FAA). In 1994, the FAA began to consider affordable innovations that might enhance the improvement of pilot performance (Beringer, 1996). The FAA focused on the use of off-the-shelf (OTS), flight training simulations that could be supported on personal computers (PCs). By 1997, the FAA had published its guidelines for approving personal computer-based aviation training devices (PCATD) for use in flight training ("Qualification and," 1997). At the time of this study, there were at least four commercial entities offering off-the-shelf PCATDs approved by the FAA (Chamberlain, 1998). Little is known about the usability characteristics for any of the currently approved PCATD systems.

The focus of this project was to conduct multiple expert usability evaluations of one selected PCATD system. Evaluations included expert usability (heuristic), cognitive walkthrough, and ethnographic analysis as applied to specific FAA training guidelines conducted on the PCATD. The identification and application of these techniques are discussed subsequently in this study.

Evaluation Goals

The first goal for this evaluation was to uncover new knowledge regarding the usability of PCATD systems by FAA Certified Flight Instructors (CFIs). CFIs selected as participating experts were highly experienced in the utilization of computer generated flight simulation. The information gained from this project was also used to

make recommendations toward the improvement of interface design, and application of the PCATD as a flight-training tool.

LITERATURE REVIEW

Historical Overview of Flight Training Device Evaluation and Related Theory

The birth of the modern flight-training simulator is often attributed to Ed Link who created the "Link Trainer" in 1929 (Gunston, Pyle, & Chemel, 1992). The Link Trainer was described as a ground-based device that pilots could use to learn the basic skills needed to fly before leaving the ground (Gunston et al., 1992). Link trainers were designed to simulate the use of flight instruments typical of aircraft being produced during that time.

The realism or "fidelity" (Caro, 1988) of the Link Trainer was very low as compared to the flight simulator produced today. Simulation fidelity has been identified as a two dimensional measurement of the realism associated with physical and functional characteristics (Hays & Singer, 1989). The ability of a simulator to accurately represent the visual, spatial or kinesthetic characteristics of the flight environment is known as physical fidelity (Hays & Singer). Hays and Singer contrast the informational, or stimulus and response characteristics, as the functional fidelity of the simulation. Physical characteristics of the early Link Trainer emphasized attributes such as the location of flight controls and limited visual (spatial) training. Later models of the Link Trainer began to incorporate more accurate representations of information displayed on instruments in response to the pilot's actions in a training situation.

Since the advent of the Link Trainer, flight simulators have evolved to a state of technology that can completely duplicate the flight environment for a specific aircraft. It is now possible to competently train flight crewmembers (pilots) to fly a specific type of aircraft without ever using the actual aircraft as a part of the training program.[1] This current level of high fidelity flight simulation was developed from a need to train crewmembers to perform tasks not previously possible or to a skill level previously unattainable (Caro, 1988). The evolution of flight simulation technology was motivated and built upon learning theories advocated by cognitive scientists such as Charles Osgood and Edward Thorndike (Caro, 1988). These theories stipulated that successful learning from simulation requires that the simulation have a one-to-one relationship to reality (Caro, 1988). This approach to simulation design seeks high levels of fidelity as the characteristic that will foster successful learning. As physical and informational simulation reaches reality, learning will be more effective. For this reason, modern simulators have reached a level of high physical and informational reality. Bill Siuru and John Busick (1994) describe today's flight simulation as computer facilitated "virtual reality." They describe virtual reality as multi-sensory flight simulation that provides three dimensional sight and sound along with feedback for touch and motion (Siuru & Busick).

Since the late 1960s, the effectiveness of high fidelity in flight simulation has been questioned by several researchers (Macfarlane, 1997; Caro, 1988; Prophet & Boyd, 1970; Grimsley, 1969). Macfarlane (1997) stated, "…the evolution of flight simulation, as a realistic representation of flight parameters, has often overshadowed the practical value of simulators and led to a number of false assumptions about their training value" (p. 59). According to Macfarlane (1997), simulation fidelity should be evaluated in

terms of "task fidelity" and "instructional fidelity" (p. 63). He defines task fidelity as the degree to which simulation is able to recreate the actual parameters of a mission, in terms of training and practice. Instructional fidelity is defined by Macfarlane as the effectiveness of the simulation, as part of an instructional system, to transfer knowledge to the training crewmembers. Macfarlane's taxonomy for fidelity does not feature the importance of physical and knowledge realities as was stressed earlier for successful simulation design. Instead, proper simulation design is based upon first asking what it is to be accomplished, then designing or selecting the simulation that best meets that need.

Macfarlane (1997) further emphasizes this strategy by stating "Simulation should not be undertaken for simulation's sake but rather for some predetermined purpose…." (p. 73). Proper instructional design and instructional systems development are essential to Macfarlane's philosophy of simulation as applied to training. Simulations should be used to support the instructional design and the related systems necessary to meet the goals of the learning objective(s). Evaluation of simulator effectiveness should focus on the relationship between desired learning objectives and the simulation fidelity required to meet those specific objectives. High fidelity as a characteristic of simulation design does not insure effective crewmember training.

Paul Caro (1998) also supports Macfarlane's reasoning by identifying the design characteristics that support effective, low fidelity flight simulators. Caro stated that fidelity should be designed around the elements of cues, discriminations, mediation, and generalizations. As an example of these criteria, consider a low fidelity computer based training (CBT) simulator. Assume that the example CBT unit has a standard computer monitor, keyboard, and mouse. The graphical user interface (GUI) depicted on the monitor only shows a few elements of the actual flight environment. According to Caro, cues are meanings assigned by the pilot to stimulus represented on the GUI. If the simulation environment offers the pilot an opportunity to learn and assign the correct meaning to the stimulus provided, then effective simulation has taken place without the need for high fidelity. Discrimination is the ability of the pilot to differentiate between various stimuli, and assign the proper meaning to each recognized stimulus. The CBT simulation need not offer realistic physical cues in order for the pilot to properly discriminate between various stimuli. As an example, the pilot could learn to discriminate and assign meaning to stimuli solely from on-screen text descriptions or audio explanations. Caro refers to simulation design elements that foster discrimination, such as on screen text descriptions, as mediations. Mediations also include generalizations, which are low fidelity representations that allow a transfer of knowledge to occur. Generalizations are elements of low fidelity that are used in simulation when the pilot already has knowledge of the element being represented by the generalization. For example, it may not be necessary to simulate the ability, or fidelity, to adjust a flight instrument if that pilot is already aware of how to adjust and use the flight instrument.

Low Fidelity Flight Training Simulation

The value and use of simulation as a training device in aviation has been well documented over the past 30 years (Beringer, 1996,). Over this period, the traditional emphasis of designing flight-training simulators with high fidelity characteristics has significantly increased the cost of aviation simulation devices. This expense has created an industry demand for lower cost, OTS low fidelity training devices (Wilson, 1998). The advent of the

personal computer (PC) has facilitated the design and implementation of lower cost, low fidelity training devices. PC-CBT devices that properly match fidelity with learning objectives are now in demand by commercial, military, and civilian aviation training facilities (Sutton, 1998).

Within the US, simulators must be approved by the FAA for use in FAA required pilot or crew training programs. The FAA's responsibility regarding PC based flight simulation is to certify that the level of fidelity is compatible with the learning objectives associated with specific FAA training objectives. In this way, the FAA "qualifies" the fidelity of the simulation and "approves" the use of the simulator for specific training curriculum (Chamberlain, 1998).

In 1995, the FAA began to approve and qualify low fidelity PC Based Aviation Training Devices (PCATD). The primary motivation for the FAA's support of PCATD was to potentially reduce the overall cost of flight training to the industry and improve pilot procedural training as related to specific FAA training guidelines (Beringer, 1996). The FAA's approval of PCATD applies to specific primary instrument training guidelines published by the FAA (Federal Aviation Administration, 1997).

Little is known about the usability of low fidelity PCATDs as applied to flight training required for the Instrument Rating. A study conducted by Dennis Beringer (1996) compared a PCATD to alternate forms of FAA approved training. In this study, Beringer (p. 11) found that the examined PCATD had "…sufficient task fidelity to motivate generalizable behavior, producing outcomes that are comparable to those obtained in other simulation devices, in fact, aircraft." Beringer's study also incorporated a component of evaluation similar to a cognitive walkthrough. A cognitive walkthrough has been described by Miller and Jeffries (1992) as an evaluation that compares the ability of the interface to the user goals and expectations. In Beringer's (1996) study, the users (pilots) were asked to compare the PCATD fidelity to the "real world" aircraft. Overall, the users found the PCATD more sensitive than the actual aircraft and harder to fly (Beringer, 1996).

A more recent study conducted by Taylor, Lintern, Hulin, Talleur, Emanuel, and Phillips (1997) measured the effectiveness of PCATD training as compared to actual flight training. This study did not specifically evaluate the usability of the PCATD in the training environment. Methodology focused on the comparison of user performance indexes and FAA published guidelines using both the PCATD and actual aircraft training environments.

Simulation Evaluation

Shneiderman (1998) has stated that the primary goal for usability evaluations "…is to force as much possible of the evolutionary development into the prerelease phase, when change is relatively easy and inexpensive to accomplish" (p. 144). This philosophy for evaluation applies during the design, or formative stage of system development. Wilson (1998) describes how current CBT aviation simulation is designed with little opportunity for formative evaluation. Instead, most low fidelity aviation CBT simulators are being offered as a "proof of concept" product, whereby evaluation is primarily summative in the form of end-user feedback (p. 28). Literature offered by Beringer (1996) and Taylor et al. (1997) seems to also support this conclusion in regards to the PCATD. Emphasis on the evaluation of the PCATD has been focused on the transfer of learning as a proof of concept, rather than the evaluation of effective PCATD design.

METHODOLOGY

Evaluation Methods

The purpose of this evaluation was to assess the usability of a PCATD as applied to selected FAA Instrument Rating training guidelines. The evaluation concluded with recommendations on the improvement for PCATD interface design and application within the aviation-training environment.

The proposed PCATD evaluation was summative and conducted within an actual training environment. Shneiderman (1998) suggests that expert reviews be conducted as summative evaluations. He offers several models for expert evaluations that are particularly viable for the PCATD. First, Shneiderman suggests the heuristic evaluation as a method to "…critique an interface to determine conformance with a short list of design heuristics" (p. 126). In this study, the design heuristics that will be used are the "eight golden rules" for interface design as also suggested by Shneiderman (1998). Shneiderman's design rules will also be supplemented with criterion for "checklist evaluations" as provided by Ravden and Johnson (1989). Miller and Jeffries (1992) found that heuristic evaluations are very successful for discovering most of the major problems inherent to the design of a user interface. The heuristic evaluation was conducted to provide evidence for specific improvements in the PCATD interface design.

Cognitive walkthroughs are also suggested by Shneiderman (1998) and Wharton, Bradford, Jeffries and Franzke (1992) for evaluating the interface while conducting a specific task. This evaluation required experts to conduct training task as defined by FAA approved guidelines. Cognitive walkthroughs are based upon the evaluation theory of "learning by doing" and focus on basic usability principles (Wharton et al., 1992). Miller and Jeffries (1992) compared the advantages of various structured evaluation processes. They determined that cognitive walkthroughs are well suited for discovering problems with the interface as related to the user goals and assumptions. Therefore, the cognitive walkthrough should provide data leading to an assessment of the cues, discriminations, mediations, and generalizations (Caro, 1998) that will be experienced by the user of the PCATD interface.

It has been recommended that in addition to structured evaluations, the potential affects of culture or the social situation should also be factored into the evaluation (Sommerville, Bentley, Rodden, & Sawyer, 1994). Sommerville, et al (1994), and Shneiderman (1998) suggest using ethnographic observation as a complement to other forms of evaluation. An ethnographer for this evaluation was present for both the heuristic and cognitive walkthrough evaluations. Ethnographic observations and interpretations were made in order to help determine the factors not inherent to the PCATD that influence the evaluations conducted in the heuristic and cognitive walkthrough evaluations.

Subjects

Two groups consisting of three FAA Certified Flight Instructors (CFIs) were asked to participate in the evaluation. One group conducted the heuristic evaluation, while the other implemented the cognitive walkthrough. Experts were solicited from a population of CFIs having over ten years experience in the application and evaluation of flight training simulators. Miller and Jeffries (1992) found that as the relative expertise of evaluators increases, the fewer the number of experts that are required for the evaluation. In this evaluation, the same number of experts participated in both the heuristic and cognitive walkthrough

evaluations as was used in previous successful studies conducted by Miller and Jeffries (1992).

Setting

The PCATD evaluations were conducted within the Aerospace Science Department (ASD) of a midwestern college. The CFIs participating in the evaluations were currently employed as faculty members of the ASD. The heuristic, cognitive walkthrough and ethnographic evaluations were conducted within the aviation simulation lab that is used by the ASD to train pilots.

Apparatus

The selected PCATD evaluated was full functioning, commercially available, and FAA approved. The PCATD simulates the flight environment for a single engine aircraft used for primary instrument flight training. The ASD is certified by the FAA to administer approved instrument flight training using the selected PCATD.

Procedure

All CFIs employed by the ASD were invited to volunteer as expert evaluators. Each participating CFI was professionally trained in human factors analysis and simulation based training. Under these circumstances, expert evaluations can be conducted within one to two days (Dumas & Redish, 1993). Each CFI was given thirty minutes to conduct their heuristic or cognitive walkthrough following a specified FAA training standard.

The FAA training standards followed were specified in the FAA's publication Instrument Rating for Airplane, Helicopter, and Airship Practical Test Standards ("Instrument Rating for," 1994) (PTS). The three CFIs conducting the heuristic evaluation were asked to select any three tasks referred to as "areas of operation" defined by the FAA's PTS. Each expert then used the PCATD to apply the three chosen operational areas of operation in any manner they deemed suitable to primary instrument instruction.

The PTS areas of operation were considered adequate for evaluation. Each area can be quickly evaluated, is stated in the user's words, provides enough information to complete the task, and is linked directly to the goals of the proposed evaluation (Dumas & Redish, 1993). As the heuristic evaluators explored their selected areas of operation, they were asked to write comments on a survey addressing "areas of concern" (Dumas & Redish, 1993). These areas of concern were related to the "eight golden rules" for interface design as suggested by Shneiderman (1998).

The CFIs participating in the cognitive walkthrough were asked to "practice" three pre-identified areas of operation contained within the PTS [2] using the PCATD. They then answered a post-task survey qualifying the PTS operations in relation to the cues, discriminations, mediation, and generalizations (Caro, 1998) inherent to the PCATD interface (see Appendix B).

A single ethnographer was also present for each of the expert evaluations. The ethnographer was an Instrument Ground Instructor with over ten years experience in flight simulator training and human factors associated with student interaction and CBT. The ethnographer observed each expert evaluator in both groups. Ethnographic examination uncovered how the common cultural values of the experts' influence their perception on the characteristics inherent to the PCATD. The primary goal of the ethnography was to detect the affect of culture on the perception, or experience [3] of using the PCATD (Gall, Borg, & Gall, 1996). As suggested by Sommerville, Bentley, Rodden and Sawyer (1994, p. 358),

no specific set of instructions were provided to the experts concerning the ethnographic observation. However, the experts were encouraged to "think out loud" and discuss any aspect of the PCATD with the ethnographer. This strategy was successfully used by Karat, Campbell, and Fiegel (1992) in a study comparing techniques in user interface evaluation.

Analysis of the Data

Qualitative analysis was applied to the results obtained from the heuristic and ethnographic evaluations (see Appendix A). Specifically, a "hermeneutic circle" (Gall et al, 1996, p. 706) analysis was applied to the concerns and issues raised by each evaluator of the PCATD. In this analysis, meaning was interpreted from the concerns or comments made by each evaluator. Meaning was also applied to the overall concerns made by each evaluator and taken as a whole (Gall et al, 1996). Conclusions and recommendations were made regarding the usability of the PCATD, as based upon the analysis. Recommendations for improving the usability of the PCATD were also made.

A questionnaire measuring each evaluator's attitude regarding aspects of the PCATD usability was provided to each member of the cognitive walkthrough (see Appendix B). The questionnaire contained an ordinal scale measuring ten (10) levels of agreement for each area of concern (Gall et al, 1996). The questionnaire was pre-tested for clarity and understanding by various CFIs within the ASD. The experts used within the evaluation did not represent a normal population. Therefore, a Kruskal-Wallis ANOVA analysis was applied to the ordinal results provided by each expert (Gall et al, 1996, p. 297). Kruskal-Wallis ANOVA analysis provided quantitative results measuring the level of agreement between each expert's cognitive walkthrough evaluation. Qualitative conclusions were made based upon the quantitative analysis. Recommendations were made regarding the usability of the PCATD, as based upon the cognitive walkthrough analysis. Recommendations for improving the usability of the PCATD were also made based upon a synthesis of all evaluations and analyses.

RESULTS

Results of the Heuristic Evaluation

Three CFIs participated in the heuristic evaluation of the PCATD. Each CFI was given approximately 30 minutes to conduct their evaluation. Each heuristic evaluator wrote comments regarding "areas of concern" as they used the PCATD to explore their selected areas of operation (see Appendix A). These comments were qualitatively evaluated and related to Shneiderman's "eight golden rules" (1998).

Issues of simulator fidelity were characterized in terms of cues, discrimination, and mediation (Caro, 1998). The following results provide each question asked on the survey along with a qualitative analysis of the comments made by all three evaluators. Relevant ethnographic analysis is also included for each question.

Heuristic and Ethnographic Results

What are your concerns regarding the clarity of objects, or information, displayed on the PCATD screen and control system?

Two of the three evaluators remarked that the icons presented in the PCATD user interface were "too small" and depicted images that were "unknown" in terms of implied meaning or utility. The third evaluator felt that all of the PCATD interface elements presented were "clear and easy to identify."

According to Shneiderman (1998), shortcuts such as icons are "appreciated by knowledgeable and frequent users" (p. 74). Although knowledgeable, the experts were not frequent users of the specific PCATD being evaluated. Evaluators expressing concern regarding the ambiguity of the icons felt that an adequate solution would be to place short, abbreviated textual descriptions under each icon. This feature offered as a user option would consider the experience level of the user, as suggested by Shneiderman (1998) when considering combining text with icon representations.

Shneiderman (1998) also suggests that each icon should be designed in a "familiar and recognizable manner." Since the PCATD technology is relatively new, it was difficult for the evaluators to relate the icons presented to any pre-existing CBT interface designs. The icons represented in the evaluated PCATD may become familiar and recognizable standards in future PCATD systems. According to Caro (1998), the addition of textual descriptions would enhance the quality mediation as supported by the PCATD interface.

Ethnographic observation revealed that one of the evaluators had limited prior experience in viewing the user-interface for the evaluated PCATD. This probably accounts for his characterization of each element being clear and easy to identify. However, upon further questioning by the ethnographer, the evaluator agreed that the novice user would benefit from textural descriptions related to each specific icon.

What are your concerns regarding the compatibility of objects, or information displayed by the PCATD, to similar attributes as experienced in actual flight?

All three evaluators agreed that simulation fidelity of the PCATD as related to the actual flight environment was quite good. Ethnographic evaluation determined that the evaluators found certain flight maneuvers as "jerky" and "too rapid" in response fidelity. As suggested by Caro (1998), the cues provide by these unrealistic fidelities might deter the student pilot from learning the correct meaning of the stimulus being provided by the PCATD. All three evaluators felt that the cues provided during these maneuvers would be of minor concern to the novice student. They also felt that although fidelities in these maneuvers were not realistic, the actual outcome of the simulation was accurate enough to provide proper understanding of the learning objective by the student pilot.

What are your concerns regarding the consistency of PCATD performance and display as applied to each PTS area of operation that was conducted?

The cues, discriminations, and mediations provided by the PCATD while simulating flight were considered adequate for successfully conducting all but one area of operation contained within the PTS. It was determined by two of the evaluators that the PCATD did not allow the student to perform a flight maneuver referred to as a "stall." This deficiency in simulation fidelity was considered by all three evaluators as a serious issue requiring attention in software redesign and upgrade by the manufacture of the PCATD. Cues (Caro, 1998) provided during the stall maneuver were considered accurate. However, the simulation was not able to provide the correct "feedback" (Shneiderman, 1998) of the instance in time that the actual aerodynamic effect of the stall occurred.

What are your concerns regarding the ease of operating the PCATD?

Ethnographic observation determined that all three evaluators viewed the PCATD as "relatively easy to use." It was generally agreed that students having a very basic understanding of the personal computer would find the PCATD very easy to use. The rule of "consistency" (Shneiderman, 1998) as compared to other PC-based software was considered very strong as applied to the overall PCATD design.

However, one of the evaluators determined that it was not possible to "multi-task," or switch to other software applications while using the PCATD. Shneiderman (1998) suggests that design elements that cause a loss of user control can build anxiety and dissatisfaction. All three evaluators agreed that this design flaw would cause potential aggravation for the instructor using the PCATD in a training environment. It was felt that the lack of multi-tasking would have minimum impact on the student's ability to use the PCATD.

What were the best aspects of the PCATD for the student pilot as a user?

All three evaluators felt that the overall fidelity of the PCATD provided a positive experience for learning and building competency in the skill required by the FAA's PTS. The ability to repeat flight-training exercises in the level of fidelity offered by the PCATD was considered its strongest attribute.

What were the worst aspects of the PCATD for the student pilot as a user?

Two of the three evaluators found the design and fidelity of the flight control hardware unsatisfactory. Confusion was observed when all three evaluators attempted to manually adjust the radio frequencies required to operate the instruments being displayed by the simulation. The ethnographer noted that negative comments were made regarding the cues, discriminations, and mediations offered by the radio hardware interface. The evaluators felt that this design would cause the students confusion over simulation consistency (Shneiderman, 1998) as compared to the actual operation of aircraft radios in the flight environment.

Is there anything else about the PCATD you would like to add?

Two of the three evaluators added comments to this question. One evaluator suggested that an additional display containing "approach chart" information be added to the screen. Approach charts are used by pilots during the arrival and landing phase of flight. The evaluator stated that this feature would simulate fidelity comparable to various flight information systems used in the flight deck of an actual aircraft. This

suggestion would potentially reduce the memory load on the user while improving the capability to assimilate information (Shneiderman, 1998) while using the PCATD.

Of particular interest were the comments made by the second evaluator responding to this question. This evaluator felt that the PCATD offered excellent fidelities for practice and instructor lead demonstrations. However, he did not believe that the fidelities for the hardware (manual controls) were sufficient to use the PCATD as training device that would meet certain experience requirements for FAA pilot certification. He stated that the PCATD was an excellent classroom-training device, but should not be used to replace any of the FAA's regulatory flight experience.

Further questioning of this evaluator determined that he had extensive experience with flight training simulators that offered the highest state-of-the-art fidelity. The prior experiences of this evaluator in regards to very high fidelity simulation technology might have biased his judgement against PCATD as sufficient in meeting the regulatory requirements of the FAA. It was also his opinion that the PCATD provided much more of a training process, rather than a simulation.

Results of the Cognitive Walkthrough Evaluation

Three CFIs participated in the cognitive walkthrough evaluation of the PCATD. Each CFI was given approximately 30 minutes to conduct their evaluation. All of the CFIs were of the male gender. Several female CFIs were invited to participate, but were unable to do so.

Each evaluator participating in the cognitive walkthrough was asked to "practice" three pre-identified areas of operation contained within the PTS [2] using the PCATD. They then answered a post-task survey qualifying the PTS operations in relation to the cues, discriminations, mediation, and generalizations (Caro, 1998) inherent to the PCATD interface (see Appendix B). Ordinal responses measured the level of agreement to each statement asked. A rank of one (1) represented an attitude of complete agreement. A rank of nine (9) represented an attitude of complete disagreement. The rank of ten (10) was used to indicate that the question was not applicable to the characteristic being evaluated. A Kruskal-Wallis ANOVA analysis was also conducted on the ordinal responses submitted on the survey for the cognitive walkthrough. This analysis measured the overall level of agreement (H statistic) between each evaluator's cognitive walkthrough.

Cognitive Walkthrough Results

None of the evaluators for the cognitive walkthrough responded with a rank of ten (10, or not applicable) to any of the survey questions (see Appendix B). The least level of agreement indicated from one of the evaluators was a four (4). This rank was assigned to the question, "The objects, or information, displayed on the PCATD screen are identifiable to those same elements as experienced in the actual flight environment" (see Appendix B). Rank level responses for the evaluation ranged from one (1) to four (4). The overall average (x) rank level of agreement to all questions by all evaluators was $x = 1.8$. Table 1 summarizes the average rank level (x_q) response for each of the questions administered in the cognitive walkthrough evaluation.

The Kruskal-Wallis ANOVA produced a relatively small test statistic ($H = 0.522$ with 2 degrees of freedom) indicating a strong level of agreement for the rank level responses provided by each evaluator of the cognitive walkthrough. It was further

assumed that a significant difference between the responses would be considered to exist if the Kruskal-Wallis probability test was $p<0.05$. Kruskal-Wallis analysis on the rank responses produced a probability of $p=0.77$, indicating no statistically significant difference between the responses made by each evaluator.

Table 1
Average Rank Level (xq) Response for each Question Administered within the Cognitive Walkthrough Evaluation

Question (Q)	x_q
(Q) 1	2
(Q) 2	1.3
(Q) 3	2.3
(Q) 4	2
(Q) 5	2
(Q) 6	1.7
(Q) 7	1.3

Note. A rank of one (1) equaled complete agreement while a rank of nine (9) equaled complete disagreement. See Appendix B section titled "Cognitive Walkthrough Evaluation – Primary Questions for Levels of Agreement" for each specific question asked during the cognitive walkthrough evaluation.

Ethnographic observations conducted during the cognitive walkthrough resulted in similar conclusions as made by the evaluators for the heuristic evaluation. The evaluators for the cognitive walkthrough felt that PCATD simulation for flights conducted at (a) "slow speeds," (b) "high pitch attitudes," and in (c) "turbulence" resulted in consistencies (Shneiderman, 1998) not found in actual flight. One evaluator remarked that the visual fidelity of the "natural horizon [earth-sky boundary] caused visual disorientation." All of the experts in this evaluation experienced the same frustration as those in the cognitive walkthrough when attempting to control and adjust the simulated radio-navigation PCATD hardware. All evaluators stated that the consistency (Shneiderman, 1998) for cues, discriminations, and mediations (Caro, 1998) was very poor in terms of radio-navigational hardware. Further questioning by the ethnographer revealed that all three evaluators questioned the decision of the FAA to certify the PCATD as a training device that can be applied to actual training flight time as required by regulation. This concern was based upon the poor fidelities associated with the hardware incorporated within the PCATD design.

CONCLUSION

Discussion

Consensus of agreement was strong among the evaluators for both the heuristic and cognitive walkthrough evaluations conducted. Heuristic and ethnographic analysis of the PCATD confirmed similar areas of concern by each evaluator regarding the usability of the simulator as a training device for primary instrument students. The relatively small value of the Kruskal-Wallis test statistic H ($H=0.522$; $p=0.733$) indicated a strong level of agreement among the evaluators participating within the cognitive walkthrough.

Ethnographic analysis determined that all experts for heuristic and cognitive walkthrough felt that the usability of the PCATD was sufficient for training primary instrument students. All evaluators expressed strong concern over the design and fidelity of the PCATD hardware used to simulate aircraft control. One evaluator from the heuristic evaluation felt that the PCATD should not have been approved by the FAA as training that applied to flight time required by regulation. Ethnographic examination during the cognitive

walkthrough discovered that all three experts had similar concerns regarding of the approval of the PCATD as an FAA approved training device.

A shortcoming to this evaluation was that all of the CFI's that volunteered to participate were of the male gender. An approximately equal number of female and male CFIs from the ASD were invited to participate in either of the evaluations. It is known that two of the ASD female CFIs had scheduling conflicts. Vardaman (1997) stated that males tend to like computers more than females. It would have been beneficial to this evaluation to incorporate the possible affect of gender and the CFI's qualification of the PCATD.

All of the experts from both evaluations felt that the PCATD was adequate for training each area of operation as described in the FAA's PTS. The only concern was the lack of consistency (Shneiderman, 1998) in fidelities for those areas of operation requiring slow flight speeds or extreme flight attitudes (position of aircraft). Results of the cognitive walkthrough evaluation support the conclusion that the PCATD is adequate for the training, and use by, student pilots pursuing primary instrument training.

Limitations of the Analysis

The primary objective of this study was to discover issues of usability regarding the PCATD design as applied to training required for the Instrument Rating. Conclusions made in this study were based upon the ethnographic observations and opinions made by expert evaluators. This study included six expert evaluators in addition to the ethnographer.

Jakob Nielson (1993) has provided evidence that on average, three to five experts offer the greatest incremental advantage for discovering issues of usability during an evaluation. However, Nielson also recommends that in the evaluation of mission critical systems, more evaluators should be used. Based upon the work by Nielson, this study recommends that future evaluations of the PCATD system should employ between seven and 15 expert evaluators. Nielson advises that on average, six evaluators will discover 80 percent of all relevant usability issues, while 15 evaluators will increase the probability to 90 percent.

A further limitation to the analysis of this evaluation was the length of time provided to conduct each evaluation. Each evaluator was provided with 30 minutes to conduct their review of the PCATD, exclusive of the time required to fill out each survey form. Nielson (1993) recommends that from one to two hours be provided for each expert to conduct their evaluation.

As a final concern, it is important to note that this study does not offer statistically significant data that can support inferential conclusions. This study did use valid methodology for exploring the issues of PCATD usability. However, it is strongly recommended that the number of expert evaluators be significantly increased for future PCATD evaluations offering conclusions supported by inferential statistics.

Recommendations

Based upon the usability investigations conducted in this study, the following recommendations are made for further investigation and potential improvement in PCATD usability and design:

1. Flight control hardware should be redesigned for consistency (Shneiderman, 1998). Particular attention should be focused on improving the cues, disseminations, and mediations (Caro, 1998) of the radio-

navigational hardware associated with the PCATD.

2. Improvements should be made in allowing the user to control the amount of interactions (Shneiderman, 1998) used to control the configuration of the PCATD simulation software. Icons, objects, and other information provided by the PCATD offered meaning that would only be understood by the experienced user of the PCATD.

3. Simulation for fidelities experienced during slow airspeeds, unusual attitudes, or turbulence require improvements in at least the visual cues (Caro, 1998) being displayed by the PCATD.

4. An option should be added allowing the PCATD user to display additional database information (such as approach charts) on a separate monitor consistent with actual flight deck configuration.

Further Recommendations for Study

Evaluators for this project expressed concern that the FAA approved the selected PCATD for use in meeting certain requirements of actual flight time required for pilot certification. This concern was based upon PCATD hardware related fidelity problems discovered in this study. Further research is recommended to determine the fidelities required for hardware design that would improve the interface consistency of the PCATD as related to the actual flight environment.

This study conducted expert and ethnographic evaluations of a selected PCATD simulator. Further research focused on the design and fidelity of the PCATD, as evaluated by the student pilot, should be considered. New efforts in research should also consider evaluating the PCATD interface during actual training conditions.

REFERENCES

Beringer, D. B. (1996, April). Use of off-the-shelf PC-based flight simulators for aviation human factors research. Washington, DC: Office of Aviation Medicine. (NTIS No. DOT/FAA/AM-96/15)

Caro, P. W. (1988). Flight training and simulation. In Weiner, E. & Nagel, D. (Eds.). Human Factors in Aviation (pp. 229-261). NY, NY: Academic Press.

Chamberlain, D. H. (1998, July/August). Qualified PC computer-based training devices take off. FAA Aviation News, 37, (5), 8-10.

Dumas, J., & Redish, J. (1993). A practical guide to usability testing. Norwood, NJ: Ablex Publishing.

Federal Aviation Administration. (1997, May 12). Qualification And Approval Of Personal Computer-Based Aviation Training Devices. (FAA Advisory Circular AC No: 61-126). Washington, D.C.: United States Department of Transportation.

Federal Aviation Administration. (1994). Instrument Rating for Airplane, Helicopter, and Airship Practical Test Standards. Newcastle, Washington: Aviation Supplies & Academics.

Gall, M., Borg, W. & Gall, J. (1996). Educational Research. NY, NY: Longman Publishers.

Grimsley, D. L. (1969). Acquisition, retention, and retraining: Group studies on using low-fidelity training devices (Tech. Rep. No. 69-4). Washington, DC: The George Washington University, Human Resources Research Office.

Gunston, B., Pyle, M., & Chemel, E. (Eds.). (1992). Chronicle of Aviation. Liberty, MO: JL International Publishing.

Hays, R., & Singer, M. (1989). Simulator fidelity in training systems de-sign: Bridging the gab between reality and training. New York, NY: Springer-Verlag.

Karat, C., Campbell, R., & Fiegel, T (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. ACM CHI'92 (pp. 397-404). Monterey: ACM.

Macfarlane, R. (1997). Simulation as an instructional procedure. In Hunt, G. (Ed.). Designing instruction for human factors training in aviation (pp. 59-93). Brookfield, USA: Avebury Aviation.

Miller, J., & Jeffries, R. (1992). Usability evaluation: science of trade-offs. IEEE Software, 9, (5). 97-102.

Nielsen, J. (1993). Usability Engineering. San Francisco, CA. Morgan Kaufmann.

O'Neil, H. F., & Baker, E. L. (1994). Technology assessment in software applications. Hillsdale, NJ: Lawrence Erlbaum.

Prophet, W. W., & Boyd, H. A. (1970). Device-task fidelity and transfer of training: Aircraft cockpit procedures training (Tech Rep. No. 70-10). Alexandria, VA: Human Resources Research Organization.

Ravden, S., & Johnson, G. (1989). Evaluating Usability of Human-Computer Interfaces. Chichester: Ellis Horwood Ltd.

Redmond-Pyle, D., & Moore, A. (1995). Graphical User Interface Design and Evaluation. NY, NY: Prentice Hall.

Siegfried, T. (1994). User interface evaluation: A structured approach. New York, NY: Plenum Press.

Shneiderman, B. (1998). Designing the User Interface. (3rd ed.). Reading, MA: Addison-Wesley.

Sommerville, I., Bentley, R., Rodden, T., & Sawyer, P. (1994). Cooperative systems design. The Computer Journal, 37, (5). 357-366.

Sutton, O. (1998, April) Training business on a roll. Interavia, 63, (619). 19-22.

Siuru, B., & Busick, J. (1994). Future Flight. Blue Ridge Summit, PA: Tab Aero.

Wharton, C., Bradford, J., Jeffries, R., & Franzke, M. (1992). Applying cognitive walkthroughs to more complex user interfaces: experiences, issues, and recommendations. ACM CHI'92 (pp. 381-388). Monterey: ACM.

Wilson, J. R. (1998, April). Wanted: Off the shelf solutions. Interavia, 63, (619). 23-28.

FOOTNOTES

[1] This is often referred to as "zero flight time" or ZFT. Under ZFT, a training program is conducted entirely within the flight simulator. The aircraft is not used until training using the simulator is complete.

[2] Area IV (A) – straight and level flight; (C) rate climbs and descents; and (F) steep turns.

[3] Gall, Borg and Gall (1996, p.608) refer to this characteristic of cultural perception as "emic" ethnography. Emic ethnography attempts to qualify the affect of culture on the human perception of reality.

APPENDIX A

Heuristic Evaluation - Primary Questions for Areas of Concern

1. What are your concerns regarding the clarity of objects, or information, displayed on the PCATD screen and control system?
2. What are your concerns regarding the compatibility of objects, or information displayed by the PCATD, to similar attributes as experienced in actual flight?
3. What are your concerns regarding the consistency of PCATD performance and display as applied to each PTS area of operation that was conducted?
4. What are your concerns regarding the ease of operating the PCATD?
5. What were the best aspects of the PCATD for the student pilot as a user?
6. What were the worst aspects of the PCATD for the student pilot as a user?
7. Is there anything else about the PCATD you would like to add?

Appendix B

Cognitive Walkthrough Evaluation – Primary Questions for Levels of Agreement

Each comment will be answered using an ordinal scale measuring levels of agreement: ex. 1 = "strongly agree," to 9 = "strongly disagree" with 10 representing not applicable (NA) (Shneiderman, 1998, p.140). After conducting the three prescribed areas of operation, the expert will answer the following questions:

1. The student pilot will be able to interpret the objects, or information, displayed on the PCATD screen.
2. The student pilot will be able to relate the objects, or information, displayed on the screen to the required knowledge areas fundamental to primary instrument flight training.
3. The objects, or information, displayed on the PCATD screen are identifiable to those same elements as experienced in the actual flight environment.
4. Adequate information is provided on the PCATD screen for the student pilot to interpret the meaning of each object or action being simulated.
5. The overall simulation of the PCATD is adequate in terms of realism as applied to primary instrument training.
6. Response of the PCATD to user control input is adequate for primary instrument training.
7. As compared to other approved flight training devices, the PCATD is acceptable for primary instrument training.