International Journal of Aviation, Aeronautics, and Aerospace

Volume 4 | Issue 2 Article 3

4-10-2017

A Usability Study for Electronic Flight Bag (EFB) Flight Planning Applications on Tablet Devices for Ab-initio Pilots

Jeff Schwartzentruber
Seneca College of Applied Arts and Technology, jeffrey.schwartzentruber@ryerson.ca

A Usability Study for Electronic Flight Bag (EFB) Flight Planning Applications on Tablet Devices for Ab-initio Pilots

Cover Page Footnote
The current study was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC), in collaboration with the Office of Applied Research and Innovation (ARI) at Seneca College. The author would like to acknowledge the support from the Seneca College School of Aviation.

Advancements in mobile technology have led to the development of electronic flight bags (EFBs). The Federal Aviation Administration (FAA) defines EFBs as an electronic display system intended for flight deck or cabin crew use, which is capable of displaying a variety of aviation data and performing basic flight calculations (FAA, 2014). Increasing affordability and availability of mobile devices are making EFB applications more popular during ab-initio training. The use of digital technology in aviation is rapidly evolving. Beyond EFBs, digital technologies are being incorporated in all sectors of aviation, such as maintenance (Pourcho, 2014), air traffic control (Croft, 2015), avionics (C.-J. Chen, Yang, & Chang, 2014; Janakiraman & Nielsen, 2016), flight safety (Ghorbani, Khatibi, FazeliFard, Naghipour, & Makarynskyy, 2016), and air operations (Williams, 2014). The increasing presence of technology in the aviation industry translates to increased pressure on aviation training institutions to prepare their students for the new era of digital aviation. As a result, training institutions/personnel are beginning to incorporate EFBs into their training curriculum as a complement to traditional methods. EFBs offer many advantages (i.e. paperless cockpit, automatic updates, etc.). However, handling an EFB as a novice pilot can be a daunting task, adding distraction and posing flight safety risks as a result of added cognitive/physical workload (ALPA, 2016). Therefore, selecting an EFB that is easily understood and manipulated by an ab-initio pilot will result in enhanced training outcomes and increased flight safety.

Literature Review

Usability is a common assessment criterion of interactive products. A usability evaluation can be divided into inspection and testing methods (Inostroza, Rusu, Roncagliolo, Jimenez, & Rusu, 2012). The testing method determines usability problems through trials and observations of user/interface interactions (Otaiza, Rusu, & Roncagliolo, 2010). The International Organization of Standardization (ISO) defines usability as the extent to which a user can operate a product – under a certain context - regarding effectiveness, satisfaction, and efficiency (ISO, 1998). Nielsen (1994) expands on this definition by identifying the five attributes of usability: efficiency, satisfaction, learnability, memorability and errors. Nielsen's model was later enhanced by the work of Harrison et al. (2013), who included cognitive load, which can be partially assessed using the NASA-TLX (task load index) developed by Hart and Staveland (Hart & Staveland, 1988). Three major factors must be considered when evaluating usability: user demographics, the goal of the task, and context of use (Harrison et al., 2013; ISO, 1998). The results of the usability scores will vary, depending on the participant demographics and tasks of interests. Kaikkonen et al. (2005) compared to field and controlled laboratory usability experiments on a generic commercial mobile application, which showed that both groups experienced the same problems, but with different frequencies. This suggests that the results from the current study would be similar to those observed in a field study.

Several authors have conducted usability studies on mobile applications. Chen and Zhu (2011) used an analytical hierarchy process to assess the usability of a mobile music application. Stoyanov et al. (2015) developed a mobile application rating scale to evaluate the usability of mobile health applications, Brachtl et al. (2001) assessed the usability of 3D navigation systems for personal digital assistant systems, and Christie et al. (2004) compared two basic interface designs. Additionally, Beck et al. (2003) tested mobile computer systems in laboratory settings,

and Masoodian and Lane (2003) compared the effectiveness of graphical and textual visualizations on different kinds of travel itinerary information.

Studies show that usability is an important metric because it is commonly associated with successful products and interfaces (Jordan, 1997; Lohse, 2000; Atyeo & Robinson, 1995; Maguire, McDonagh, Hekkert, van Erp, & Gyi, 2004), and has financial implications for the commercial and industrial sectors (Mack & Sharples, 2009). Lack of usability can cause problems in varying degrees of severity, such as frustration and time wasting (Mack & Sharples, 2009). Beyond the general usability of a product, it has been shown that aesthetics, emotion, and expectations also influence the user's experience (Lindgaard & Dudek, 2003). Several studies suggest that other features have priority over usability. Lightner (2003) has examined the importance of usability in e-commerce via online shopping sites. Her results showed that customers valued security, information quality, and data quantity as the most important attributes. Jordan and Thomas (1995) have shown that print quality (accuracy) was evaluated as the most important characteristic of a printer by users purchasing a printer online, rather than its ease of loading (usability). Although the paradigm of user product selection is complicated - with the usability of a product not always the top priority - the safety risks associated with a product must also be considered. In terms of EFBs, usability and accuracy of flight planning applications must be a top priority, considering the flight safety implications of distraction and increased workload.

Norman (1990) has highlighted the practical difficulties associated with assessing a product's usability before purchase. Davis (2002), as referenced by Mack and Sharples (2009), has shown the importance of product aesthetics in purchases. When comparing two mobile phones, Davis (2002) discovered that the participants initially chose a phone based on 'look.' However, after having the opportunity to use the phone, complete various tasks and examine the quality, the participants changed their initial decisions. This study clearly illustrated the importance of appearance prior to using a product. Regarding mobile applications, a majority of paid applications provide some free trial period, allowing users to assess the product more thoroughly. However, Anderson (2009) states that 95% of users use the free service, with the remaining 5% of the users willing to pay for the premium features. This suggests that a significant portion of the demographic does not experience the product's full potential and truly quantify its usability, thus partially basing their assessment on aesthetics rather than usability. Conducting a rigorous usability study on flight training applications for ab-initio users will assist flight training personnel and student pilots by providing a baseline assessment that informs their product buying decisions.

Chandra et al. have conducted considerable research in assessing, evaluating, and standardizing EFBs (D. Chandra; D. C. Chandra & Kendra, 2010; D. C. Chandra & Yeh, 2006a, 2006b), much of which has been implemented by the FAA. The majority of Chandra's work focuses on understanding the human factor influences associated with the integration and use of EFBs. Usability studies of EFB flight planning applications are limited. The current study conducts a usability study on commercially available flight planning applications in the context of ab-initio pilot training. The results of the study aim to provide information on applications and EFBs to ab-initio flight training units, institutions, and individuals.

DOI: https://doi.org/10.15394/ijaa.2017.1162

Methods

This study used a summative usability approach to compare flight planning applications for mobile devices (Albert & Tullis, 2013). When comparing products, the context in which they are used greatly affects their usability. Additionally, developers also design their applications based on different goals; for example, maximizing efficiency or creating an exceptional user experience. The usability score of the applications is only applicable for the various contexts in which they are scored (i.e. user, environment, etc.).

Software and Apparatus

All of the applications considered under the current study focus on providing the required utilities and information to prepare an ab-initio pilot for a cross-country flight. The three applications being considered are:

1. ForeFlight (Version 7.1.1 (1708), ForeFlight, LLC. Houston, TX, USA),
2. Garmin Pilot (Version 7.3.1, Garmin Inc., Chanhassen, MN, USA), and,
3. FltPlan Go (Version 4.0.2, Flight Plan LLC. CT, USA)

All applications were subscription/unrestricted versions, with complete updates for Canadian airspace. The mobile tablet device used in the study was an iPad Air 2, 64 GB version with cellular capabilities (Apple Inc., CA, USA). The device was chosen based on its compatibility with the aforementioned flight applications.

Participant/User Inclusion Criteria

The sample for this study was comprised only of students enrolled in the Seneca College Bachelor of Aviation Technology Program. Additionally, participants had to meet the appropriate inclusion criteria to be eligible. It should be noted that user attributes play a major role in quantifying usability and that the usability rating of the application may differ with demographic compositions (Harrison et al., 2013). The participant inclusion criteria were:

- Minimal/nil exposure to EFB or flight applications
- Less than 100 total flight hours, but with at least one completed cross-country flight
- Some experience with Apple products/touch screen interfaces

The constraints of the inclusion criteria ensured that the candidates would be familiar with the tasks asked during the usability study while reducing the noise due to varying demographics. As expected, the inclusion criteria eliminated a majority of the available students. Faulkner (2003) reports that a minimum of 95% of the usability issues was discovered with 20 users, and the variation between the groups was small. Kaikkonen et al. (2005) state that an ordinary usability test often consists of 5-10 users per test round, which is supported by Albert and Tullis, who report that meaningful results can be extracted from 8-10 participants (Albert & Tullis, 2013). Although larger sample sizes are ideal, Albert and Tullis (2013) recommend a minimum sample size of 30 if resources are limited. Thus, a sample size of 30 was chosen; however, based on the variation of usability tests, and the consistency among the test results, a sample size of 30 was shown to provide sufficient accuracy.

Data Collection

An experimental moderator both collected data and administered usability tests. The data was compiled, de-identified, and statistically analyzed using IBM SPSS statistical software. Ethics approval for the study was granted by the Seneca College Research Ethics Board. Candidate participation was voluntary, and recruitment conducted via email and in-class discussions; candidates received gift cards for participating in the study.

Experimental Procedure

The experimental procedure was separated into two categories based on the various collection methods: quantified performance (usability) and self-reported metrics. In the first phase of the test, which quantifies the performance of the flight applications, participants completed a series of tasks. The tasks were quantified by the following sub-categories:

i. task success
ii. time on task
iii. efficiency
iv. learnability
v. cognitive load

These sub-categories were chosen based on the formulation of Nielsen (1994) (errors, efficiency, learnability) with the additions of cognitive load, which has been shown to be an important aspect of usability (R. Adams, 2007; R. G. Adams, 2006).

The study followed the typical usability testing experience, in which the moderator asked the candidate to complete tasks that are common throughout all applications. The experiment consisted of six tasks that are common practices of ab-initio pilots planning cross-country flights. These tasks include identifying various locations on a map, evaluating the weather, and developing a flight route.

Before the session began, the candidate was briefed on the objective of the experiment, which was to assess their completion of various tasks. The applicant was told that they would not be given any assistance to complete the tasks and that they could concede to a task at any time. No external resources were permitted during the testing process. The candidate was supplied with a list of the airport names, identifiers, and locations for reference during the session. The overarching research objective – comparison of flight apps - was withheld from the applicant to avoid biased behavior or increased performance anxiety. The order in which the participants used the applications was randomized based on the number of possible permutations, which was six. The use of a cross-over methodology alleviated order/learning effects of the subjects. The implementation of the cross-over methodology, in combination with the inclusion criteria, provided a clean dataset for meaningful statistical analysis.

When participants arrived, the applications were opened on the screen, showing the world map in full view. All applications have the ability to navigate the world maps. The moderator-initiated the quantitative testing procedure by having the candidate complete a series of relatively simple tasks, as discussed in the following subsection (Usability Analysis). The moderators

dictated the task as written in the quantitative analysis section, and provided clarification on the task objectives without giving additional/assisting information. After the quantitative portion, the candidate was then requested to complete various self-reported metrics on the application's performance. The quantitative analysis of the self-reported metrics is elaborated on in further subsections (Self-reported Metrics). The experimental procedure is summarized in Figure 1.

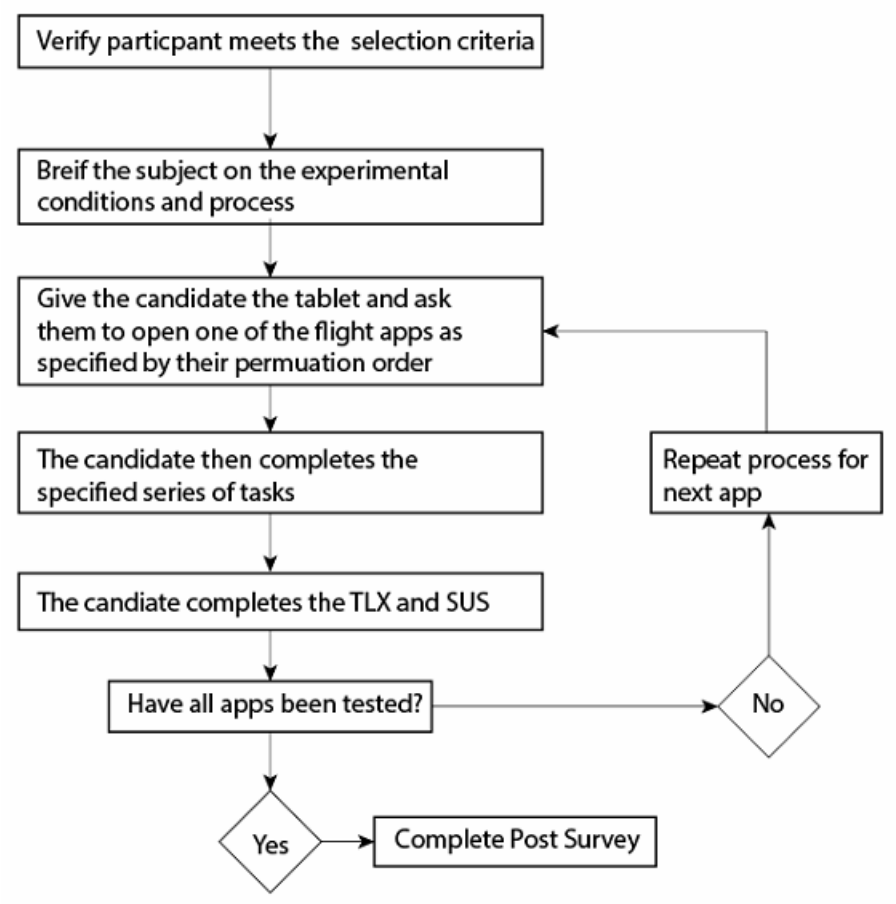

Figure 1. Experimental procedure flow chart.

Usability Analysis

This section outlines the various tasks and metrics used in the quantitative analysis of the experiments. The tasks used to assess the usability of the applications in their respective order are:

1. Determine current location of Bella Coola Airport - CYBD (Victoria, British Columbia, Canada).
    a. In the first task, the applicant is required to locate the position of Victoria International Airport. This airport will be the starting location of the flight route developed in later tasks. Since the participants attended Seneca College in Ontario, Canada, it was assumed that the students would have little/nil prior geographical

knowledge of Western Canadian airspace, since all cross-country training flights are restricted to Ontario and Eastern Canada.

2. Determine the METAR at Victoria International Airport – CYYJ (Victoria, British Columbia, Canada).
    a. Determination of the weather at various locations is a pivotal part of flight planning. Additionally, this task is repeated three times throughout the experiment to determine the learnability of the application.
3. Create a flight route from Victoria International Airport, to Kelowna Airport, to Bella Coola Airport (CYYS-CYLW-CYBD).
    a. This can be classified as the most complicated task in the usability study but is critical in determining the usability of the applications under the current study.
4. Determine the METAR at Kelowna Airport – CYLW (Kelowna, British Columbia, Canada).
5. Find the tower frequency for Victoria International Airport – CYYJ (Victoria, British Columbia, Canada).
6. Determine the METAR at Bella Coola Airport (Bella Coola, British Columbia, Canada).

The quantifiable metrics used to assess the tasks are task success, time on task, efficiency, and learnability.

Task Success. Task success measures how effectively users are able to complete a given set of tasks (Albert & Tullis, 2013). This metric is similar to effectiveness defined by ISO 9241 (1998) and Harrison et al. (2013). The task(s) should be relevant to the objective of the study, and must have clearly defined success criterion, which can be either binary or leveled. In addition to the success criterion, the study must also specify a failure criterion. Failure criteria can consist of the applicant producing the wrong result, taking too long in completing the task, having the moderator stop the participant from completing the task due to frustration, anxiety, or lack of progress, and the participant conceding.

For the current study, task success utilized the binary method to determine if the participant was capable of completing the task, with "1" indicating a task success and "0" representing a failure. The task success criteria corresponding to the aforementioned tasks were as follows:

1. Determine the current location of Victoria International Airport – CYYJ (Victoria, British Columbia, Canada).
    a. The participant displays the location of Victoria International Airport.
2. Determine the METAR at Victoria International Airport – CYYJ (Victoria, British Columbia, Canada).
    a. The participant correctly identifies the current METAR for Victoria International Airport.
3. Create a flight route from Victoria International Airport, to Kelowna Airport, to Bella Coola Airport (CYYS-CYLW-CYBD).
    a. The participant constructs a flight route to the proper airports, in the correct order.
4. Determine the METAR at Kelowna Airport – CYLW (Kelowna, British Columbia, Canada).
    a. The participant correctly identifies the current METAR for Kelowna Airport.

DOI: https://doi.org/10.15394/ijaa.2017.1162

5. Find the tower frequency for Victoria International Airport – CYYJ (Victoria, British Columbia, Canada).
    a. The participant correctly identifies the correct tower frequency (119.1) for Victoria International Airport.
6. Determine the METAR at Bella Coola Airport - CYBD (Bella Coola, British Columbia, Canada).
    a. The participant correctly identifies the current METAR for Bella Coola Airport.

According to Albert & Tullis (2013), in most situations, faster completion time is better (inclusive of the current study, but for some applications such as gaming or e-learning, longer times might be beneficial). Time on tasks is more important for tasks that are repeated by the participant/customer.

In the current study, the moderator used a stopwatch to time the completion of tasks. The participants were asked to verbally declare when they were beginning or finishing a task, and prompted by the moderator as needed. After the participant declared they were starting the task, the moderator would begin timing. The moderators were trained on the flight applications by experienced personnel and had a strong competency with all of the EFB applications being tested. After the participant had declared they completed the task, or conceded, the moderator would stop timing, and validate if the participant was successful. Subsequently, the moderator would record the time and success results without revealing the results to the participant.

Efficiency. Efficiency is the amount of effort required to complete a task (i.e. Number of gestures/actions required to complete the task) (Albert & Tullis, 2013; Harrison et al., 2013; ISO, 1998). There are two types of effort – cognitive and physical. Cognitive is thinking about the action (i.e. finding the button to press), whereas physical is touching the button. The cognitive loading is analyzed in preceding subsections. In this study, efficiency is represented as a combination of task success and time. This is done by using the 'Common Industry Format (CIF) for Usability Test Reports, (ISO/IEC 25062:2006) (ISO, 2006) and defined as:

$$Efficiency = \frac{Task\ Success}{Time\ on\ Task}$$

Based on the task success and time on task measures recorded by the moderator, the efficiency metric was calculated post study.

Learnability. Learnability is the extent to which something can be learned efficiently (Harrison et al., 2013). It is an important factor in the current application of EFBs since pilots will be using the device repeatedly. Quantifying learnability was accomplished by utilizing repeat tasks within the same session, but with breaks between each task (Albert & Tullis, 2013). This methodology was accomplished by having the participant determine the METAR information at three different airports, at three separate times during the session. The application's learnability metric is compared to the rate of learnability. The learnability was assessed by measuring the change in time on task for the three METAR tasks, with the expected trend of decrease in task time, with an increase in task repetition.

Self-reported Metrics

The self-reported metrics questionnaires (i.e. SUS, TLX, post-survey) were administered immediately after the quantitative usability testing. The SUS, TLX and post-survey were used to measure the users' cognitive load, usability, and overall satisfaction. Each participant completed three SUS and TLX questionnaires, which were administered immediately after each round of usability task testing, and assessed post study.

Cognitive Load. Harrison et al. (2013) define cognitive load as the amount of cognitive processing required by the user to use the application. This attribute was measured with the NASA Task Load Index (TLX) method. The TLX method assesses the workload the participant experiences when performing a task(s) (Sandra G. Hart, 2006).

Usability. Usability, as a self-reported metric, is commonly measured by the System Usability Scale (SUS) (Brooke, 1996). The SUS consists of ten phrases (five positively worded and five negatively worded) to assess users' usability experience. The overall results of the SUS should correspond with the results of quantitative usability testing.

Satisfaction Survey. The satisfaction survey reports on the various characteristics of the applications. Unlike the TLX and SUS, the satisfaction survey is only administered once at the very end of the session and summarizes various characteristics perceived by the participant. The survey is broken up into three sections:

i. How important is a factor: price, support, and functionality/features when selecting a flight application?
ii. Which application has the most appealing graphical user interface (GUI)?
iii. Which application did the student like the most?

Results and Discussion

Participant Demographics

All participants were first-year students in a collegiate aviation program in Canada. Based on a pre-selection survey, candidates rated themselves (median) as having had 'minimal' experience with EFBs or flight planning applications. All candidates had completed their cross country training and had less than 100 hours of flight time. Thus, it was inferred that they possessed the required flight planning knowledge, but were considered ab-initio pilots. The candidates rated themselves as experienced (median) with Apple products and touch interfaces; however, the mode of the results indicated that the participants considered themselves very experienced. This suggests that the results of the usability study are heavily dependent on the applications themselves, not on the interface. The mean age of the group was 20.9 +/- 2.7 years.

DOI: https://doi.org/10.15394/ijaa.2017.1162

Usability

A multi-variant analysis of variance (MANOVA) using SPSS 20 was conducted to compare the usability scores among three groups. The results of the MANOVA revealed several significant differences. The post-hoc analysis involved a series of two-way analyses of variance (ANOVA). The F-value, significance, and power are shown in Table 1.

Table 1

Statistically significant differences among the three groups (Foreflight, Garmin Pilot, Flight Plan Go) based on usability metrics.

Usability Test Number	F-Value	Significance	Power
Task 1 - Time on Task	8.832	.000	.967
Task 3 - Success	3.274	.043	.608
Task 3 - Time on Task	5.350	.006	.892
Task 3 - Efficiency	10.037	.000	.982
Task 4 - Time on Task	8.658	.000	.964
Task 4 - Efficiency	14.347	.000	.998
Task 5 - Time on Task	4.624	.012	.767
Task 5 - Efficiency	7.852	.001	.946

α=0.05

Eight of the 18 tasks produced significant differences among the three groups, which will be discussed in the following subsections.

Success. Of the exercises, only one task (Task 3: generating a flight route from Victoria International Airport to Kelowna Airport to Bella Coola Airport) produced a significant difference among the three groups. Foreflight had the most successful result (mean = .90), followed by Garmin (mean = .80), and lastly, FltPlan Go (mean = .63). As expected, Task 3 produced a significant difference regarding success, due to the exercise's complexity when compared to Tasks 1 and 2, compounded with the effect that participants were still relatively unfamiliar with the application. The lack of significant differences among the success metrics for each application suggests that all applications are capable of successfully completing the task correctly. Additionally, considering the relatively high significance value and low power (Table 1), it can be concluded that there is a negligible difference among the success metric of each application.

Time on Task. Figure 2 shows the time-on-task results for the three groups. The results show that Foreflight consistently had the lowest time on task results, suggesting that Foreflight is capable of producing the desired output faster than its counterparts. The task times for Garmin and FltPlan Go are split, with Garmin performing better in Tasks 1 and 2, and FltPlan Go performing better in Tasks 4 and 5. The results also show a consistent trend among all tasks, i.e. relatively low Task 1 time, followed by a spike in Task 3 time, and a decrease in Tasks 4 and 5. A possible explanation for the decreasing trend is learning effects, due to increased familiarity with the application, and the simplicity of the task.

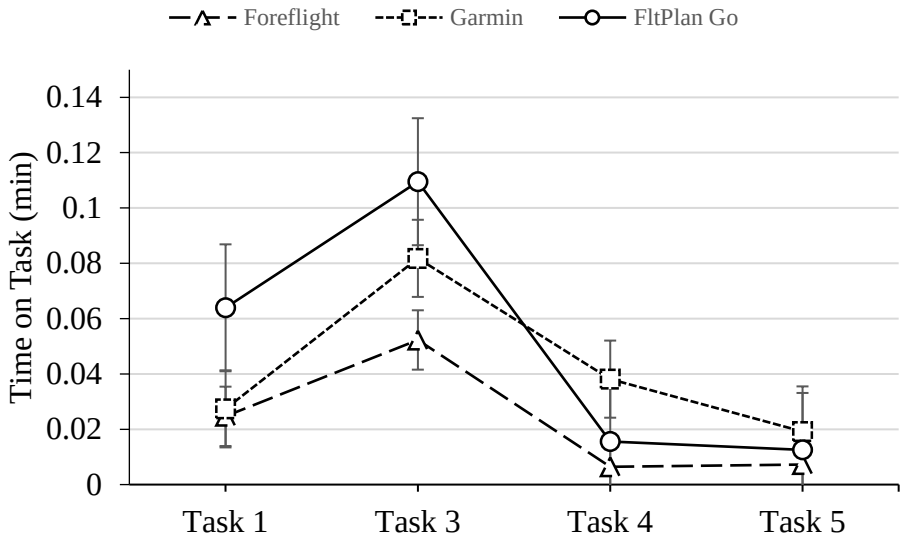

Figure 2. Time on task for significant results among the three groups.

Efficiency. Figure 3 displays the efficiency results for significant tasks among the three groups. The results show that Foreflight was the most efficient, followed by FltPlan Go, and lastly, Garmin. The results followed a similar trend as the time-on-tasks results, with FltPlan Go outperforming Garmin. This is due to the formulation of efficiency, which is dependent on success and time-on-task; of which, both FltPlan Go and Garmin have similar success rates for Tasks 3, 4 and 5.

DOI: https://doi.org/10.15394/ijaa.2017.1162

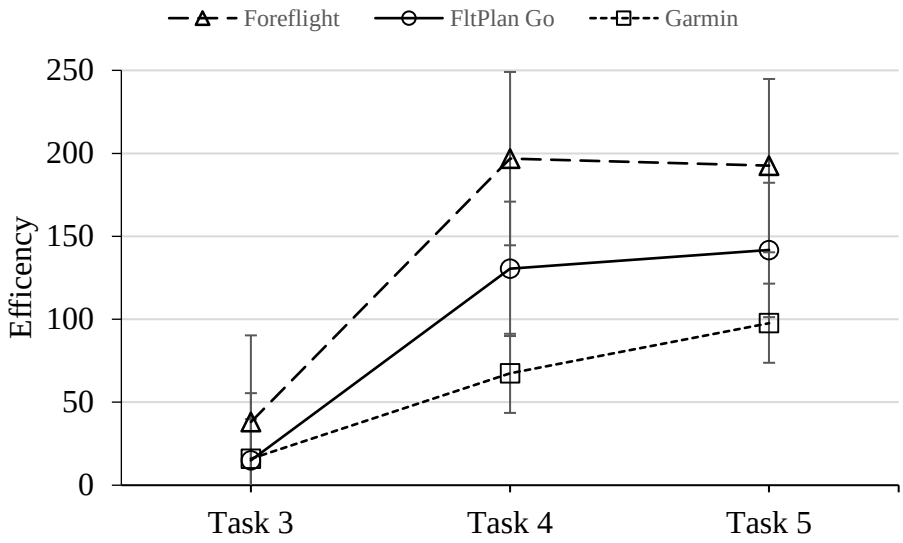

Figure 3. Efficiency for significant results among the three groups.

All applications follow a consistent trend. As application exposure increased, efficiency increased as well; possibly due to learning effects and increased familiarity. Based on these trends, it appears that Foreflight and FltPlan have reached a steady state efficiency value, while Garmin is still growing.

Learnability. Figure 4 presents the mean values of the time-on-task metric for the learnability tasks. The results show that Foreflight had the fastest learnability, followed by FltPlan Go, and lastly, Garmin. All applications show an overall decrease in time-on-task with increased repeatability; however, both FltPlan Go and Garmin show negative learnability effects on the second task repeat, compared to Foreflight, which has decreasing scores with each repetition.

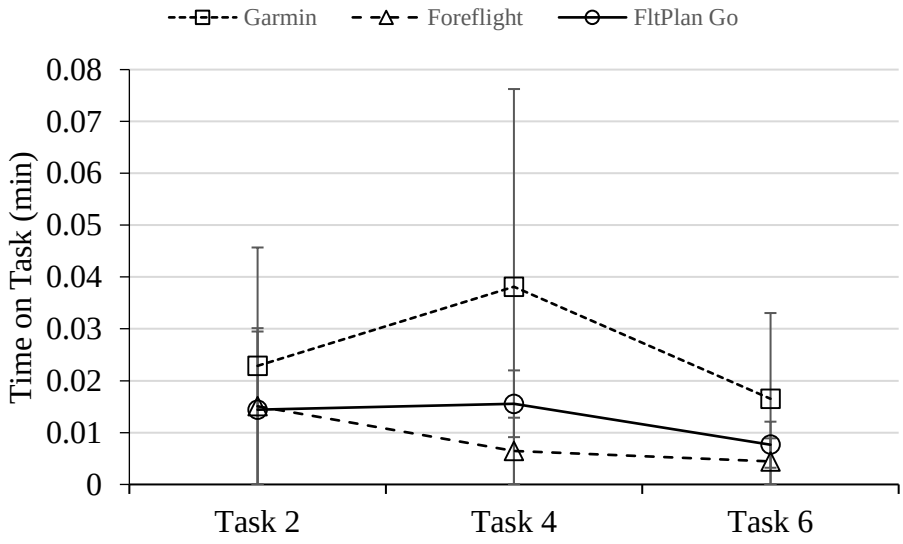

Figure 4. Learnability results.

Task Load Index

A MANOVA was conducted to compare the three groups and revealed several significant differences. The dependent variables were then analyzed separately, using a two-way ANOVA. The F-value, significance, and power are shown in Table 2.

Table 2

Statistically significant differences among the three groups (Foreflght, Garmin Pilot, Flight Plan Go) based on the TLX test results.

TLX Metric	F-Value	Significance	Power
Mental Demand	13.302	.000	.997
Physical Demand	4.098	.020	.712
Temporal Demand	4.764	.011	.780
Performance	4.111	.020	.714
Effort	13.776	.000	.998
Frustration	16.646	.000	1.000

α=0.05

The results of the TLX are shown in Figure 5, and demonstrate that the participants unanimously rank the applications - in descending order - Foreflight, Garmin, and FltPlan Go, across all categories, respectively. The performance metric of the TLX supports the grand mean (success metric mean of all tasks, significant and not), and significant results from the usability study, which showed that the Foreflight (grand mean: 0.967) had the best performance, followed by Garmin (grand mean: 0.939), then FltPlan Go (grand mean: 0.894).

The temporal demand metric shows that the user thought FltPlan Go required the most time to achieve the desired output, followed by Garmin, and then Foreflight. These results correlate with the grand mean time-on-task results (Foreflight: 0.018 min, Garmin:0.034 min and FltPlan Go:0.037 min).

The physical demand characteristic gives an indication of the number of tactile user inputs required per task. The results show that FltPlan Go had the largest physical demand (followed by Garmin and Foreflight), suggesting that it required more user inputs per task solution.

FltPlan Go required the most mental demand, effort, and had the highest level of frustration. These metrics can be related to cognitive loading, suggesting that FltPlan Go may have reduced memorability effects, compared to the other applications.

DOI: https://doi.org/10.15394/ijaa.2017.1162

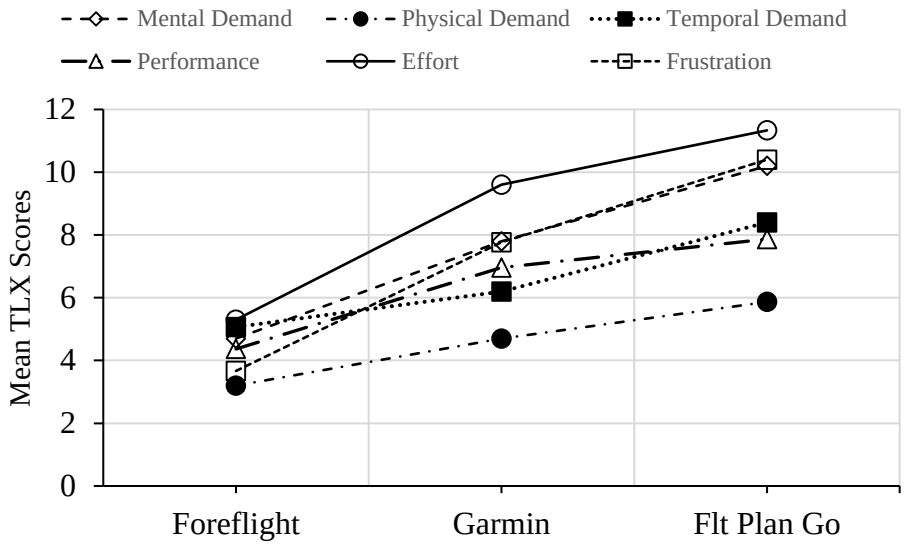

Figure 5. TLX results.

System Usability Scale

A MANOVA was conducted to initially compare the three groups and revealed several significant differences. The dependent variables were then analyzed separately, using a univariate two-way ANOVA. Table 3 shows the F-value, significance, and power.

Table 3

Statistically significant differences among the three groups (Foreflight, Garmin Pilot, Flight Plan Go), based on the SUS test results.

SUS Question Number	F-Value	Significance	Power
Question 1	21.127	.000	1.000
Question 2	9.452	.000	.976
Question 3	16.672	.000	1.000
Question 4	7.806	.001	.945
Question 5	21.780	.000	1.000
Question 6	7.531	.001	.937
Question 7	7.315	.001	.930
Question 8	10.250	.000	.984
Question 9	22.412	.000	1.000
Question 10	7.551	.001	.938

α=0.05

Figure 6 shows the SUS scores from each participant. Based on their mean values, Foreflight, Garmin, and FltPlan Go had a mean score of 82.6 +/-11.9, 63.1 +/-22.2, and 49.1+/-22.7,

respectively. As previously mentioned, it was expected that these results would correlate well with the quantitative results, thus further validating the experimental results/design.

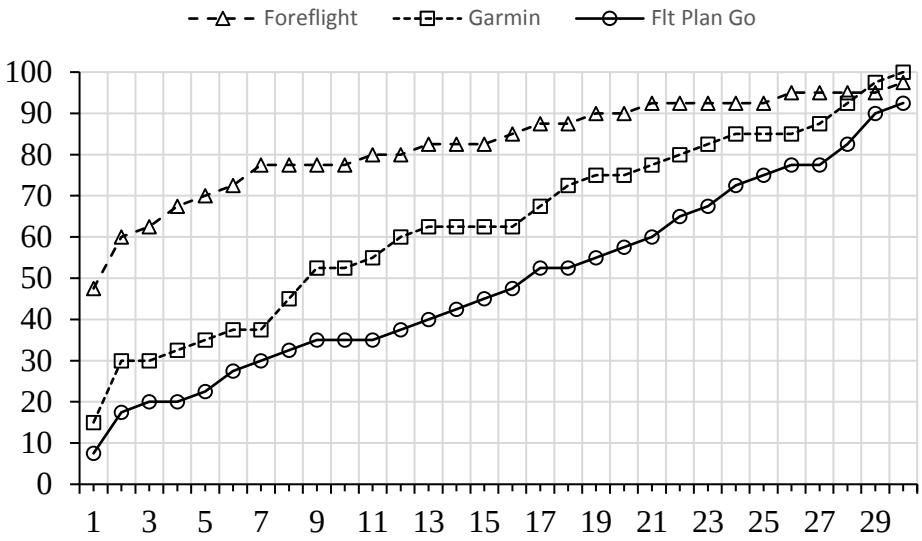

Figure 6. SUS participant scores.

Post Selection Survey

Of the three applications, Foreflight (53.3%) was rated as having the most appealing graphical user interface, followed by Garmin (40.0%), and lastly FltPlan Go (6.7%). The most-liked application among the participants was Foreflight (73.3%), followed by Garmin (23.3%), and lastly FltPlan Go (6.7%). Based on the importance of application price, 56.7% of the participants considered the cost factor important, followed by 26.7% reporting that price was critical, and 10.0% being neutral on the matter. The number of participants that considered regular updates and support crucial when selecting a flight application was 43.3%.

Conclusion

This paper examines the results of a usability study for consumer EFB flight planning applications. The flight applications considered were ForeFlight, Garmin Pilot, and FltPlan Go. The participant demographics consisted of 30 first-year students in a collegiate aviation program, who had completed their cross-country training, had less than 100 hours of total flight time, and had minimum experience with flight applications - with an average age of 20.9 +/-2.7 years. The usability of the applications was based on success, time-on-task, and efficiency usability metrics. The participants also assessed the applications using the NASA TLX scale, SUS, and a post data collection survey. The data of the usability tests, TLX, and SUS were statistically analyzed using a two-way ANOVA. The results from the statistical analysis showed that Foreflight had the best scores for all significant effects in the usability study, TLX, and SUS, followed by Garmin and FltPlan Go. Based on the post data collection survey, participants favored Foreflight over Garmin

and FltPlan Go, with 56.7% and 43.3%, respectively, considering price factor and regularly updated/support important when purchasing flight applications.

A limitation of the study is the relatively small sample size. This limitation was due to a lack of accessibility to participants that met the inclusion criteria. Statistically, a sample size of 30 is the minimum viable sample size that should be used in a usability study. As a result, several quantitative usability metrics did not produce significant differences, despite being close to the 0.05 significance level threshold. Similarly, the participants were drawn from the same training institution, adding a possible bias to the results. Another limitation of the study was the determination of the efficiency metric and physical demand metric. These metrics can be more accurately represented by measuring the number of actions/steps the user took to complete a task. Due to lack of available technology (i.e. iPad API that could record the number of user inputs per period), the author opted for a more generalized approach.

Building on present work, future investigations could assess the variation of the usability of EFBs with variation in flight experience. In the current study, a specific training demographic was considered. Further understanding could be gained by conducting similar usability methodologies on other training demographics, such as the instrument flight rating, multi-engine, and airline transport pilot license. Examining the possible morphology of usability preferences on experience/training levels could provide valuable insights for training institutions and application developers. Similarly, comparative studies that assess flight planning skills between EFBs and paper chart methods are crucial in identifying the positive and negative effects of increased aviation technology use in the pilot training environment. Further validation/enhancement of EFB research could be attained by replicating the current study at alternate training institutions – preferably at an international level. The results of the study could be used to further validate the current work, or provide understanding for discrepancies between ab-initio usability studies.

Acknowledgements

The current study was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC), in collaboration with the Office of Applied Research and Innovation (ARI) at Seneca College. The author would like to acknowledge the support from the Seneca College School of Aviation.

References

Adams, R. (2007). Decision and stress: Cognition and e-accessibility in the information workplace. *Universal Access in the Information Society, 5*(4), 363-379.

Adams, Ray G. (2006) Applying advanced concepts of cognitive overload and augmentation in practice: The future of overload. Schmorrow, Dylan D. and Stanney, Kay M. and Reeves, Leah M, (Ed.), In: *Foundations of augmented cognition: Augmented cognition: Past, present, and future* (pp. 223-229). Arlington, Vi: Strategic Analysis Inc.

Albert, W., & Tullis, T. (2013). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics*. Waltham, MA: Elsevier Science.

ALPA. (2016). Safety and training: Partners in advancing aviation. Retreived from http://www.alpa.org/news-and-events/news-room/2016-08-22-air-safety-forum-safety-training

Anderson, C. (2009). *Free: The future of a radical price*. New York, NY: Hachette Books.

Atyeo, M., & Robinson, S. (1995). Delivering Competitive Edge. In K. Nordby, P. Helmersen, D. J. Gilmore & S. A. Arnesen (Eds.), *Human—Computer Interaction: Interact '95* (pp. 384-385). Boston, MA: Springer US.

Beck, E., Christiansen, M., Kjeldskov, J., Kolbe, N., & Stage, J. (2003). *Experimental evaluation of techniques for usability testing of mobile systems in a laboratory setting.* Paper presented at the OzCHI, Brisbane, Australia.

Brachtl, M., Šlajs, J., & Slavík, P. (2001). PDA based navigation system for a 3D environment. *Computers & Graphics, 25*(4), 627-634.

Brooke, J. (1996). SUS- A quick and dirty usability scale. *Usability Evaluation in Industry, 189*(194), 4-7.

Chandra, D. (2003). *A tool for structured of electronic flight bag usability.* Paper presented at the Digital Avionics Systems Conference, Indianaoplis, IN.

Chandra, D. C., Kendra, A. J. (2010). *Review of safety reports involving electronic flight bags*. Washington, DC: Air Traffic Organization Operations Planning, Human Factors Research and Engineering Group.

Chandra, D. C., & Yeh, M. (2006). *Evaluating electronic flight bags in the real world*. Cambridge, MA: Volpe National Transportation Systems Center.

Chandra, D. C., & Yeh, M. (2006). *A tool kit for evaluating electronic flight bags*. Washington, DC: U.S. Department of Transportation, Federal Aviation Administration.

DOI: https://doi.org/10.15394/ijaa.2017.1162

Chen, C.-J., Yang, S.-M., & Chang, S.-C. (2014). A model integrating fuzzy AHP with QFD for assessing technical factors in aviation safety. *International Journal of Machine Learning and Cybernetics, 5*(5), 761-774.

Chen, Z., & Zhu, S. (2011). *The research of mobile application user experience and assessment model.* Paper presented at the Computer Science and Network Technology (ICCSNT), Harbin: China.

Christie, J., Klein, R. M., & Watters, C. (2004). A comparison of simple hierarchy and grid metaphors for option layouts on small-size screens. *International Journal of Human - Computer Studies, 60*(5), 564-584.

Croft, J. (2015, Aug 16). FAA seeking nextgen ATC digital voice recorder, *The Weekly of Business Aviation,* p. 4. Retreived from http://aviationweek.com/awin-only/faa-seeking-nextgen-atc-digital-voice-recorder

Davis, M. (2002). *A usability test by 3G lab. Comparison of two camera phones: Nokia 7650 vs. Sony-Ericsson T68i*. Cambridge, UK: 3G Lab Limited.

FAA. (2014). *Guidelines for certification, airworthiness and operational use of electronic flight bags*. (120-76C).

Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers, 35*(3), 379-383.

Ghorbani, M. A., Khatibi, R., FazeliFard, M. H., Naghipour, L., & Makarynskyy, O. (2016). Short-term wind speed predictions with machine learning techniques. *Meteorology and Atmospheric Physics, 128*(1), 57-72.

Hart, S. G. (2006). Nasa-Task Load Index (Nasa-TLX); 20 years later. *Human Factors and Ergonomics*
Society Annual Meeting Proceedings, 50(9), 904-904.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology, 52,* 139-183.

Inostroza, R., Rusu, C., Roncagliolo, S., Jimenez, C., & Rusu, V. (2012). *Usability heuristics for touchscreen-based mobile devices.* Paper presented at the Ninth International Conference on Information Technology: New Generations (ITNG), Las Vegas, NV.

ISO. (1998). Software ergonomics requirements for office work with Visual Display Terminal (VDT). *Guidance on Usability* (pp. 22). Geneva: ISO.

ISO. (2006). Software Engineering -- Software Product Quality Requirements and Evaluation (SQuaRE) -- Common Industry Format (CIF) for usability test reports (pp. 46). Geneva: ISO.

Janakiraman, V. M., & Nielsen, D. (2016, 2016). *Anomaly detection in aviation data using extreme learning machines.* Paper presented at the 2016 International Joint Conference on Neural Networks (IJCNN), Vancover, BC.

Jordan, P. (1997). *Usability evaluation in industry: Gaining the competitive advantage.* Paper presented at the From Experience to Innovation. Volume 2. Proceedings of the 13th Triennial Congress of the International Ergonomics Association. Tampere, Finland.

Jordan, P., & Thomas, D. (1995). ... But how much extra would you pay for it? An informal technique for setting priorities in requirements capture. In Robertson, S. (Ed.), *Contemporary Ergonomics* (pp. 145-148). London: Taylor & Francis.

Kaikkonen, A., Kekäläinen, A., Cankar, M., Kallio, T., & Kankainen, A. (2005). Usability testing of mobile applications: A comparison between laboratory and field testing. *Journal of Usability studies, 1*(1), 4-16.

Lightner, N. J. (2003). What users want in e-commerce design: Effects of age, education and income. *Ergonomics, 46*(1-3), 153-168.

Lindgaard, G., & Dudek, C. (2003). What is this evasive beast we call user satisfaction? *Interacting with Computers, 15*(3), 429-452.

Lohse G.J.L. (2000). Usability and profits in the digital economy. In: McDonald S., Waern Y., Cockton G. (Ed.) *People and Computers — Usability or Else*!. London: Springer.

Mack, Z., & Sharples, S. (2009). The importance of usability in product choice: A mobile phone case study. *Ergonomics, 52*(12), 1514-1528.

Maguire, M., McDonagh, D., Hekkert, P., van Erp, J., & Gyi, D. (2004). *Does usability= attractiveness?* Paper presented at the Design and Emotion. Ankara, Turkey.

Masoodian, M., & Lane, N. (2003). *An empirical study of textual and graphical travel itinerary visualization using mobile phones.* Paper presented at the Proceedings of the Fourth Australasian User Interface Conference on User interface. Sydney, Australia.

Nielsen, J. (1994). *Usability engineering*. Cambridge, MA: Elsevier.

Norman, D. A. (1990). *The design of everyday things*. New York, NY: Basic Books.

Otaiza, R., Rusu, C., & Roncagliolo, S. (2010). *Evaluating the usability of transactional web sites.* Paper presented at the Third International Conference on Advances in Computer-Human Interactions. St. Maarten, Netherlands.

Pourcho, J. B. (2014). *Augmented reality application utility for aviation maintenance work instruction.* (Published master's dissertation). Purdue, West Lafayette, IN.

Stoyanov, S. R., Hides, L., Kavanagh, D. J., Zelenko, O., Tjondronegoro, D., & Mani, M. (2015). Mobile app rating scale: A new tool for assessing the quality of health mobile apps. *JMIR mHealth and uHealth, 3*(1), e27, 1-9.

DOI: https://doi.org/10.15394/ijaa.2017.1162

Williams, J. K. (2014). Using random forests to diagnose aviation turbulence. *Machine Learning, 95*(1), 51-70.