<table>
<thead>
<tr>
<th>Manuscript Number:</th>
<th>ASEM3372R1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Title:</td>
<td>Pilot Performance: Assessing How Scan Patterns &amp; Navigational Assessments Vary by Flight Expertise</td>
</tr>
<tr>
<td>Article Type:</td>
<td>Research Article</td>
</tr>
</tbody>
</table>
| Corresponding Author: | Ji Hyun Yang, Ph.D.  
Naval Postgraduate School  
Monterey, CA UNITED STATES |
| Corresponding Author's Institution: | Naval Postgraduate School |
| First Author:     | Ji Hyun Yang, Ph.D. |
| Order of Authors: | Ji Hyun Yang, Ph.D.  
Quinn Kennedy, Ph.D.  
Joseph Sullivan, Ph.D.  
Ronald D. Fricker, Jr., Ph.D. |
| Order of Authors Secondary Information: |
Frederick Bonato, Ph.D.
Editor-in-Chief
Aerospace Medical Association

Dear Dr. Bonato,

Enclosed please find the first revision of our manuscript ASEM 3372 entitled “Pilot Performance: Assessing How Scan Patterns & Navigational Assessments Vary by Flight Expertise”. The initial title of the manuscript was “Expertise effect on gaze patterns, navigation accuracy, and subjective assessment in overland navigation on varying route difficulty.”

We reply to all comments and questions of the reviewers and the editor in the following pages. We have restated the original review questions and comments in italic, which are followed by our answers and arguments in normal font. In an effort to update our paper, we have accepted most of the suggestions for enhancements and requests for correction. We believe that our revised paper answers all questions and clears all major objections of the reviewers.

We deeply appreciate all comments and suggestions from the editor-in-chief and reviewers.

Sincerely,

Ji Hyun Yang

Ji Hyun Yang, Ph.D.
Research Associate Professor
The MOVES Institute
Naval Postgraduate School
Monterey, CA 93943

Office: (831) 656-3004
Cell: (617) 877-4108
E-mail: jyan1@nps.edu
Answers to the Comments from Reviewer 1

This manuscript poses some interesting questions concerning the integration of pilot experience (in terms of their total flight hours) and behavioral patterns in the cockpit. However, there are some serious issues with the experimental design of this study which bring to question the results and conclusions obtained. My over-riding concern here is that the conclusions drawn (or not drawn for that matter) in this study are not necessarily substantiated by the data. I perceive a possible flaw in the experimental design of this study. I strongly feel that this study is very likely significantly under-powered. The authors have only 12 subjects comprising a possibly skewed sample from a very likely skewed population. I may be wrong, and if so fine, but this desperately needs clarification in the manuscript. With only 12 subjects tested, the limited data does not support the conclusion that TFH is "too crude to use as a measure of expertise for task specific activities".

We understand the reviewer’s concern about the conclusion and have deleted the phrase that TFH is "too crude to use as a measure of expertise for task specific activities" from the abstract and main text.

We appreciate the reviewer’s concern about power and sample representativeness. As explained below, the sample size actually does have reasonable power for interesting and useful effects of the size we desired to detect. In addition, we note that our sample is appropriate in the sense that it contains a relevant range of flight experience, from no actual piloting experience to instructor pilot experience. Given that we are trying to compare how experienced pilots’ behaviors differ from novice pilots’, the sample is relevant and appropro. We explain the power analysis and the sample demographics in response to the reviewer’s main concerns #1 and #2 (below).

It may very-well not be; however, this limited data from a small sample size doesn’t really support such a broad statement. This study should have been designed and conducted with a significantly larger sample size. I would strongly encourage the authors to go back and test additional subjects to make their data set more robust, of sufficient power, and representative of the natural population of helicopter pilots. Only then would I feel that the observations and conclusions made in this manuscript be valid.

We understand the reviewer’s concern on the small sample size. Our power calculations show that \(n=12\) is of sufficient size to detect moderate correlations (see response to the reviewer’s main concern #1), and the experimental subjects spanned the range of pilots from novice to expert and thus results from this sample are informative to a larger population as shown by demographic information (see our response to the reviewer’s main concerns #2).

Main concerns are:
1. With respect to the sample size: was there a power test done? If so, the assumptions it was carried out under, as well as the test and results need to be presented in this paper to justify the limited sample size.

Yes, power calculations were done and they show that a sample size of \(n=12\) provides adequate power for detecting correlation effects of the magnitude observed in the data. To clarify this, we added the following text: “Post-hoc statistical power analysis (3, 14) showed that the power of
correlation analysis ranges between .51 and .76 given sample size = 12, $\alpha = .05$, and observed $\rho = .52 - .69$.”

Furthermore, in our previous paper (Sullivan et al. 2011) we also included a power analysis, in which the corresponding power of our linear model ranged between 0.67 and 0.87 given sample size = 12, significance level $\alpha=0.05$, and observed $R^2 = 0.35 \sim 0.48$.

Finally, we note that a sample size of $n=12$ was adequate to achieve quite a number of statistically significant results in this experiment.

2. **What were the specific sample demographics?** I am not too concerned about the differences between males and females (although that would be interesting), but I am concerned about the sample possibly being skewed. I notice that this study occurred at the Naval Postgraduate School where usually only the very senior pilots attend, as typical with most advanced military education. If the sample size was skewed toward very senior subjects, and the sample was of limited size (which it is) then one may not see a significant effect of Total Flight Hours. Where all of the test subjects Navy/Marine pilots, or were there members from the other services? In the paper it was noted that: "TFH varied from 0 to 3100 hrs" and that "overland-flight-hours varied from 0 to 2,500hrs". This is an extremely varied sample, especially for it being so small. Also, where some of the participants not pilots at all (flight hrs = 0) as implied in the methods? This definitely needs to be clarified. While the Spearman’s rank correlation is a non-parametric test used for populations with an other-than-normal distribution, I still feel that the sample size may be limited in size and skewed. Further clarification of the sample demographics is needed.

First, we note that students at NPS have a wide range of aviation skills and experience levels. Though we don’t mention it in the paper, the vast majority of NPS students are O-3s and O-4s and, as such, they are junior to mid-grade officers. This composition reflects the majority of the officer corps as well as of military pilots.

The minimum skill requirement for the study was completion of at least one overland navigation class, meaning that subjects have overland navigation ability. Three of our subjects had only taken an overland navigation class and never actually flown. Since our task is focused on visual navigation, we thought subjects with only overland navigation were eligible.

We have included the following in the manuscript “ Seven subjects were from the U.S. Navy, two from the U.S. Marine Corps, one from the U.S. Air Force, one from the U.S. Army, and one was unknown.” All of the subjects were male. Histograms of TFH and ages are shown below.
3. In the third sentence, 1st paragraph of the "Subjects" heading in the Methods there is a mention of "fourteen subjects". Which is it, 12 or 14 subjects?

We have corrected fourteen to twelve.

4. RMS error is well-known in the aviation community; however it was never really defined in the beginning of the manuscript. The mathematical calculation of it was presented in the statistic sections of the Methods, but for the diverse readership of this Journal it really should be defined the first time it is used, like any other abbreviation.

We have added the RMS definition “root mean square” in the abstract and the first time it appeared in the main text.

5. What precisely was the overall goal of this study? I believe I understand, and I believe the authors touched on it when they said: "From a training perspective, understanding expertise differences in the link between navigation?..." in the Introduction. This is a very minor point, but I feel that it would be appropriate to
state this up front in the Introduction- it would allow the readers to follow the train-of-thought with respect to the experimental design a bit better.

Based on the reviewer's suggestion, the following sentence now appears in the Introduction section: "Thus, the purpose of this study was to attempt to understand underlying cognitive strategies used by experts that aid in superior performance."

6. It might be more appropriate to express the hypotheses stated at the end of the Introduction in terms of the "null" hypothesis- especially since there was no significant effect seen. This is usually how scientific studies are described.

As the reviewer points out, the scientific method is focused on rejecting or retaining the null hypothesis. However, in our experience, hypotheses in scientific journal articles typically are stated in terms of the alternative hypotheses and we have thus explicitly so labeled them. Please see the end of the Introduction section.

7. Under the section on "Navigation task" in the Methods, who exactly was this subject matter expert? Was it an IP? What were his/her qualifications that made him/her a SME?

SME was an instructor pilot (IP) as well as a Standardization Pilot (SP) responsible for training instructors.

8. 2nd line, 1st Paragraph of Statistical Analysis in Methods section: The authors refer the reader to a 2011 paper by Sullivan on regression analysis; but never really state if they did indeed use regression analysis in this study. It seems that the Spearman rank correlation was the only statistic discussed in the Results/Discussion.

In this study, we did not use regression analysis. Spearman’s rank correlation and paired t-tests were the statistics discussed in the Results and Discussion section. We did not see any further advantage to using regression analysis in this study.

9. Results section, Paragraph 2 - sentence 3. Not sure of the significance of this observation that the pilots tend to spend more time looking out the window than at the map? Shouldn’t all pilots spend more time looking out the window and flying the aircraft rather than fixating on the map?

Indeed, pilots were told to maintain their OTW scan – what we were looking at was the ratio of OTW scan to MAP scan, and tested whether the route difficulty affected the OTW percentage, e.g., the more difficult the route, the more MAP scan the pilot did compared to OTW scan. However, the route difficulty was not associated with significant changes in OTW-MAP scan ratio. In our previous study (Sullivan 2011), experts tended to have less OTW scan than MAP, which was opposite to our expectation.

10. Results Section - Paragraph 3, sentence 5. With respect to the significant correlation between post-test assessment and Navigation accuracy- was this a positive or negative correlation. I would recommend on all of the results where there is a mention of a "correlation" that the authors describe what kind of correlation (positive or negative) it is.
Post-test assessment and navigation accuracy were negatively correlated. Following the reviewer’s suggestion, we added the direction of correlation (positive or negative) when mentioning correlation in Results section.

11. Results Section: HYPOTHESIS 1, First sentence. I have several issues with this sentence. First, the statistical test itself does not support (partially or not) a hypothesis. The statistical test validates the data that actually rejects or supports a null hypothesis. Second, how can the data "partially support" the hypothesis as stated. The statistical analysis is either significant or not, so the data either supports or it does not support the hypothesis. One either accepts the null hypothesis or rejects it and accepts the alternative hypothesis. Based on this, from what I can gather, the data from this study does not support the hypothesis of the relationship as described by this sentence. Again, this is very likely due to the study being under-powered and limited by the small sample size. I do not understand how the authors can say in the next sentence that TFH and navigation accuracy were "marginally associated". If one accepts a standard statistical cutoff of 0.05, then p=0.39 P=.103 is not significant. Also, with respect to this, I did not see in the Statistical Section of the Methods where the authors stated their accepted level of significance for the analysis. It was listed in the first line of the results, as alpha=0.05, but it should have been stated in the Methods.

We understand the reviewer’s point, and we appreciate the need for a predetermining a significance level in advance of a confirmatory analysis. However, it is also well-accepted practice to report p-values and interpret them as “evidence against the null hypothesis” (c.f. R.A Fisher, The Logic of Inductive Inference (with discussion), Journal of the Royal Statistical Society, 1935) as we did in our original draft. This practice is particularly useful in exploratory data analysis and for allowing journal readers to assess the hypothesis test outcomes for other choices of significance level.

One common convention is that P<0.001 means the null hypothesis is “strongly rejected,” 0.001<P<0.05 means the null is rejected, and 0.05<P<0.1 means the null is weakly rejected. (See, for example, “Interpreting P-values” at http://bmbolstad.com/teaching/Stat20.F03/2003Oct20.pdf or “Introduction to Hypothesis Testing at http://www.wright.edu/~thaddeus.tarpey/stt630chap6.pdf.) In a similar vein, in our original draft we used the term “marginally associated” when the p-value was around 0.1. We certainly are aware that this did not represent a statistically significant result under assumption of α = 0.05.

To eliminate any confusion, we have removed the sentence containing the marginal association phraseology. Furthermore, we have revised the paper so that all reject/fail to reject conclusions from the hypothesis tests are based on a fixed significance level of α = 0.05 and we have explicitly stated that significance level in the Methods section.

In terms of other specific changes, we have modified the first sentence as follows: “The experimental data did not support either Hypothesis 1.a regarding the relationship between TFH and navigation accuracy or 1.b regarding the relationship between TFH and RMS error.”

We grouped several similar hypotheses together so that we can report our results more concisely. However, this resulted more confusion. Thus, we have re-stated our hypothesis so that we can either support or reject them. Now, the hypothesis reads as follows:
1. Correlation between TFH and navigation accuracy vs. TFH and RMS error: On both easy and difficult route sections,
   a. TFH will be positively correlated with navigation accuracy.
   b. TFH will be correlated with RMS error.
2. Correlation between TFH and gaze parameters/scan patterns: On both easy and difficult route sections,
   a. Higher TFH will be associated with shorter OTW dwell times.
   b. Higher TFH will be associated with shorter MAP dwell times.
   c. Higher TFH will be associated with greater frequency of OTW fixations.
   d. Higher TFH will be associated with greater frequency of MAP fixations.
   e. Higher TFH will be associated with a greater number of view changes between OTW and MAP.
   f. Higher TFH will be associated with less OTW scan duration.
3. TFH & subjective assessment:
   a. TFH will have a stronger correlation with the pre-survey than with actual navigation accuracy.
   b. TFH will have a stronger correlation with the post-survey than with actual navigation accuracy.

12. Next section: HYPOTHESIS 2. I am a bit confused by this paragraph. The authors state here that "Spearman's correlation analysis supported our hypothesis on the association between TFH gaze parameters"; although, they present no data supporting this. A table or figure with the data showing this would be helpful. The relationship between TFH and dwell time is shown in Table III; however, the authors make no reference to that until later in the context of "on-track" vs. "off-track". If this is where the data they are referring to is presented then this table should be referred to at that time.

We have clarified the section with re-organized hypotheses. We also added Table II Spearman’s correlation between TFH and gaze parameters for better understanding.

13. In the same paragraph the authors state: "These results indicate that pilots with more TFH showed a more efficient scan characterized by shorter overall dwell, shorter median OTW dwell, less number of fixations per OTW and more number of OTW-MAP view changes." Again, where is the data shown here? Also, was this a statistically significant finding? It was also mentioned in the next sentence that THF x gaze parameter interactions were found. Again, it would be extremely helpful to have a table illustrating the data.

As stated above, we have clarified the section and added Table II. We expect these additional information will help readers better understand our results including TFH x gaze parameter interactions.

14. Results section, "HYPOTHESIS 3" Paragraph. First sentence. How can a hypothesis be "partially supported"? Again, one either accepts or rejects the null hypothesis of no significant difference based on the statistical analysis of the data. There is no “partial” to it. The same applies to the next sentence: "TFH hours were marginally associated with less differences?". In the same sentence it clearly
states that there was no significant association (P<.1). There is either a statistically significant relationship or not.

Please see our response to item #11. We have changed the first sentence to “not supported” and removed a sentence stating marginal association from the main text.

15. Last paragraph of Results. The authors state that they did not conduct statistics on their data presented in Table 3 because of their own admitted small sample size. Then why report the data? One does not pick and choose the data to conduct statistical analysis on- that represents the worst form of bias (and there seems to be considerable bias in this study).

Returning to our points in response to item #11, we emphasize that we were very careful here not to mislead readers into believing we had conducted some sort of confirmatory analysis when, in fact, no hypotheses were stipulated prior to the experiment in this area. Rather, Table III (Table IV in updated manuscript) is presented as exploratory analyses in order to try to better understand the somewhat surprising results and to provide hypothesis formation for future studies. In so doing, those future studies will either confirm or reject these findings, as appropriate, and we like to think we have no particular bias either way.

16. Discussion- First paragraph. This paragraph is simply a re-statement, already presented in the results. This section should be a “discussion” of the results, not a repetition of them. Recommend deleting this paragraph.

We have deleted the first paragraph of the Discussion as suggested.

17. Discussion: 2nd Paragraph. The authors offer possible explanations as to why they found a lack of a relationship; however, they make no mention of the most likely reason- that the study is underpowered, and their sample distribution may be skewed. Navigation performance should improve with experience. It is also possible that if the sample distribution is skewed toward more experienced pilots that they have "plateaued" in their navigation skills say for example (1500-200hrs) and this might explain why a significant difference was not found- again secondary to having a small sample size.

We have included the following sentences in the 2nd paragraph: “First, there is the possibility of this being an underpowered study due to a small sample size. Post-hoc statistical power analysis (3, 14) showed that the power of correlation analysis ranges between .51 and .76 given sample size = 12, α = .05, and observed ρ= .52 ~.69.” The sample demographic has been discussed in response to the reviewer’s main concerns #2.

18. The last paragraph of the discussion states the conclusion that "TFH predicted gaze parameters but ?. was too crude to use as a measure of expertise for task specific activities". Again, based on the limited sample size, this CANNOT be concluded.

We have deleted this phrase from the manuscript.

19. The authors state in paragraph 5 of the Discussion that their sample size was too small to reach any conclusion. And that "the statistics suggest future hypotheses
to be tested with larger sample sizes." Why wasn't a power study done to tell this before the study was undertaken?

A power study has been included as stated before.

We highly appreciate to Reviewer 1 for constructive comments.
Answers to the Comments from Reviewer 2

Reviewer #2: SUMMARY:

The manuscript builds upon the previous work of Sullivan, et al (2011) assessing the correlation between 'expertise' and gaze measurements, navigation accuracy, and subjective assessment for simulated helicopter overland navigation.

GENERAL COMMENTS:
The paper is very good and the topic apropos. It is my hope that the authors' attention to the following comments might make the manuscript better and certainly more applicable to a wider audience. The authors make the case in the Introduction that it is important to understand how 'experts' become expert (e.g., "do they do things in a qualitatively different way, by perhaps sampling different sources of information?"). In understanding this "why," perhaps we can do better with teaching, training, and standardization. This is indeed, an important issue, but my major criticism of the article is in defining (and conducting data analysis based on) their definition of 'expert.' The authors use TFH (as is often the case for a measure of expertise), but there are obvious limitations (e.g., one would not expect a pilot with 1,500 hrs in a traffic pattern or only flying IFR airways to be an expert in map-only overland navigation). So, when looking specifically at overland navigation, why is it important to determine that TFH was associated with certain gaze patterns if the pilot wasn't able to navigate any better than a novice? It seems that if you wanted to know what gaze patterns are associated with an 'expert' in overland navigation, define an 'expert' as those who navigate the best instead of simply having the most flight hours.

We agree with the reviewer that using TFH as a measure of flight expertise has limitations, although it is often an index for a measure of expertise – particularly for tasks such as overland navigation. However, as Reviewer 1 pointed out, our sample of military subjects is homogenous for several aviation experiences that could be used as markers of expertise: they went through very similar training, were all IFR rated, and all had overland navigation experience. We therefore felt that, in addition to being a typical measure of expertise, TFH would provide a wider range of expertise than other aspects of aviation experience for this particular sample. It also is important to note that we found a similar pattern of results when number of overland hours was used as a measure of expertise as described in the Discussion.

The reviewer’s point that defining an expert by those who performed best is intriguing, and exactly pinpoints why we conducted exploratory analyses between those who performed the best vs. those who performed the worst. These exploratory analyses will help us in formulating future hypotheses and studies. However, in this study, our a priori hypotheses were based on a definition of expertise as TFH. We therefore, feel that it would be inappropriate to change our hypotheses post hoc.

SPECIFIC COMMENTS:

- Abstract: define RMS before using the acronym. In fact, check all acronyms to ensure they are defined at least once prior to using, for I think there are others in the manuscript besides RMS (e.g., OTW at the end of the introduction).

We have added the RMS definition “root mean square” and OTW definition “out-the-window” in the abstract and the first time it appeared in the main text.
- Introduction, para 1: the issue of RMS with navigation deserves more attention: sometimes a pilot MUST stay on a very exact course (e.g., to avoid surface-to-air threats, remain within the confines of a specified flight corridor, or deconflict with other traffic), while sometimes he/she may deviate (e.g., follow a train feature for ease of out-the-window navigation). So a "?different measure of expertise beyond RMS error is needed" may be true for the later, but not the former. What was the specific guidance given to your subjects?

Because our experimental tasks included only overland navigation tasks and no other tasks associated with specific threats were included, they were not required to stay on a very exact course. However, their main goal was to navigate by going through pre-planned waypoints, which means they didn’t need to stay right on the waypoint but if they were too far away from the planned route, then it would be hard to stay oriented.

- Introduction, para 2: the discussion regarding "why pilots' performance differs by expertise level?" also might deserve a bit more explanation. There are many types of performance, and I would argue that the skills required are likewise varied. For example, there is a big difference between simple aircraft control (right altitude/airspeed/attitude) and mature judgment and decision making when it comes to "performance."

The reviewer correctly distinguishes different types of pilots’ performance. We have clarified that sentence by specifying the type of performance to which we are referring: “In order to better explain why pilots’ overland navigation accuracy differ by expertise level and to find cues for assessing their cognitive states, we suggest observing human behaviors (e.g., where they look) which influence their performance (e.g., how they navigate).”

- Introduction, para 2: at the conclusion of the paragraph, I would argue your comment that "no standardized scan training has been systematically constructed, yet" is incorrect. In fact, training on scan pattern is regularly taught to helicopter pilots, but perhaps our teaching is insufficient?

The training tells you what type of features you should look at (the features that are easier to distinguish). Training also tells you what gauges to use and how to orient and reference the map. It also tells you generally how much time to spend looking where. So 'yes' some level of scan pattern training is covered. The question is: how specific is the training – could it be improved with some model and insight into expert performance. Thus, we have removed “no standardized scan training has been systematically constructed, yet” from the text and added the following: “Standardized methods and scan patterns can be described to students at a high level and in general terms, however, actually assessing the appropriateness of a student's scan in relation to the in-situ training environment and their performance is not well supported. There is little support for instructors to provide carefully tailor feedback specific to a pilot and the immediate training environment.”

- Introduction, para 3: this paragraph seems identical in many respects (as in ? cut and paste) to the Sullivan, et al (2011) manuscript?

It was our inattention. We removed the identical paragraph and it is now replaced as “See Sullivan et al. (15) for descriptions of eye-tracking studies in several domains including aviation
(1,8), ground transportation (13), different cognitive states (11), and visual processing load (16)."

- The term "expert" is used throughout the introduction without defining?TFH? No history of accidents? Years of flying? Expert as determined by a standardization check flight? Etc.

We have now clarified how expert pilots were defined in previous work in the first paragraph of the Introduction: “Expert pilots, defined by total flight hours or FAA ratings, consistently perform these tasks better than less experienced pilots (1,9,12)”

- Introduction, para 4: why are helicopter overland navigation tasks more cognitively demanding and complex than fixed-wing tasks? Because helicopters use maps while FW uses airways? Then the difference is the 'type' of navigation, not in the aircraft platform.

We have included the following sentences in the paragraph: “The distinguishing characteristic that makes it cognitively demanding is height above terrain. Given the altitude of helicopters above ground, pilots rely much more on terrain relief than their fixed-wing counterparts. From higher altitude at which fixed-wing pilots fly, the visible terrain more closely matches the map representation, whereas helicopter pilots rely more on terrain relief at lower altitudes. Also, from higher altitudes more features are in view for a longer time.”

- Introduction, hypothesis: what do you mean by "efficient" scan pattern in #2? Shorter median dwell? Fewer number of view changes? Etc? Why is it important to have a more "efficient" scan pattern (who cares?) if the pilot's navigation isn't any better? Again, why not see who performs the best? That might then be defined as the "best" scan pattern for overland navigation. Scan is more of surrogate endpoint, but the desired endpoint for pilots (and related issues of teaching, training, and standardization of pilots) is better overland navigation.

We have changed hypothesis 2 to read as follows:

2. Correlation between TFH and gaze parameters/scan patterns: On both easy and difficult route sections,
   a. Higher TFH will be associated with shorter OTW dwell times.
   b. Higher TFH will be associated with shorter MAP dwell times.
   c. Higher TFH will be associated with greater frequency of OTW fixations.
   d. Higher TFH will be associated with greater frequency of MAP fixations.
   e. Higher TFH will be associated with a greater number of view changes between OTW and MAP.
   f. Higher TFH will be associated with less OTW scan duration.

Regarding the reviewer’s comments on hypothesis 2, we formulated our hypotheses regarding eye scan patterns and performance based on previous research indicating that expert pilots’ efficient eye scan is associated with better flight control and faster reaction times. The relatively poor performance of the pilots on the difficult route section, regardless of TFH or overland hours, was surprising.
The goal of our research is to improve training for challenging aviation tasks by providing instructors with real-time information regarding what the trainee is thinking. A large body of research demonstrates that eye scan parameters successfully predict different cognitive states, and we have found them to be able to detect an underlying cognitive strategy specific to overland navigation. Although navigation performance provides a marker of who completed the task well, it does not provide insight into the strategies that the person used to complete the task successfully.

- **Methods:** why 12 subjects? Was there a power calculation? Are these just the same 12 subjects used for the previous Sullivan article?

They were the same 12 subjects used in the previous Sullivan et al. (2011) article. Sullivan et al. (2011) analyzed the data only up to the 6th waypoints and did not consider route difficulties. This paper extends the study considering route difficulties and pilots’ subjective assessments. Ad hoc power analysis was calculated and included as follows in Discussion: “Post-hoc statistical power analysis (3, 14) showed that the power of correlation analysis ranges between .51 and .76 given sample size = 12, $\alpha = .05$, and observed $\rho = .52 \sim .69$.”

- **Methods:** The third sentence mentions "fourteen subjects" versus 12 in the first sentence--where two excluded?

We have corrected fourteen to twelve.

- **Equipment:** might be worth having a picture of the experimental set up for those unfamiliar with the equipment (or reference the Sullivan article if there is a picture there and the identical set up was used).

We agree with the reviewer’s concern. As far as we understand, the journal does not allow authors to include a photo of the experimental set up. If the editor allows us, we can include one of the experimental setup pictures or an experimental diagram shown below.

![Experimental setup: OTW display, MAP display, Cockpit, faceLAB eyetracking stereo cameras and infrared sensor, faceLAB control panel. Flight control stick is now shown in the figure.](image-url)
Experimental setup diagram.

- **Navigation task:** what do the thicker versus thinner lines mean connecting the waypoints in Figure 1?

The two thick lines indicate wp2-4 and wp5-7, which correspond to easy and difficult route sections. We have included following in the text: “The two thick lines indicate the easy route section (wp2-4) and difficult route section (wp5-7).”

- **Navigation task:** I think the issue that the pre-study determination of “easy” versus “difficult” deserves more attention. What are the limitations of using only a single SME to establish this? What criteria did the SME use? What makes him/her an SME? Where there any objective criteria used? This is important, as some results are significant for the easy versus hard sections. It is noted that RMS error increased from easy to hard and subject estimation also seemed to agree (which is reassuring), but how this was defined a priori is important, as it determines how you group data for the analysis.

We did informal trials with several SMEs as part of preliminary studies. For this study, the SME selected routes based on number, prominence and length of time that key landscape features would be in view. Segments of the routes were selected based on number of salient features. SME would then invite experienced navigators to fly the route and informally rate the difficulty. During these trial runs, the pilot subjects would describe the features they expected to be prominent from preflight map study. In flight they would describe their thought process and the landscape features they were relying on. This gave an indication of the degree to which landmarks along the route could be differentiated. Post-flight subjective assessment of the difficulty also provide insight.

- **Navigation task:** where the subjects told they must not deviate from the specified course (presumably so if you assessed “on-track”) or where they free to deviate as long as they hit the waypoint?
We have included the following in the text: “They were free to deviate as long as they remained oriented. Pilots were considered on-track when they stayed within .5km from the wps and off-track when deviated more than .5 km.”

- Procedure: the route looks identical to the one used for the 2011 Sullivan paper. If these were the same subjects (see above), was there a learning affect by flying the exact same route again? Or, was this a novel route for the subjects? If so, state.

Yes. The route is the same and the data came from the same experiment – not the repetition of it. The 2011 paper only discussed first half of the data. Thus, no learning effect needs to be considered.

- Results: if you specify an alpha significance of 0.05, why asterisk nonsignificant results (e.g., p<0.1 in Table II)? It just makes it more confusing.

We removed the asterisks.

- Table I and III title: spell out DV (presumably for Dependent Variables)

Yes. We spell out DV into Dependent Variables in Table I and III, which changed to Table I and IV in the updated manuscript.

- Results: the use of the terms "navigation accuracy" versus "navigation performance" is unclear. "Performance" in the Methods section is defined using the term "accuracy." Are you using these terms interchangeably? If so, suggest sticking to one term. If not, explain the difference in methods.

Yes, we were using the terms synonymously. We revised the text to only use the term “navigation accuracy” to avoid and confusion.

- Results: (similar to general comments) why is it important that pilots with more TFH had a more efficient scan pattern (surrogate endpoint) if they couldn't navigate any better (desired endpoint)? Of course, I'm hinting at Table III. Recommend putting sample size/subject #'s in Table III in addition to the text.

We have included sample size in Table III (Table IV in the updated manuscript.)

The goal of our research is to improve training of challenging aviation tasks by providing instructors with real-time information regarding what the trainee is thinking. A large body of research demonstrates that eye scan parameters successfully predict different cognitive states, and we have found it to be able to detect an underlying cognitive strategy specific to overland navigation. Although navigation performance provides a marker of who completed the task well, it does not provide insight into the strategies that the person used to complete the task successfully.

The above paragraph has been added as third paragraph of Introduction to discuss why we claim it is important to understand scan parameters than just monitoring navigation performance.
Clarify "Three subjects were in the on-track group and two subjects were in the on-track group each." Is that 3 for easy and 2 for hard? It was somewhat disappointing for me as a reader that the issue of looking at performance on the test was almost an afterthought (exploratory analysis), as this was the result in which I was most interested.

We corrected the typo and now the sentence reads “Three subjects were in the on-track group and two subjects were in the off-track group.” In the beginning of the paragraph, it says “subjects were divided into two groups according to their navigation accuracy (on-track vs. off-track) in both route sections.” We italicized both to make it easily understood to the reader that the three subjects were on-track in both easy and hard routes whereas the two subjects were off-track in both routes.

We understand the reviewer’s disappointment. It was exploratory because of small sample size (three and two). We actually thought to examine the correlation between performance and gaze parameters rather than that between TFH and gaze parameters. However, we cannot really control participant’s navigation performance –whether they get lost or not- as we controlled route difficulty. One possible idea for a future study is to design a medium-difficulty route so that about half of participants stay on-track whereas the other half get off-track.

That said, we intend to do follow-on experiments to more formally assess some of these exploratory results.

- Discussion: in the paragraph following Figure 2, how are you defining "subject expertise" here? TFH? If your analysis showed that TFH was not predictive of navigation accuracy or RMS error (for either easy or difficult routes), then why would the map scanning strategy of these so-called experts (who did not perform any better on the navigation task) in Figure 3 be of interest? I would be more interested in the scan patterns of those who performed the task best.

Figure 3 showed OTW scan for subjects who were on-track. OTW scan is highly subject to helicopter flight trajectory, thus plotting off-track subjects’ OTW scan does not provide a useful comparison.

- Again, very interesting topic and a good manuscript. If I have mentioned items or had questions that were, in fact, stated, consider that it could be made clearer.

We very much appreciate Reviewer 2 for constructive comments.
Answers to the Comments from the Editor

***

COMMENTS FROM THE EDITOR

TITLE
We prefer to start with important words rather than stock phrases such as “Comparison of?” or “Effects of?” Length is limited to 100 characters and spaces. Please make the change in both Editorial Manager and the manuscript.

The title has been changed to “Pilot Performance: Assessing How Scan Patterns & Navigational Assessments Vary by Flight Expertise”

ABSTRACT
The abstract needs to be 250 words or less. Please include some descriptive numerical results but not p-values.

The abstract is 250 words. We have included descriptive numerical results.

INTRODUCTION
Please remove subheadings.
We do not allow strings of four or more citations in the text because they are of limited use to readers; additional information should be given about the importance of each reference or smaller groups of them. Or perhaps a review article could be cited instead. Any reference that appears only in such a string and is never discussed in detail should be considered for deletion.

The updated manuscript does not have subheadings in Introduction. Strings of four or more citations in the text has been removed from the manuscript by just keeping properly-discussed references.

METHODS
There are too many subheadings. Please use standard subheadings such as subjects, equipment, procedure, and statistical analysis.

We have removed non-standard subheadings and the updated manuscript has only standard subheadings.

***

REQUIRED FORMATTING CORRECTIONS
(checklist courtesy of Debra Sventek):

MANUSCRIPT FORMAT:
Please place page numbers in the upper right hand corner of each page. Remove any other information that appears in the header.

FIGURES:
PLEASE ASSURE THAT YOUR FIGURES MEET THE ASEM’S QUALITY STANDARDS FOR PUBLICATION AND THAT DIGITAL ART CONFORMS TO THE APPROPRIATE RESOLUTION, COLOR MODE, AND FILE FORMAT.
PLEASE GO TO OUR HOME PAGE, HTTP://ASEM.EDMGR.COM AND SCROLL DOWN TO "DOCUMENTS TO DOWNLOAD." CLICK ON "DIGITAL ART GUIDELINES" - DOWNLOAD AND BE SURE YOU HAVE FOLLOWED THE INSTRUCTIONS IN THIS DOCUMENT.

The figures in your paper appear in color. Please indicate your intent to pay the charges for the color figures (see below), OR indicate to us IN WRITING your agreement that they be printed in black and white.

We agree that all figures be printed in black and white.

Color Figures require submission of a signed author agreement to pay associated charges. Color figures in print cost $900 for the first figure and $450 for each additional page of color. (Go to our home page, http://asem.edmgr.com, scroll down to Documents for Download, Agreement to Pay Charges for Figures).

ON LINE MATERIAL:

Please submit material to go on line as a separate file that accompanies the manuscript so that the reviewers can see it. In the main text you should use a phrase such as "detailed information can be viewed on line" [CITE AUXILIARY MATERIAL HERE]. The capitalized phrase in brackets will alert the copy editor to insert the exact web address.

After the paper is accepted, the auxiliary material will be sent directly to Ingenta as separate electronic file for uploading to the web. If you have references that are only used in the auxiliary materials, then they will need a separate reference list.

There is a charge for on-line material that is not part of the published article. The charges are $40/page or portion thereof for tables; $40/figure for B/W; $60/figure for color. Please go to our home page, http://asem.edmgr.com and scroll down to documents to download. Click on "Agreement to pay extra charges." Print the form, sign the form and either fax or scan and e-mail the form.

***

SUBMISSION OF REVISED MANUSCRIPT

1. Please review the editorial comments together with the INSTRUCTIONS FOR AUTHORS located on the menu bar of our web site. Careful attention to details will speed further processing.

2. TO REVISE YOUR MANUSCRIPT, first use your computer to prepare revised electronic files in accordance with the suggestions of the editor and referees. PLEASE REMEMBER TO UPDATE THE INFORMATION IN THE MANUSCRIPT'S TITLE PAGE.

**NOTE: You will not be able to change the separate Abstract that you originally pasted into the text box.**

3. YOU SHOULD ALSO PREPARE A LETTER describing point-by-point the changes you have made. If you elect not to comply with some of the suggestions, please state the justification for your position.

4. WHEN ALL OF THE FILES ARE READY, log in to http://asem.editorialmanager.com. Click on AUTHOR. Go to SUBMISSIONS NEEDING REVISION and click on SUBMIT REVISION. You will see several screens like those you used to submit the original manuscript. Carefully
enter all required changes and attach your revised files. Be sure that the TITLE and list of AUTHORS exactly match those in the manuscript itself.

YOU MUST SEND THE FOLLOWING FILES:
--> Manuscript which includes a Title Page, Abstract, Text, Acknowledgement (if any), References, Tables and Figure Legends (if any).
--> Figures (if any) as separate files named Fig 1, Fig 2, etc.
--> Letter of Explanation describing your response to the reviews and any editorial corrections.

5. Then click on SEND to build a revised PDF. You will need to VIEW and APPROVE the revised PDF for submission to the journal.

REQUIRED FORMATTING CORRECTIONS
(checklist courtesy of Sarah Pierce-Rubio):

REFERENCES:
After completing all revisions to the text, please verify that all listed references are cited in the text.

They have all been verified.
Pilot Performance: Assessing How Scan Patterns & Navigational Assessments Vary by Flight Expertise

Ji Hyun Yang, Ph.D.¹, Quinn Kennedy, Ph.D.¹,²,
Joseph Sullivan, Ph.D.¹, and Ronald D. Fricker, Jr., Ph.D².

¹ MOVES Institute
Naval Postgraduate School
Monterey, CA, 93943

² Operations Research Department
Naval Postgraduate School
Monterey, CA, 93943

Correspondence concerning this article should be address to Ji Hyun Yang, Ph.D., MOVES Institute, Naval Postgraduate School, CA, 93943, Email: jyan1@nps.edu

Running Head: Scan patterns and navigation
ABSTRACT

Introduction: Helicopter overland navigation is a cognitively complex task that requires continuous monitoring of system and environment parameters and many hours of training to master. This study investigated the effect of expertise on pilots’ gaze measurements, navigation accuracy, and subjective assessment of their navigation accuracy in overland navigation on easy and difficult routes. Methods: Twelve military officers who ranged in flight experience, as measured by total flight hours (TFH), completed a simulated overland task. They first completed map study of a route that including easy and difficult route sections, and then had to ‘fly’ this simulated route in a fixed-base helicopter simulator. They also completed pre-task estimations and post-task assessments of the navigational difficulty of the transit to each waypoint in the route. Their scan pattern was tracked via eye tracking systems, which captured both the subject’s out-the-window (OTW) and topographical map scan data. Results: TFH was not associated with navigation accuracy and RMS (root mean square) error for either route section. For the easy routes, experts spent less time scanning out the window ($\rho=-.61$), had shorter OTW dwell ($\rho=-.66$), For the difficult routes, experts appeared to slow down their scan by spending as much time scanning out the window as the novices, while also having fewer MAP fixations ($\rho=-.65$) and shorter OTW dwell ($\rho=-.69$). However, TFH was not significantly correlated with more accurate estimates of route difficulty. Discussion: This study found that TFH did not predict navigation accuracy or subjective assessment but was correlated with some gaze parameters.

Keywords: expertise; scan strategy; cognition; subjective assessment
INTRODUCTION

A common goal in training is to teach novices to behave and think like experts so that they can more quickly attain satisfactory levels of performance and decision making skills (10). In aviation, performance generally is assessed by level of flight control, typically defined by RMS (root mean square) error of flight trajectory, accuracy of flight decisions, and depth of understanding of the issues surrounding the decision. Expert pilots, defined by total flight hours or FAA ratings, consistently perform these tasks better than less experienced pilots (1,9,12).

Helicopter overland navigation is a particularly challenging aviation task for trainees and instructors as it entails additional cognitively demanding tasks above and beyond flight control. Furthermore, RMS error of flight trajectory does not predict expertise levels in helicopter overland navigation (15) as it does in other aviation tasks. This is because helicopter pilots are trained to adapt their between-waypoints navigation solution based on current observation. For example, pilots may elect to deviate from a straight-line connection between waypoints to take advantage of a guiding feature that was not readily apparent in pre-flight planning. (15). Thus, in training helicopter pilots, a different measure of expertise beyond RMS error is needed.

Another limitation of using RMS error as a measure of flight expertise is that it does not provide information regarding experts’ underlying cognitive strategies while flying or how these strategies may change with accrued experience. Currently, little is known about the learning process underlying improvements in flight control and navigation. For example, do experts simply demonstrate more precise control or do they do things in a qualitatively different way, by perhaps sampling different sources of information (1, 7)? In order to better explain why pilots’ overland navigation accuracy differ by expertise level and to find cues for assessing their cognitive states, we suggest observing human behaviors (e.g., where they look) which influence
their performance (e.g., how they navigate). Even for one of the most common causes of mishaps, the breakdown in cockpit scan, developing a good scan strategy has not been given high priority during training (2). Standardized methods and scan patterns can be described to students at a high level and in general terms, however, actually assessing the appropriateness of a student's scan in relation to the in-situ training environment and their performance is not well supported. There is little support for instructors to provide carefully tailor feedback specific to a pilot and the immediate training environment. Thus, the purpose of this study was to attempt to understand underlying cognitive strategies used by experts that aid in superior performance.

The goal of our research is to improve training of challenging aviation tasks by providing instructors with real-time information regarding what the trainee is thinking. A large body of research demonstrates that eye scan parameters successfully predict different cognitive states, and we have found it to be able to detect an underlying cognitive strategy specific to overland navigation. Although navigation performance provides a marker of who completed the task well, it does not provide insight into the strategies that the person used to complete the task successfully.

Among several candidate psychophysiological measures for human cognitive states in real time, eye movements are relatively easy to collect in actual operational environments, and recent eye-tracking technology provides non-intrusive devices to collect ocular data (5). See Sullivan et al. (15) for descriptions of eye-tracking studies in several domains including aviation (1,8), ground transportation (13), different cognitive states (11), and visual processing load (16). Thus, by knowing expert pilots’ scan patterns for different aviation tasks and decisions, training novice pilots can be improved by (1) teaching them how to scan the environment more
effectively, and (2) detecting experts’ underlying cognitive strategies based on their scan pattern; these strategies can then be taught to novices.

Previous aviation studies that used eye-tracking did not investigate expertise and visual scan differences in helicopter overland navigation tasks, which are considered to be more cognitively demanding and continuously complex than fixed wing aircraft operating tasks. The distinguishing characteristic that makes it cognitively demanding is height above terrain. Given the altitude of helicopters above ground, pilots rely much more on terrain relief than their fixed-wing counterparts. From higher altitude at which fixed-wing pilots fly, the visible terrain more closely matches the map representation, whereas helicopter pilots rely more on terrain relief at lower altitudes. Also, from higher altitudes more features are in view for a longer time. Recently, Sullivan et al. (2011) demonstrated that when pilots were on track during an overland navigation task, flight expertise predicted gaze parameters and scan management skills but did not predict flight performance measures, such as RMS error. However, it is unknown whether this pattern of results also occurs when pilots are faced with more difficult navigation routes in which they are more likely to be off track. It also is unknown how well experts’ estimates of route difficulty match their actual performance. If experts know ahead of time which sections of the route are difficult to navigate, they may alter their visual scan strategies accordingly during these sections. From a training perspective, understanding expertise differences in the link between navigation accuracy, pilots’ subjective assessment of how they are doing, and visual scan patterns would greatly enhance current training procedures. We thus focused on improving our understanding of cognitive processing associated with helicopter overland navigation by analyzing gaze measurements, navigation accuracy, subjective estimation and assessment, route difficulties, and expertise level of pilots.
In this study, we have designed overland navigation tasks in a flight simulator integrated with eye-tracking systems and performed human-in-the-loop experiments with pilots who ranged in total flight hours. The simulated navigation tasks entailed ‘flying’ to 12 waypoints depicted on a map. In our previous work (15), we examined only waypoints 2 – 5, in which all pilots were on track. In this study, we extend upon these results by also examining expertise differences in actual and self-reported performance for waypoints 5 – 7, which were rated as much more challenging than waypoints 2 – 5. The Results section focuses on these two route sections; notable points from other waypoints data are described in Discussion for an organized reporting.

Pre-experiment, we made the following alternative hypotheses for helicopter overland navigation tasks regarding route difficulty and expertise represented by TFH:

1. Correlation between TFH and navigation accuracy vs. TFH and RMS error: On both easy and difficult route sections,
   a. TFH will be positively correlated with navigation accuracy.
   b. TFH will be correlated with RMS error.

2. Correlation between TFH and gaze parameters/scan patterns: On both easy and difficult route sections,
   a. Higher TFH will be associated with shorter OTW (out-the-window) dwell times.
   b. Higher TFH will be associated with shorter MAP dwell times.
   c. Higher TFH will be associated with greater frequency of OTW fixations.
   d. Higher TFH will be associated with greater frequency of MAP fixations.

1 Parts of the result in this paper will be presented in Human Factors and Ergonomics Society 2012 Annual Meeting (19).
e. Higher TFH will be associated with a greater number of view changes between OTW and MAP.

f. Higher TFH will be associated with less OTW scan duration.

3. TFH & subjective assessment:
   a. TFH will have a stronger correlation with the pre-survey than with actual navigation accuracy.
   b. TFH will have a stronger correlation with the post-survey than with actual navigation accuracy.

As part of this project we developed a visualization tool, Flight and Eye Scan visualization Tool (FEST), designed to provide a representation of spatial and temporal correspondence among features scanned in OTW (3D) and Map (2D) views in relation to the actual aircraft location (15).

**METHODS**

**Subjects**

There were 12 male military personnel, 29 to 40 years of age who participated in the study. The minimum skill requirement for the study was completion of at least one overland navigation class. Among the twelve subjects, three subjects were helicopter flight instructors and two subjects had other navigation-related instructing experience but no flight experience. Expertise was defined in terms of total flight hours (TFH), where higher TFH values were used as a proxy measure indicating increased pilot expertise. TFH varied from 0 to 3,100 hrs (avg = 1,488 hrs, std = 1,104 hrs) and overland-flight-hours varied from 0 to 2,500 hrs (avg = 612 hrs, std = 853 hrs). Seven subjects were from the U.S. Navy, two from the U.S. Marine Corps, one
from the U.S. Air Force, one from the U.S. Army, and one was unknown. No special neurological, visual acuity, or spatial ability tests were performed. The study was approved by the Naval Postgraduate School (NPS) Institutional Review Board. Subjects were recruited via an e-mail advertisement to NPS e-mail account holders. All the subjects were given written informed consent to participate, with the right to withdraw at any time.

**Equipment**

The basic experimental apparatus included the flight simulator X-Plane 8.6, a 46” wide screen to present OTW view, a 40” wide display for the map and instrument display, two stereo cameras and associated faceLAB 4.6 software for collecting eye data, and cockpit-style seat with side-mounted joystick. Data from the flight simulator were sent to an Image Generator (IG), which provided an OTW and a map view combining an OpenSceneGraph terrain model of Twentynine Palms, CA.

The helicopter was designed to be on an automated terrain-following mode at fixed 150' above ground level (AGL) flying at 60 knots. However, the pilot was able to control the heading of the aircraft using the lateral control of the joystick. The joystick pitch control (up/down) was programmed to change the up/down view of the OTW, not the actual pitch angle of the aircraft. The display presented a 1:50,000 topographical land map typically used for flight planning and execution. The map was fixed in position about the pair-wise mean of the waypoints, whereas the orientation of the map was synchronized to the aircraft’s heading to maintain a track-up orientation. The bottom portion of the screen contained instruments to support navigation: the left-most instrument display was a compass typical of legacy Navy H-60 (SH/HH-60F/H) displays. To the right of the compass display were typical barometric and radar altimeters. The rightmost portion of the instrument cluster contained a digital-style elapsed time clock. We had
two separate faceLAB systems (two sets of stereo cameras with 12.5 mm lenses, three Infra Red strobe lights) for tracking eye gaze for OTW and map displays.

**Procedures**

The navigation task was to fly over 12 waypoints (indicated as black circles on Figure 1), after studying the area utilizing Falcon View flight planning software, a system widely employed by diverse communities within DoD. The first waypoint (wp) is located slightly south of the map so it is not shown in the figure. Each waypoint pair has a “doghouse” that indicates (from top to bottom): the next waypoint number, the recommended heading to reach that waypoint from the previous one, the distance between waypoints, and the amount of time it takes to traverse the distance assuming a speed of about 60 knots. They were free to deviate as long as they remained oriented. Pilots were considered on-track when they stayed within .5km from the wps and off-track when deviated more than .5 km.

Waypoints were set very close together, and the terrain tends to be ambiguous, so subjects needed to make course corrections based on visual cues from both the OTW and map screens (their goal being to bring their perceived location closer to their actual location). The task was purposely designed so that some legs would be more challenging than others. The difficulty of each leg was assessed by a subject matter expert (SME) who designed the whole route. The SME determined that the legs from wp 2 – 4 were easy, whereas the legs between wp 5 – 7 were difficult. We refer to wp 2 – 4 as the easy route section and wp 5 – 7 as the difficult route section. The two thick lines indicate the easy route section (wp2-4) and difficult route section (wp5-7).

[Fig. 1 here]
A pre-task questionnaire asked subjects to indicate the level of navigation difficulty for each of the 11 legs on a scale from 1 to 5, in which 1=completely trivial, 2=somewhat difficult, 3=moderately difficult, 4= very difficult, and 5=not at all possible. The scale was represented as a straight horizontal line and participants were told to draw a vertical line to indicate their level of perceived difficulty. The numbers 1-5 were evenly dispersed above the line. Thus, perceived level of difficulty was calculated by measuring the marker distance from the left most point of the scale (line) divided by the total length of the line. Multiplied by 100, self reported perceived difficulty was thus quantified on a percentage scale.

A post-task questionnaire asked subjects to indicate the level of navigation difficulty that they experienced for each of the 11 legs on the same scale used for the pre-task questionnaire. A demographic survey had questions regarding subjects’ age, gender, branch of military service, total flight hours, overland navigation hours, days since last flight, instructor experience, and years of aviation experience. After a brief introduction, subjects were asked to read and sign an informed consent form. They then completed the demographic survey. The next step was a calibration of faceLAB stereo cameras to verify that the visual scan data was usable (error less than 3 degrees) before subjects started the navigation tasks. Subjects were asked to sit in the simulator chair, where eye-tracking cameras had been mounted in between the chair and the simulator screen. Once the calibration was done, the simulated flight environment was explained to the subjects (e.g., altitude and speed maintained by Autopilot, forward/backward movement of the flight stick controls the view of the helicopter, the digital map stay oriented automatically, etc.) and then they flew a practice route. The practice run took about seven to eight minutes, giving subjects enough time to get familiar with the simulated environment and the simulator itself.
Following the calibration phase and equipment familiarization navigation route exercise, subjects were briefed on the main navigation route (CleghornWest, Figure 1) for up to 20 minutes. After the brief, subjects completed the pre-task questionnaire and then were directed back to the flight simulator and evaluators re-verified calibration. Subjects then flew the main route (6 min long) while evaluators collected eye-scan data and flight information. If a subject went too far off course, the experimenter would verbally intervene, giving them a course to guide the subject back to a waypoint. Subjects then completed the post-task questionnaire and were debriefed. Total experiment time varied from one hour to 1.5 hour.

**Statistical Analyses**

We used Spearman’s rank correlation to see if expertise is associated with flight performance and/or visual scan characteristics. We used a significance level $\alpha=0.05$ for determining whether to reject the null hypotheses, though here we report all the $p$-values for those who might prefer an alternate significance level. For a regression analysis on the easy route section between TFH and gaze parameters, we refer the reader to Sullivan et al. (15).

The main outcome measures for the flight and navigation performance were 1) RMS error of the flight trajectory and 2) navigation accuracy, i.e., whether pilots were on-track (within .5 km from the wps) or off-track (deviated more than .5 km). The RMS error is defined as

$$\text{RMS error}_{k,k+1} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i^a - x_i^o)^2},$$

where for $n$ data points between waypoints $k$ and $k+1$ $x_i^a$ is the actual flight position and $x_i^o$ is the corresponding reference trajectory point for the $i_{th}$ point.
Navigation accuracy was assessed on the easy route section and the difficult route section respectively. Navigation accuracy was quantified as a 2 if the pilot was on-track for both legs of the section (e.g., on-track for wp 2-3 and wp 3-4 in wp 2-4), 1 if the pilot was on track for only 1 leg (e.g., on-track only for wp 5-6 in wp 5-7), and a 0 if they were off-track for both legs. Being on track was determined based on whether or not the subject was closely located (threshold was .5 km) to designated wps and by the subject’s debrief. Navigation accuracy is a relaxed variation of the conventional RMS error, which allows acceptable deviation and captures “good-enough” or “satisfying” characteristics of tracking tasks (7,20), whereas RMS error penalizes any errors deviated from wps.

The main outcome measures for visual scan patterns were 1) median of dwell duration, 2) OTW scan time, 3) number of OTW-MAP view changes, and 4) number of fixation points per unit time. Dwell duration (or the duration of fixations) is calculated as a period between consecutive saccades (12). Because the navigation tasks had two different views (OTW and MAP), the variables, OTW and MAP scan time ratio and number of OTW-MAP view changes, were included to account for how many features pilots scanned per view. Data from faceLAB, X-plane and IG were combined into a text file and all data were processed in MATLAB R2010a. The main outcomes from the survey data were self reported level of navigation difficulty.

RESULTS

Preliminary Analyses. Spearman’s rank correlation is denoted by ρ and the corresponding p-value is shown as P. As would be expected, TFH was positively correlated with overland flight hours (ρ =.64 and P=.02), days since last flight (ρ =.77 and P=.003) and days since last overland flight (ρ =.79 and P=.002) , but not with any other demographic variables, such as age or branch
of service. Results from the pre-task surveys indicated that subjects estimated wps 5-7 (i.e., the difficult route section as determined by the SME) would be more difficult than wps 2-4 (i.e., the easy route section determined by the SME) prior to the navigation task \( (t(11) = 3.163, P < .01) \). After the task completion, subjects still assessed the difficult route as more difficult than the easy route \( (t(11)=8.300, P < .001) \).

Route difficulty affected actual flight and navigation accuracy. As expected, and as can be seen in Table 1, RMS error increased and navigation accuracy decreased from the easy route section to the difficult route section \( (t(11) = 5.171, P < .001 \) and \( t(11) = 3.924, P < .01 \) respectively. Ten pilots were on course for the easy route whereas only three pilots were on course for the difficult route. These results confirmed the SME’s evaluation. Comparing subjects’ pre-task estimation with their post-task assessment, we found that pre- and post-reports were consistent for the easy route, but that the difficult route was under-estimated in the pre-task estimate compared to the post-task assessment \( (t(11)=2.901, P < .001) \). Table I shows mean and standard deviation of each dependent measure on the easy route section and the difficult route section respectively. Dwell parameters in the helicopter navigation tasks were in the range of results previously reported (17). Also, the distribution of dwell duration was skewed to the left. We therefore used the median dwell duration in statistical analyses rather than using mean dwell duration.

| [Table I here] |

None of the gaze parameters were significantly different between the two route sections, possibly due to wide range of variability in all gaze parameters, with the most variability occurring with median Map dwell duration. Of note, the number of fixations per OTW view was more than that of the Map view in both routes (easy route: \( t(11)=3.067, P < .01 \) and difficult
route: $t(11)=3.586$, $P<.005$ and OTW scanning time was more than 50% for both routes. This result indicates that regardless of route difficulty, pilots tend to spend more time looking and fixating OTW relative to the MAP view.

Navigation accuracy was correlated negatively with two gaze parameters on the easy route (median dwell, $\rho=-.45$, $P<.1$; median OTW dwell, $\rho=-.52$, $P<.05$; pilots who were on-track had shorter dwell times on the easy route) whereas no significant correlation was found in difficult route with any gaze parameters. Spearman’s rank correlation coefficients among flight, navigation, gaze, and subjective data were calculated for both route sections and are shown in Table II. Lower half of the table corresponds to Spearman’s rank correlation coefficient between dependent variables in Leg 1 and upper half to that of Leg 2. Navigation accuracy was correlated negatively with RMS error, OTW dwell duration, and post-task route assessment in easy route ($\rho=-.52$, $P<.05$; $\rho=-.52$, $P<.05$; $\rho=-.55$, $P<.05$). On the other hand, navigation accuracy was only correlated negatively with post-task route assessment in difficult route ($\rho=-.53$, $P<.05$). As would be expected, most gaze parameters were correlated with each other on both the easy and difficult routes; for example, OTW dwell and OTW-MAP view changes were correlated negatively in both legs ($\rho=-.66$, $P<.05$ and $\rho=-.69$, $P<.001$ respectively). The subjects' pre-task estimation was correlated negatively with OTW scan duration in easy route ($\rho=-.61$, $P<.05$) and post-task assessment was correlated positively with number of fixations per MAP in difficult route ($\rho=57$, $P<.05$). Pre-task estimation and post-task assessment were correlated negatively in difficult route ($\rho=-.69$, $P<.001$), whereas no correlation was shown for the easy route.

[HYPOTHESIS 1: The experimental data did not support either Hypothesis 1.a regarding the relationship between TFH and navigation accuracy or 1.b regarding the relationship between]
TFH and RMS error. That is, TFH was not a significant predictor of either navigation accuracy or RMS error for both easy and difficult route sections. Of course, failure to prove the alternative hypothesis does not mean the null is true, though we note that the lack of association between TFH and RMS error is consistent with our previous work (15) and a post-hoc power analysis (see the Discussion section) indicates the sample size provided reasonable power to detect correlations similar to those observed between other factors in this study.

HYPOTHESIS 2: The experimental data supported hypotheses 2.a, c, and e on the association between TFH and gaze parameters. Specifically, TFH predicted median OTW dwell, Number of fixations per OTW, and Number of OTW-MAP view changes in both easy and difficult route sections ($\rho=-.66, P<.01$; $\rho=-.62, P<.05$; and $\rho=.59, P<.05$ for easy route sections and $\rho=-.69, P<.01$; $\rho=-.59, P<.05$; and $\rho=-.65, P<.05$ for difficult route sections). Table III contains the correlations between TFH and scan parameters by route section. As illustrated in Table III, these results indicate that pilots with more TFH showed a more efficient scan pattern characterized by shorter median OTW dwell, less number of fixations per OTW and more number of OTW-MAP view changes. Hypothesis 2.b, d, and f were not supported, but TFH-by-gaze parameter interactions (e.g., TFH-by-OTW scan duration, TFH-by-fixations per MAP view) were found. In particular, TFH was negatively associated with OTW scan duration for the easy route ($\rho=-.61, P<.05$), whereas no differences in OTW scan duration were found for the difficult route section. On the other hand, TFH was negatively associated with the number of fixations per MAP view only on difficult route section ($\rho=-.65, P<.05$). The interactions suggest that more experienced pilots make subtle changes to their scan pattern when route difficulty increases, where they spend more time scanning out the window and look less often at the map. In contrast, less experienced pilots do not change their scan pattern when navigation difficulty changes.
HYPOTHESIS 3: The hypothesis on association between TFH and route difficulty estimation was not supported. Regardless of TFH, pilots tended to underestimate the difficult route compared to post-task assessment. Interestingly, pre-task estimation and post-task assessment were negatively correlated for the difficult route ($\rho = -0.69, P < .001$), which indicates that pilots who estimated the leg to be easy/difficult, after completing the navigation task, then assessed it as more/less difficult respectively.

As an exploratory analysis, subjects were divided into two groups according to their navigation accuracy (on-track vs. off-track) in both route sections. The purpose of the grouping was to see if on-track subjects can be characterized differently from off-track subjects in terms of gaze parameters. Table IV shows dependent measures comparison between these two groups. The descriptive statistics suggest differences between the two groups, but we did not conduct statistical analyses due to the small sample size. Three subjects were in the on-track group and two subjects were in the off-track group. The rest of the subjects had a combination of on- or off-track navigation accuracy, thus they are not included in this exploratory analysis.

DISCUSSION

There are a few possible explanations for the lack of a relationship between TFH and navigation accuracy and gaze parameters. First, there is the possibility of this being an underpowered study due to a small sample size. Post-hoc statistical power analysis (3, 14) showed that the power of correlation analysis ranges between .51 and .76 given a sample size = 12, $\alpha = .05$, and observed $\rho = .52 \sim .69$. Second, TFH may not be an accurate measure of
expertise for task specific activities. Even instructor-experienced pilots, which is an alternative proxy for pilot expertise, did not predict gaze and navigation accuracy on both legs. A better measure of overland navigation expertise may be total overland hours, particularly in this cohort of military pilots, some of whom have most of their flight hours over water. However, overland flight hours did not predict gaze parameters better than TFH either. Third, it could be that the difficult routes were very challenging even for the experienced pilots. Evidence supporting this view is that mean level of navigation accuracy for the difficult route was quite low, .62 out of a maximum score of 2.0. Additionally, during the difficult route, more experienced pilots showed a scan pattern that was more representative of a novice scan pattern: longer scan time out the window and fewer fixations. Finally, even the more experienced pilots underestimated how challenging the difficult route would be, suggesting that they were unprepared when confronted with that part of the navigation route.

Other surprising results were that gaze parameters only partially predicted navigation accuracy and changes in route difficulty. Pilots with better navigation accuracy in the easy route had lower median OTW dwell times. As shown in Table I, no significant change was shown in OTW scanning time between easy and difficult route sections. However, increased variability in OTW scanning time during the difficult route could have masked any significant relationship between OTW dwell time and navigation accuracy for this route.

The above results lead to two questions: (1) Can we characterize those pilots who had high levels of navigation accuracy; that is, those that showed task specific expertise? (2) What types of mistakes were pilots making during the flight?

To address the first question, we compared descriptive statistics between pilots who scored 100% on navigation accuracy across all legs and pilots who had very poor navigation
accuracy. Although the sample sizes are too small to reach any general conclusions, the statistics suggest future hypotheses to be tested with larger sample sizes. The high performance pilots are characterized by more THF, more overland hours, lower RMS, shorter overall and map dwell duration, more time spent looking out the window, and more accurate pre-task estimates of route difficulty.

Regarding the second question, whether subjects perceive their whereabouts correctly is critical for successful mission completion. Common frequent visual misperceptions among pilots were observed throughout the study. Some expert pilots successfully located waypoint six and made a 90 degree left turn into a narrow valley toward waypoint seven. However, nine out of twelve pilots missed this narrow valley mainly due to a field of view angle limitation. Once they passed waypoint six without realizing it, another valley appeared on their left. Pilots who missed waypoint six made a left turn into this valley believing they were on track.

As shown in Figure 2, subject 5 missed waypoint 6 and took a left turn into this valley (6'). Then, he flew north of the intended trajectory (7' and 8'), believing he was on waypoints 7 and 8. Initially planned waypoints are shown in black whereas the subject’s estimation is shown in blue. On his way from waypoint 6' to 7', he saw a valley on the right side of the flight heading direction in the OTW scene. If he had been on track (i.e., between 6 and 7), he would have been surrounded by hills and should not have been able to see any saddle or valley and his heading would have been much different. Even though his gaze data showed that he scanned the valley, the pilot did not question his orientation. This information indicates the pilot rejected the visual cues that were not compatible with his current belief, which could not have been correct. Thus, the subject did not question his orientation or status, indicating that he overweighed those visual cues that fit into his mental picture by giving little attention (subconsciously) to cues conflicting
with it. This type of bias, carrying over initial bias, has also been seen in a cognitive task that tapped inductive biases on cultural evolution (6). Cowden et al. (4) investigated the misperception and showed pilots’ perception was wrong 77.86% the time when they were “off-track.”

[Figure 2 here]

Our gaze pattern analysis thus far has focused on temporal aspects of the data, such as gaze duration, number of fixations, etc. On the other hand, we can also study spatial aspects of gaze parameters, e.g., where in the OTW or MAP subjects were looking when navigating. A specific MAP scanning strategy used by experts to maintain course was introduced in Sullivan et al (2011). Figure 3 shows OTW gaze histogram of wp 2-5 depending on subject expertise sorted by TFH. Red cells indicate the location where experts had more fixations than novices, whereas blue cells represent novice pilots gazing on that area more so than experts. The figure clearly shows that where experts and novices looked were different. For example, experts looked on the left side of the travel direction (hilly terrain) while novices looked on the right (plain area) near wp 2. From wp 3 to wp 4, novices tended to stay and look more to the left while experts tended to look more to the right. The OTW gaze location is, of course, highly subject to helicopter trajectory.

[Figure 3 here]

We can conclude TFH predicted gaze parameters but, in this cohort of military pilots, we cannot reach any firm conclusion regarding the association between TFH and expertise. As future work, how an expertise’s scan strategy induced better navigation accuracy and how expert pilots obtained the desirable scan strategy should be studied. We should be able to characterize/predict who will perform task well based on eye gaze pattern, vs. those who have
scan breakdown. This research is particularly important towards preventing CFIT (Controlled Flight Into Terrain) and mid-air collisions while conducting low level VFR operations. Scan strategy also differ by task; therefore a “portfolio” of successful scan strategies by aviation task could be developed.

ACKNOWLEDGEMENTS

This work is funded by Naval Modeling Simulation Office (NMSO) and Office of Naval Research (ONR). We are grateful to Noah Lloyd-Edelman for helping us in calibrating the experimental device and to Michael Day for providing technical support for experimental setup.
REFERENCES


decision making and flight control in a flight simulator. Aviat Space Environ Med 2000;
11. Marshall SP. Identifying cognitive state from eye metrics. Aviat Space Environ Med
2007; 78(5, Suppl.):B165-75.
12. Morrow DG, Miller LMS, Ridolfo HE, Magnor C, Fisher UM, Kokayeff NK, Stine-
Morrow EAL. Expertise and age differences in pilot decision making. Aging Cogn
Human Factors (1972); 14 (4): 325-35.
16. Van Orden KF, Limbert W, Makeig S, Jung TP. Eye Activity Correlates of Workload
17. Velichkovsky B, Dornhoefer S, Pannasch S, Unema P. Visual fixations and level of
attentional processing. In: Proceedings of the 2000 Symposium of Eye Tracking Research
& Applications; 2000 Nov 06-08; Palm Beach Gardens, FL; 79-85.
18. Yang J, Kennedy Q, Sullivan J, Day M. Bayesian modeling of pilot belief and visual
misperception in helicopter overland navigation. In: Proceedings of IEEE International
Conference on Systems, Man and Cybernetics; 2011 Oct 09-12; Anchorage, AL.

Table I. Mean, median and standard deviation (std) of DV (dependent variables)

<table>
<thead>
<tr>
<th></th>
<th>Leg 1 (easy, wp2-4)</th>
<th></th>
<th></th>
<th>Leg 2 (difficult, wp5-7)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>median</td>
<td>std</td>
<td>mean</td>
<td>median</td>
<td>std</td>
</tr>
<tr>
<td>Navigation accuracy (max = 2.0)</td>
<td>.92</td>
<td>1.0</td>
<td>0.19</td>
<td>0.62</td>
<td>0.5</td>
<td>0.22</td>
</tr>
<tr>
<td>RMS error</td>
<td>11.5 ft</td>
<td>9.05 ft</td>
<td>7.8 ft</td>
<td>30.6 ft</td>
<td>30.5 ft</td>
<td>14.2 ft</td>
</tr>
<tr>
<td>Median dwell duration</td>
<td>229.1 msec</td>
<td>215.8 msec</td>
<td>47.3 msec</td>
<td>212.8 msec</td>
<td>208.6 msec</td>
<td>34.1 msec</td>
</tr>
<tr>
<td>Median OTW dwell duration</td>
<td>226.5 msec</td>
<td>227.1 msec</td>
<td>38.7 msec</td>
<td>213.9 msec</td>
<td>207.6 msec</td>
<td>43.1 msec</td>
</tr>
<tr>
<td>Median MAP dwell duration</td>
<td>297.5 msec</td>
<td>224.8 msec</td>
<td>159.0 msec</td>
<td>257.5 msec</td>
<td>230.1 msec</td>
<td>91.4 msec</td>
</tr>
<tr>
<td>Num. of OTW fixations per view</td>
<td>4.1</td>
<td>3.0</td>
<td>2.6</td>
<td>3.3</td>
<td>2.4</td>
<td>1.8</td>
</tr>
<tr>
<td>Num. of MAP fixations per view</td>
<td>1.74</td>
<td>1.79</td>
<td>.65</td>
<td>1.78</td>
<td>1.55</td>
<td>.61</td>
</tr>
<tr>
<td>OTW scanning time</td>
<td>61%</td>
<td>60%</td>
<td>12%</td>
<td>56%</td>
<td>56%</td>
<td>9%</td>
</tr>
<tr>
<td>Num. of OTW-MAP view changes per second</td>
<td>1.35</td>
<td>1.34</td>
<td>.63</td>
<td>1.30</td>
<td>1.20</td>
<td>.56</td>
</tr>
<tr>
<td>Route difficulty Estimation (max = 75)</td>
<td>19.4</td>
<td>19.0</td>
<td>7.4</td>
<td>32.5</td>
<td>37.0</td>
<td>11.9</td>
</tr>
<tr>
<td>Route difficulty Assessment (max = 75)</td>
<td>15.4</td>
<td>13.0</td>
<td>9.9</td>
<td>50.4</td>
<td>47.3</td>
<td>11.1</td>
</tr>
</tbody>
</table>
Table II Spearman’s correlation between TFH and gaze parameters

<table>
<thead>
<tr>
<th></th>
<th>Easy route</th>
<th>Difficult route</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median OTW dwell duration</td>
<td>-.66**</td>
<td>-.69**</td>
</tr>
<tr>
<td>Median MAP dwell duration</td>
<td>-.47</td>
<td>-.02</td>
</tr>
<tr>
<td>Num. of OTW fixations per view</td>
<td>-.62*</td>
<td>-.59*</td>
</tr>
<tr>
<td>Num. of MAP fixations per view</td>
<td>.12</td>
<td>-.65*</td>
</tr>
<tr>
<td>Num. of OTW-MAP view changes</td>
<td>.59*</td>
<td>-.65*</td>
</tr>
<tr>
<td>OTW scanning time</td>
<td>-.61*</td>
<td>-.30</td>
</tr>
</tbody>
</table>

* p < .05, ** p < .01
Table III Spearman’s rank correlation coefficient $\rho$ between navigation accuracy, flight performance, gaze parameters, and subjective measures, lower half and upper half of the table corresponds to easy route and difficult route respectively.

<table>
<thead>
<tr>
<th>Easy Route</th>
<th>Navigation Accuracy</th>
<th>RMS error</th>
<th>Median dwell</th>
<th>Median OTW dwell</th>
<th>Median MAP dwell</th>
<th>Num. of fixations per OTW</th>
<th>Num. of fixations per MAP</th>
<th>Scan Duration OTW</th>
<th>Scan Duration MAP</th>
<th>OTW-MAP view changes</th>
<th>OTW-MAP</th>
<th>Route Diff Estimation</th>
<th>Route Diff Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigation accuracy</td>
<td>-</td>
<td>-.31</td>
<td>.03</td>
<td>.20</td>
<td>-.20</td>
<td>-.03</td>
<td>-.31</td>
<td>.08</td>
<td>.08</td>
<td>.22</td>
<td>-.53*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMS error</td>
<td>-.52*</td>
<td>-</td>
<td>.30</td>
<td>.08</td>
<td>.22</td>
<td>-.24</td>
<td>-.15</td>
<td>-.01</td>
<td>-.04</td>
<td>-.04</td>
<td>-.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median dwell</td>
<td>-.45</td>
<td>.03</td>
<td>-</td>
<td>.82*</td>
<td>.61**</td>
<td>.31</td>
<td>.22</td>
<td>.62*</td>
<td>-.69**</td>
<td>.32</td>
<td>-.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median OTW dwell</td>
<td>-.52*</td>
<td>.13</td>
<td>.94**</td>
<td>-</td>
<td>.38</td>
<td>.34</td>
<td>.36</td>
<td>.54*</td>
<td>-.69**</td>
<td>.37</td>
<td>-.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median MAP dwell</td>
<td>-.39</td>
<td>.00</td>
<td>.87*</td>
<td>.76**</td>
<td>-</td>
<td>.33</td>
<td>-.01</td>
<td>.45</td>
<td>-.60*</td>
<td>.22</td>
<td>-.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Num. of fixations per OTW</td>
<td>.26</td>
<td>-.27</td>
<td>.58*</td>
<td>.45*</td>
<td>.55</td>
<td>-</td>
<td>.73**</td>
<td>.08</td>
<td>-.78**</td>
<td>.21</td>
<td>.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Num. of fixations per MAP</td>
<td>.32</td>
<td>-.15</td>
<td>.03</td>
<td>.01</td>
<td>-.07</td>
<td>.29</td>
<td>-</td>
<td>-.08</td>
<td>-.70**</td>
<td>-.10</td>
<td>.57*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scan Duration OTW</td>
<td>.13</td>
<td>-.20</td>
<td>.43*</td>
<td>.44</td>
<td>.44</td>
<td>.69**</td>
<td>-.24</td>
<td>-</td>
<td>-.40</td>
<td>.23</td>
<td>.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OTW-MAP view changes</td>
<td>.00</td>
<td>.29</td>
<td>-.78**</td>
<td>-.66*</td>
<td>-.73**</td>
<td>-.83**</td>
<td>-.29</td>
<td>-.49</td>
<td>-</td>
<td>-.07</td>
<td>-.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Route Diff Estimation</td>
<td>.00</td>
<td>.10</td>
<td>-.33</td>
<td>-.27</td>
<td>-.10</td>
<td>-.32</td>
<td>.25</td>
<td>-.61*</td>
<td>.11</td>
<td>-</td>
<td>-.69**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Route Diff Assessment</td>
<td>-.55*</td>
<td>.33</td>
<td>.20</td>
<td>.09</td>
<td>.38</td>
<td>-.13</td>
<td>-.39</td>
<td>-.08</td>
<td>-.12</td>
<td>.20</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* $p < .05$, ** $p < .01$
Table IV. Mean, median and standard deviation (std) of DV (dependent variables) for subjects who were on-track or off-track for both easy and difficult route sections

<table>
<thead>
<tr>
<th>DV</th>
<th>On-track subjects (three subjects)</th>
<th>Off-track subjects (two subjects)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>median</td>
</tr>
<tr>
<td>TFH</td>
<td>1780 hrs</td>
<td>1600 hrs</td>
</tr>
<tr>
<td>OFH</td>
<td>867 hrs</td>
<td>850 hrs</td>
</tr>
<tr>
<td>RMS error</td>
<td>16.4 ft</td>
<td>14.2 ft</td>
</tr>
<tr>
<td>Median dwell duration</td>
<td>215.9 msec</td>
<td>196.7 msec</td>
</tr>
<tr>
<td>Median OTW dwell duration</td>
<td>228.6 msec</td>
<td>214.6 msec</td>
</tr>
<tr>
<td>Median MAP dwell duration</td>
<td>220.3 msec</td>
<td>213.3 msec</td>
</tr>
<tr>
<td>Num. of OTW fixations per view</td>
<td>3.7</td>
<td>4.3</td>
</tr>
<tr>
<td>Num. of MAP fixations per view</td>
<td>1.5</td>
<td>1.6</td>
</tr>
<tr>
<td>OTW scanning time</td>
<td>63%</td>
<td>62%</td>
</tr>
<tr>
<td>Num. of OTW-MAP view changes per second</td>
<td>1.3</td>
<td>1.4</td>
</tr>
<tr>
<td>Route difficulty Estimation</td>
<td>28.2</td>
<td>29.8</td>
</tr>
<tr>
<td>Route difficulty Assessment</td>
<td>27.8</td>
<td>27.3</td>
</tr>
</tbody>
</table>
List of Figure Captions

Figure 1. Flight route showing 2\textsuperscript{nd} to 12\textsuperscript{th} waypoints with corresponding dog houses: wp2-4 and wp5-7 are shown in thick lines (15).

Figure 2. Subject 5’s actual flight trajectory (blue line) and planned route (circles with connecting lines) (18)

Figure 3. Visual scan difference between expert and novice
Visual scan differences between experts and novices