Expertise Differences in Attentional Strategies Related to Pilot Decision Making

Angela T. Schriver[1], Daniel G. Morrow[1], Christopher D. Wickens[2], and Donald A. Talleur[1]
[1] Human Factors Division, University of Illinois at Urbana-Champaign
[2] MAAD Division of Alion Science & Technology Corporation, Boulder, CO

While much is known about differences in decision making outcomes related to pilot expertise, less is known about the processes that underlie these differences. We explored expertise differences in decision making processes by simultaneously measuring expert and novice pilots' attention, using eye-tracking, and their decision outcomes in a realistic context. We also investigated how expertise differences in pilots' attentional strategies were influenced by cue properties of diagnosticity and correlation. Fourteen expert and 14 novice pilots flew brief simulated flights. Half the flights contained failures that required diagnosis and an action (i.e., a decision). The environmental cues that signaled these failures varied in diagnosticity and/or correlation. We found that experts made better decisions than novices in terms of speed and accuracy. Both groups made faster correct decisions when cues were higher in diagnosticity. Only experts made faster correct decisions when cues were correlated. Experts attended more to cues relevant to the failure when a failure was present. Findings suggest that expertise differences in decision outcomes partly reflect attentional strategies relevant to problem diagnosis.

INTRODUCTION

The benefits of experience for pilot decision making have been repeatedly demonstrated, but a firm understanding of the psychological processes underlying these benefits is still unavailable. The current study investigates expertise differences in attentional patterns (using eye-tracking) while also measuring decision outcomes of expert and novice pilots flying simulated flight scenarios requiring fault diagnosis and corresponding action (i.e., decision making).

Expertise-related differences may occur at "cue seeking", "diagnosis" or "action choice" stages of decision making (Wickens & Hollands, 2000). For example, at the cue seeking stage, experts may be better able to direct attention to appropriate cues for decision making. Bellenkes, Wickens and Kramer (1997) found that expert pilots are able to direct their attention in a manner conducive to selecting flight-relevant information. At the diagnostic phase, Stokes, Kemper and Marsh (1992) found that experts made more accurate diagnoses of flight-related problems than novices did because they relied on more elaborate aviation-related knowledge stored in long-term memory (LTM). These studies also suggest that expertise advantages exhibited in attention and knowledge may interact in that experts' attention is directed by mental models based on domain knowledge.

Given that expert-related differences on pilot decision making tasks partly reflect differences in attention and domain knowledge, predictions about how the cue properties of diagnosticity and correlation impact expertise differences in decision making follow. Before presenting these predictions, we briefly review the supporting literature.

Diagnosticity of Cues

Cue diagnosticity is defined as the degree to which a piece of information specifies a particular state. For example, a reading near empty on a gas gauge in an automobile can be said to be high in diagnosticity because it indicates the state of the automobile being low on gas and almost never indicates any other state (with the rare exception of a malfunctioning gauge). In this sense, information that is high in diagnosticity can also be thought of as relevant to correctly diagnosing the state of the situation. Diagnostic information should be selected in the cue seeking phase and integrated in the diagnosis phase of decision making (Wickens & Hollands, 2000).

More experienced pilots may be better able to select and attend to diagnostic information than less experienced pilots. Stokes, et al. (1992) found that, compared to novice pilots, experts identified a higher number of diagnostic cues, fewer irrelevant cues, and had a higher ratio of diagnostic to total cues identified. The percentage of diagnostic cues identified was the best predictor of decision optimality, once the certificate held by pilot was removed from the regression analysis. This result supports a hypothesized causal link that paying attention to diagnostic cues leads to more optimal decision making performance. Wiggins and O'Hare (2003) also found that expert pilots, defined as those pilots with more hours of cross-country flight experience, used more diagnostic cues when making weather-related decisions. Morrow et al. (2008) found that expert pilots spent more time than novices reading cues that were independently determined to be diagnostic for solving problems in unfolding descriptions of those problems. Such findings are consistent with Shanteau's (1992) proposal that experts are more discriminating than novices when it comes to diagnostic information because learning cue diagnosticity is context-specific and can only develop with experience.

Correlated Cues

Correlated cues are "the co-occurrence of features in perception" (Wickens & Hollands, 2000). Correlated cues form a recognizable pattern such that the decision maker need not consider each piece of information separately and use working memory resources to integrate that information. For example, if a physician knows that two symptoms ALWAYS co-occur, then upon finding one symptom, she need not look for the other. In some cases, all of the cues may not signal the state independently, but their co-occurrence signals the state. Experts should better recognize these patterns because, based on experience, they have more stored information about co-occurrence in LTM against which to compare current cue patterns (Wickens & Hollands, 2000). These ideas form the basis of Klein's (1993) recognition-primed decision making, which accounts for experts outperforming novices in dynamic situation assessment because experts rapidly compare the characteristics of the present situation to schemas stored in LTM.

Some evidence suggests that experienced pilots take advantage of cue correlation to diagnose problems during flight. Stokes, et al. (1992) found that experts outperformed novices on a memory-intensive flight decision making task, but not on domain-general measures of short term working memory or on a test of declarative aviation knowledge, suggesting that experts could quickly match current information with stored patterns (e.g., a schema of correlated cues), which in turn allowed them to choose an appropriate action. Novices had to consider each cue individually, using the capacity-limited resources of working memory. As such, they were also more prone to error. However, few studies have investigated whether experts benefit more than novice pilots do from cue diagnosticity and correlation when detecting and responding to problems during flight, or the attentional processes underlying these differences.

Predictions

Expert and novice pilots flew simulated flights, half of which contained a failure that required diagnosis and a decision about how to respond. The diagnosticity and correlation of the cues to these problems were manipulated. Pilots' attention allocation was measured (by eye-tracking) as well as actions taken to diagnose and respond to the failures. Our predictions concern overall expertise-related differences in decision appropriateness and latency and associated attentional strategies, as well as how these differences are moderated by cue properties. We investigated the following cue conditions: 1) A **single,** highly diagnostic cue (**S**), 2) Multiple cues that are each **high** in diagnosticity and **correlated** to form a diagnostic pattern (**HC**), 3) Multiple cues that are each **low** in diagnosticity and **correlated** to form a diagnostic pattern (**LC**), and 4) Multiple cues with **mixed** diagnosticity (e.g., some high and some low in diagnosticity), that are **uncorrelated** with one another (**MU**). We made the following predictions:

* General expertise performance differences (across all cue conditions) should occur.

(1) Experts will make more appropriate decisions more quickly than novices, although (2) this difference will be smallest in the S condition because both novices and experts should be sensitive to one highly diagnostic cue.

* Expertise and cue diagnosticity.

(3) Both experts and novices will perform better when cues are high vs. low in diagnosticity.

(4) Expertise benefits will be greater in the low diagnosticity condition because the cues are correlated and experts should be more sensitive to this information in the absence of diagnosticity.

* Expertise and cue correlation.

(5) Both experts and novices will perform better when cues are correlated than when they are uncorrelated.

(6) Expertise benefits will be greater when cues are correlated than when they are uncorrelated.

The final predictions relate to expertise differences in attentional strategies related to diagnosing the problems. Attention allocation was measured by mean percent dwell time (MPDT) on areas of interest that contain the cues to the failure.

* Both experts and novices will (7) spend more time attending to problem-relevant cues post-failure onset in trials when a failure occurs than in trials when no failure occurs. (8) This effect will be amplified for experts.

* In the **MU** condition, (9) experts are more likely than novices to attend to highly diagnostic cues and ignore low diagnostic cues.

METHOD

Fourteen expert and 14 novice pilots participated (ages 19-44). All novices had a private pilot's license but were not instrument-rated. All experts had a commercial pilot's license with instrument ratings and had received flight instructor training. Expert pilots had more total hours (481.9 vs. 110.5) and instrument hours (80.5 vs. 10.8) than novice pilots, and outperformed them on the aviation knowledge test (adapted from an FAA written exam for instrument rating); experts and novices did not differ in mean performance on any of a set of cognitive ability measures administered prior to the experiment. Thus, experts differed from novice pilots in terms of domain knowledge and experience, but not domain-general cognitive abilities.

Participants each flew 16 brief simulated flights in a Frasca 142 flight simulator. An out-the-window display was projected on three 7' x 10' connected screens with a 135° field of view and Air Traffic Control (ATC) messages were simulated using audio recordings. Eye tracking data (frequency and duration of fixations) were recorded within 26 areas of interest (AOIs) comprising all of the gauges within the instrument panel, the radio rack, switch banks, an information display panel (presenting static weather and, when applicable, take-off conditions), and the three screens depicting the outside world. Depending on the scenario, different AOIs contained the various manipulated cues.

The eight failure scenarios experienced by a participant were drawn from a pool of 11 scenarios, created and validated by experienced flight instructors. Table 1 displays the scenarios in each condition experienced by each of three counterbalanced groups. Equal numbers of expert and novice pilots were assigned to each group. Performance during the failure scenarios was monitored by the experimenter who recorded actions taken by the pilot on a checklist. Eye data were time-stamped any time an action was recorded on the checklist. Non-failure scenarios were used as baselines for analysis of eye data.

Counterbalanced Groups

	Group A	Group B	Group C
S	Pitot Ice Failure, Avionics Failure	Pitot Ice Failure, Avionics Failure	Pitot Ice Failure, Avionics Failure
LC	Low Oil Pressure Failure, Carburetor Ice Failure	Low Oil Pressure Failure, Static System Failure	Carburetor Ice Failure, Static System Failure
HC	VOR Navigation Failure, Electrical Failure	VOR Navigation Failure, Significant Power Loss Failure	Electrical Failure, Significant Power Loss Failure
MU	Improper Loading Failure, Broken Altimeter Failure	Improper Loading Failure, High Elevation Performance Deficiency Failure	Broken Altimeter Failure, High Elevation Performance Deficiency Failure

Table 1. Scenarios experienced by each of three groups.

RESULTS

Performance measures

Diagnostic accuracy was defined as the percentage of failure scenarios that a pilot responded to correctly, determined by a subject matter expert (note: of the eight original scenarios, the pitot ice scenario within the S condition was dropped because no pilots responded correctly). *Notice latency* was defined as the time from the onset of a failure until the pilot first noticed the problem. The latter time was determined as follows: within the set of all fixations in a trial longer than 500 msec (the minimum amount of time required to extract new information, Mumaw et al., 2000), the subject matter expert coded 1) when the pilot was looking at an AOI that contained problem-relevant information and 2) whether the pilot subsequently modified his or her gaze pattern in a way that signaled he or she noticed a problem. If both these criteria were met, the fixation on the AOI containing the diagnostic information was coded as the notice point. *Action latency* is the time from the onset of the failure to the time the participant performed his or her first action in response to failure. Mean values for these three dependent variables, for the 4 cue conditions X expert and novice groups, are shown in Table 2.

		S	HC	LC	MU
Diagnostic Accuracy (percent)	**E**	78.6	75.0	75.0	89.3
	N	28.6	39.3	46.4	67.9
Notice Latency (sec)	**E**	20.4	6.8	36.2	28.1
	N	5.2	9.1	60.2	32.1
Action Latency (sec)	**E**	29.8	16.2	50.5	48.6
	N	25.9	35.5	79.6	45.4

Table 2. Performance measures.

General Expertise Differences in Decision-making Performance

To address predictions about overall expertise differences in performance, we conducted an expertise (novice, expert) x cue condition (S, HC, LC, MU) repeated measures ANOVAs on diagnostic accuracy (i.e., appropriate decision), notice latency, and action latency (for trials in which the participant responded appropriately to the problem). The main effect of expertise was most relevant to prediction 1, *experts make more appropriate decisions more quickly than novices*. Experts made appropriate decisions more often than novices (E = 79.5%, N = 45.5%, $F(1,26) = 27.7$, $p < .001$). Experts also responded more quickly to failures than novices (E = 36.3 seconds, N = 46.6 seconds, $F(1,22) = 8.9$, $p < .01$). The pattern of means suggested they also noticed the problem faster than novices, but this difference was not significant ($F(1,22) = 1.6$, $p > .10$).

The expertise x cue condition interaction was most relevant to prediction 2, *the expertise performance benefit will be smallest in the S condition*. The interaction was not significant for diagnostic accuracy ($F(3,78) = .9$, $p > .10$), indicating experts and novices showed similar patterns in accuracy among the four cue configurations. ANOVAs performed on notice latency and action latency yielded significant interactions ($F(3,66) = 10.3$, $p < .001$, $F(3,66) = 8.5$, $p < .001$ respectively), however the mean values shown in table 2 reveal that novices actually performed better than experts in the S condition, contrary to prediction 2.

Expertise and cue diagnosticity

To test predictions 3 and 4, expertise (novice, expert) x cue diagnosticity (LC, HC) repeated measures ANOVAs were performed on diagnostic accuracy, notice latency and action latency measures. Prediction 3, *both experts and novices will perform better when cues are high vs. low in diagnosticity*, was not supported by the analysis of accuracy, as the main effect of cue diagnosticity was not significant ($F(1,26) = .157$, $p > .10$). However, analysis of notice latency (LC = 48.2 seconds, HC = 8.0 seconds, $F(1,22) = 142.4$, $p < .001$) and action latency (LC = 65.1 seconds, HC = 25.8 seconds, $F(1,22) = 181.4$, $p < .001$) both supported prediction 3 by showing that participants noticed and acted in response to failures faster when cues were high in diagnosticity than when they were low in diagnosticity.

Prediction 4, *expertise benefits will be greater when diagnosticity is low,* was supported by an expertise x cue diagnosticity interaction for notice latency ($F(1,22) = 10.4$, $p<.005$) and a non-significant trend for the same interaction in the same direction for action latency ($F(1,22) = 2.8$, $p = .106$). Both showed larger expertise benefits when cues were low vs. high in diagnosticity. The ANOVA performed on accuracy showed no interaction ($F(1,26) =.157$, $p >.10$), thus eliminating the possibility that the above-described interactions were the result of a speed-accuracy tradeoff.

Expertise and cue correlation

In order to test predictions about the effects of correlated cues, the MU condition was compared to an average value from the two correlated conditions, HC and LC (i.e., HC_LC). Expertise (novice, expert) x cue correlation (HC_LC, MU) repeated measures ANOVAs on diagnostic accuracy, notice latency and action latency were used to test predictions 5 and 6. Prediction 5, *both experts and novices will perform better when cues are correlated*, was not supported by the results. The effect of correlation was significant, but in the opposite direction for accuracy ($F(1,26) = 10.8$, $p <.005$), and was not significant for notice latency ($F(1,22) = .6$, $p >.10$) or action latency ($F(1,22) = .2$, $p >.10$).

However, prediction 6, *expertise benefits will be greater when cues are correlated*, was supported. The critical expertise x condition interaction was marginally significant for notice latency ($F(1,22) = 3.0$, $p = .099$) and was significant for action latency ($F(1,22) = 18.0$, $p <.001$). For action latency, planned t-tests showed that experts did in fact perform better when cues were correlated (MU = 48.6, HC_LC = 33.3, $t(22) = 2.7$, $p <.05$), but novices performed worse in this condition (MU = 45.4, HC_LC =57.6, $t(22) = 2.8$, $p <.05$). No interaction was found for the accuracy ($F(1,26) = .807$, $p>.10$).

Expertise differences in attentional processes underlying decision making

To evaluate prediction 7, *both experts and novices spend more time attending to problem-relevant cues post-failure onset in failure vs. non-failure trials*, and prediction 8, *this attentional strategy is exaggerated for experts*, mean percent dwell time (MPDT) in problem-relevant AOIs for failure and non-failure trials was analyzed. MPDT allocated to AOIs that contained cues that indicated a failure was measured from the time of failure onset to the time the participant took his or her first action in response to the failure in both failure and non-failure trials. In non-failure trials, MPDT was measured within the same time interval as in their corresponding failure trials. An expertise (novice, expert) x trial type (non-failure, failure) repeated measures ANOVA was then performed on the MPDT for trials in which participants made appropriate decisions. A marginally significant effect of trial suggested more attention to failure trials (NF = 7.6%, F = 8.6%, $F(1,22) = 3.4$, $p = .08$). While experts and novices did not differ overall (N = 7.6%, E = 8.5%, $F(1,22) < 1.0$), a significant expertise x trial type interaction ($F(1,22) = 4.8$, $p <.05$) suggested experts but not novices allocated more attention to problem-relevant AOIs when a failure occurred ($t(22) = 2.5$, $p < .05$). The results of this analysis lend some support to both predictions 7 and 8.

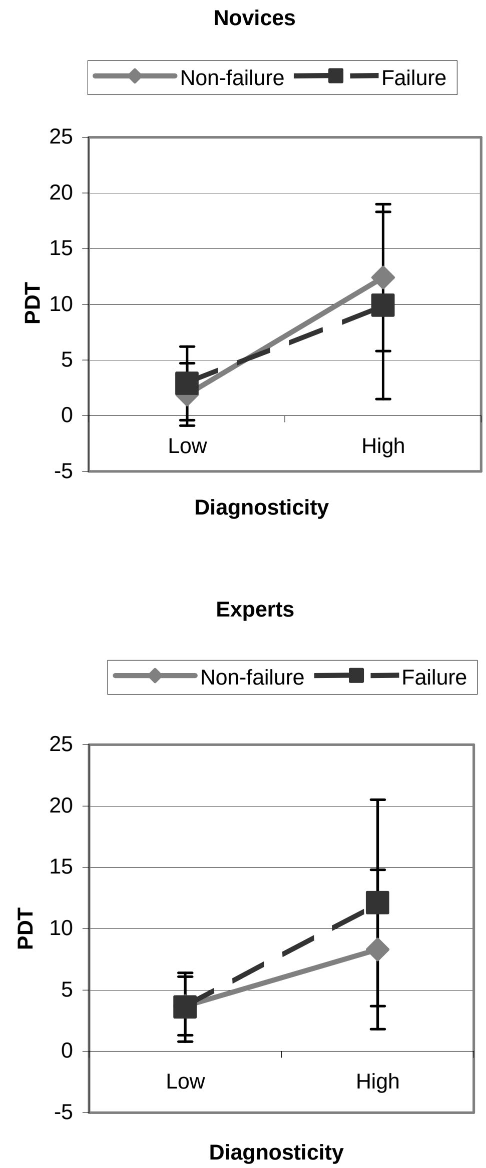

Figure 1. PDT in AOIs containing cues low or high in diagnosticity in non-failure and failure trials. Standard deviation is shown

To test prediction 9, *experts are more likely than novices to attend to highly diagnostic cues and ignore low diagnostic cues in the MU condition,* an expertise (novice, expert) x cue diagnosticity (low, high) x trial type (non-failure, failure) ANOVA was conducted on MPDT computed within an interval defined from the failure onset to the time when the participant took his or her first action. This analysis was restricted to the three scenarios in the MU condition, and

MPDT was computed only for those AOIs that contained low or high diagnostic cues for the failures. Consistent with the prediction, a significant expertise x cue diagnosticity x trial type interaction ($F(1,22) = 9.1$, $p < .01$) showed that experts allocated more attention to the more diagnostic cues during failure trials, and less to low diagnostic cues, while novices showed the reverse pattern (see Figure 1).

DISCUSSION

We explored how expertise differences in pilot decision making were reflected in both the outcome and process of the diagnosis stage of decision making. In doing so, we hoped to find evidence that links expertise differences in attention allocation - presumably directed by knowledge structures stored in long-term memory – to diagnosis and choice outcomes in a realistic context.

As a validity check, we note that experts, as operationally defined by hours of experience and rating type, were indeed better decision makers, both in terms of speed and accuracy. *But how were they better?* We found several differences consistent with expertise effects in other domains.

Importantly, expertise differences in pilot decision-making were sometimes moderated by the properties of problem-relevant cues, suggesting how experts outperformed novices. First, experts noticed and responded to problems more quickly when cues were correlated, whereas novices did not. This expertise-correlation effect was reinforced by the findings that experts, but not novices, appeared to better sustain rapid diagnosis when the diagnostic value of the set of cues was lower; but they were correlated nevertheless (P4). With correlated cues of higher diagnosticity, any cue could be selected to yield the correct diagnosis; it is only when diagnosticity is lowered that several cues need be considered. Experts could do this more rapidly than novices, presumably by having good mental models of the cue correlations.

Second, consistent with our predictions, experts exhibited greater attentional sensitivity to the more diagnostic cues when cues varied in diagnosticity, as revealed by the process measure of visual scanning. Importantly, as revealed by the three-way interaction, it was only when the information sources were necessary for problem solving that they received more attention. We may assume that this process measure was, at least in part, responsible for the overall benefit in accuracy of experts in the MU condition; a benefit which replaced that of correlation knowledge, manifest in the HC and particularly the LC conditions.

Some findings, however, were inconsistent with predictions. Most prominent was the finding that expertise benefits were smaller, not greater (as predicted), on the multi-cue vs. single cue condition. Other predictions, while not contra-indicated by the results, were not supported either, because of null statistical effects. For example, accuracy did not show an overall benefit for correlated vs. uncorrelated cues or for more vs. less diagnostic ones. In general, the fact that accuracy showed few effects probably reflects the lower statistical power of this measure, based as it was, upon a proportion score from 0 to 7.

CONCLUSION

Our study, like previous work, showed that expert pilots are better decision makers than novices. Further, we found that expert pilots moderate their attention by attending more to the relevant AOIs than novices do when failures requiring diagnosis are present. Together, these findings provide support for the hypothesized link between attention and decision making (see O'Hare, 2003).

Because we found superior decision making performance among experts and some evidence that this performance was related to their attention allocation strategies, our findings related to how cue properties of diagnosticity and correlation influence expertise differences in decision making take on particular importance. We did in fact find a performance (notice and action latency) benefit for cue diagnosticity, such that novices and experts benefited when all cues were high-diagnostic, and a specific expertise benefit for cue correlation, such that only experts were able to take advantage of correlation.

Overall, the results show expertise differences in sensitivity to the cue properties in terms of attention allocation and associated decision making quality. Achieved in a high-fidelity flight simulator, these results also address the issue of lack of realism present in previous work.

REFERNCES

Bellenkes, A. H., Wickens, C. D., & Kramer, A. F. (1997). Visual scanning and pilot expertise: The role of attentional flexibility and mental model development. *Aviation, Space, and Environmental Medicine, 68*(7), 569-579.

Klein, G. A. (1993). A recognition-primed decision (RPD) model of rapid decision making. In G. A. Klein, J. Orasanu, R. Calderwood & C. E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 138-147). Norwood, NJ: Albex.

Morrow, D. G., Miller, L. M. S., Ridolfo, H. E., Magnor, C., Fischer, U. M., Kokayeff, N. K. & Stine-Morrow, E. A. L. (2008). Expertise and age differences in pilot decision making. *Aging, Cognition, & Neuropsychology*. In press.

Mumwaw, R., Sarter, N., Wickens, C., Kimball, S., Nikolic, M., Marsh, R., et al. (2000). *Analysis of pilots' monitoring and performance on highly automated flight decks* (Final Project Report No. NAS2-99074). Moffett Field, CA: NASA Ames Research Center.

O'Hare, D. (2003). Aeronautical decision making: Metaphors, models, and methods. In P. S. Tsang, & M. A. Vidulich (Eds.), *Principles and practice of aviation psychology* (pp. 201-237). Mahwah, NJ: Lawrence Erlbaum Associates.

Shanteau, J. (1992; 1992). How much information does an expert use? Is it relevant? *Acta Psychologica, 81*(1), 75-86.

Stokes, A. F., Kemper, K. L., & Marsh, R. (1992). *Time-stressed flight decision making: A study of expert and novice aviators* (Tech. Rep. No. ARL-93-1/INEL-93-1). Savoy, IL: University of Illinois, Aviation Research Laboratory.

Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Wiggins, M. W., & O'Hare, D. (2003). Expert and novice pilot perceptions of static in-flight images of weather. *The International Journal of Aviation Psychology, 13*(2), 173-187.