Federal Aviation Administration

DOT/FAA/AM-10/6
Office of Aerospace Medicine
Washington, DC 20591

Effects of Video Weather Training Products, Web-Based Preflight Weather Briefing, and Local Versus Non-Local Pilots on General Aviation Pilot Weather Knowledge and Flight Behavior, Phase 2

William Knecht[1]
Jerry Ball[1]
Michael Lenz[2]

[1]Civil Aerospace Medical Institute
Federal Aviation Administration
Oklahoma City, OK 73125

[2]FAA Headquarters
Washington, DC 20591

March 2010

Final Report

OK-10-0077-JAH

NOTICE

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-10/6		

4. Title and Subtitle	5. Report Date
Effects of Video Weather Training Products, Web-Based Preflight Weather Briefing and Local Vs. Non-Local Pilots on General Aviation Pilot Weather Knowledge and Flight Behavior, Phase 2	March 2010
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Knecht WR, Ball J, Lenz M	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
FAA Civil Aerospace Medical Institute P.O. Box 25082 Oklahoma City, OK 73125	11. Contract or Grant No.

12. Sponsoring Agency name and Address	13. Type of Report and Period Covered
Office of Aerospace Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, DC 20591	14. Sponsoring Agency Code

15. Supplemental Notes

Work was accomplished under approved task AM-A-07-HRR-521

16. Abstract

This research had two main phases. Phase 1 investigated three major questions, one of which was whether or not video weather training products could significantly affect general aviation (GA) pilot weather knowledge and flight behavior in marginal meteorological conditions. Fifty GA pilots took a general weather knowledge pre-test, followed by exposure to either one of two weather training videos (the Experimental groups) or to a video having nothing to do with weather (the Control group). They next took a post-test to measure knowledge gain induced by the training product. Finally, they planned for and flew a simulated flight mission through marginal weather from Amarillo, TX, to Albuquerque, NM. Multivariate modeling implied that a *combination* of higher pilot age, receiving either weather training product, and takeoff hesitancy could significantly, correctly predict 86.7% of diversions from deteriorating weather and 77.8% of full flight completions.

The question then became whether or not this model would be robust over time. In the present study (Phase 2), after a time lapse of 3-4 months, 44 of the 50 original Phase 1 pilots returned for further testing. Again, they were tested for weather knowledge and flew a simulated flight mission similar to Phase 1's.

No significant change in weather knowledge was evident from Phase 1 to 2, nor were any significant differences seen between the three treatment groups. Additionally, the 3-factor model of Phase 1 failed to significantly predict flight diversions or flight completions in Phase 2.

The combined results of Phases 1 and 2 imply that the effects on weather knowledge and flight behavior of a single 90-minute training video seem minimal in comparison to the complexities of weather itself and flight into weather. This is consistent with intuition. Moreover, what small effects are produced seem to decay with time.

None of this is unexpected. It merely means that weather is complex, and effective weather training must be intensive to begin with and ongoing to remain effective.

17. Key Words	18. Distribution Statement
Weather, Training, Pre-Flight Briefing, Weather Knowledge, Flight Behavior	Document is available to the public through the Defense Technical Information Center, Ft. Belvoir, VA 22060; and the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	24	

ACKNOWLEDGMENTS

Stanley Roberts, Manager, AFS-630 was invaluable in providing item difficulties to FAA test questions. Tammy Harris, FAA, and Janine King, Sally Glasgow, and Suzanne Thomas of Xyant Technology, Inc. provided key scheduling and subject payment support. Again, we extend our heartfelt gratitude to the pilots who made this effort possible.

CONTENTS

INTRODUCTION 1
Purpose of This Research 1
METHOD 1
Background 1
Research Design—Original Assignment to Group and Order of Treatments 1
Participants and Attrition 2
Apparatus 2
Procedure 2
RESULTS 4
Data Normality, Phase 2 4
Specific Effects 5
Modeling Flight Behavior 13
DISCUSSION 14
Phase 1, Brief Summary 14
Phase 2, Brief Summary 16
Findings Common to Both Studies 16
REFERENCES 17
APPENDIX A: Frequency Histograms A-1
APPENDIX B: Phases 1 & 2 Correlational Structures B-1

Effects of Video Weather Training Products, Web-Based Preflight Weather Briefing, and Local Versus Non-Local Pilots on General Aviation Pilot Weather Knowledge and Flight Behavior, Phase 2

INTRODUCTION

Purpose of This Research

FAA Civil Aerospace Medical Institute (CAMI) researchers were tasked by the FAA Flight Standards division (AFS-810) to explore a number of issues in general aviation, including:

1. Do video weather training products significantly affect pilot weather knowledge and flight behavior in the face of potential instrument meteorological conditions (IMC)?
   a. If so, what are the immediate effects?
   b. Do these effects persist over time?
2. How are modern Web-based weather products used during preflight briefing?
3. Do local[1] Oklahoma pilots differ appreciably from non-local pilots in either weather knowledge or weather-related flight behavior?

Issue 1 is consequential because bad weather is a perennial concern to general aviation, and therefore remains a continual focus of FAA investigation.

Issue 2 begins human factors study of what promises to be the future of preflight weather briefing—self-briefing by pilots using World-Wide Web-based tools.

Issue 3 addresses the question of whether or not Oklahoma pilots are representative of U.S. pilots in general. Presumably, they are similar but, so far, this has not been directly investigated. Since many of the FAA's general aviation studies are conducted by the Oklahoma-based Civil Aerospace Medical Institute, this is a statistical validity issue calling for study.

METHOD

Background

This research was conducted in two phases. Phase 1 examined data collected from January to July, 2008, and was the subject of a prior report by similar name. Phase 2 data were collected from July to September, 2008, and are the subject of the current report. Most of the current methodology is detailed in the Phase 1 report. Therefore, only key findings from Phase 1 will be reiterated here.

The reason for conducting a longitudinal study was to investigate the effect of time on any learning taking place due to the video training product during the Phase 1 study. Phase 1 found no statistically significant single, direct effect of the training product on general weather knowledge or hazardous flight behavior (in simulo). Instead, what was found through modeling was a significant *multiple effect* of a *combination* of

- Training product(the two training products differed from the Control, but not from each other)
- Pilot age
- Takeoff hesitancy(pilots' initial yes/no decision whether they would even attempt the flight)

on subsequent completion of the entire flight (although not directly on objectively observable flight hazard variables).[2] Since risk was arguably cumulative, an argument could be made that completion of the entire flight was an indicator of increased flight hazard.[3]

Research Design—Original Assignment to Group and Order of Treatments

Because Phase 2 was a follow-up study, no further experimental treatments were introduced. Table 1 shows the original experimental structure, which was maintained in Phase 2. Pilots had been assigned to one of 3 primary *Training product* groups, each of which was sub-divided into 2 secondary *Instrument rating* groups, each of which was sub-divided into 2 *Pilot residence* groups containing "local" (Oklahoma residents) versus non-local residents. This produced a 3x2x2 mixed design with 12 treatment cells, 4 Ss per cell. Pilots had been assigned to cells in order to equilibrate cell means for age and flight hours.

[1]A "local" pilot was defined as one living in Oklahoma at the time of the study. For the most part, this meant long-term Oklahoma residents. However, there were a few instances of pilots living in Oklahoma whose state-of-legal residence was not Oklahoma because they were attending local flight schools.

[2]Severe non-normality of flight hours and the hazard variables prevented doing standard continuous-data linear regression modeling. Instead, a more-tolerant stepwise binary logistic regression was done on the yes/no outcome of whether or not pilots completed the entire flight.

[3]The way the Phase 1 flight scenarios were physically constructed, there was a -.384 ($p = .006$) correlation between how close pilots finally got to the destination (Albuquerque, NM) and the number of minutes spent < 500' AGL (above ground level).

Table 1. Experimental structure.

Independent variables (IV)					Dependent variables (DV)					
Training product	Instrument-rated pilot?	Pilot's state of residence	Age	Flight hr	Flight duration	Minimum distance to ABQ	Minutes spent in IMC	Minutes scud running	Minutes spent < 500' AGL	Takeoff decision (Y/N)
1	No	OK	12 treatment cells were equilibrated for age and flight hr.							
1	No	Non-OK								
1	Yes	OK								
1	Yes	Non-OK								
2	No	OK								
2	No	Non-OK								
2	Yes	OK								
2	Yes	Non-OK								
Control	No	OK								
Control	No	Non-OK								
Control	Yes	OK								
Control	Yes	Non-OK								

Table 2. Phases 1-2 attrition characteristics for cell ns, median age, median flight hours, and % instrument-rated.

	Training Product 1				Training Product 2				Control				Combined group			
	n	Age	FH	%IR	n	Age	FH	%IR	n	Age	FH	%IR	N	Age	FH	%IR
Phase 1	16	39	280	53	16	38	235.5	47	18	42	262.5	50	50	39	268	50
Phase 2	15	39	300	60	15	34	227	53	14	42	225	57	44	39	280	57

Participants and Attrition

Fifty GA pilot volunteers had originally been selected to participate with informed consent, but six were unable to participate in Phase 2, dropping the final N to 44. Table 2 describes key demographic factors and changes. None of these changes was statistically significant.

Apparatus

Weather training products/control materials. Two well-known video weather training products had been selected from a list of candidate products. The Phase 1 report details the selection method. The authors of these products graciously provided them on condition of confidentiality; therefore their wishes for confidentiality shall be maintained in this report as well.

Weather Knowledge Tests. In Phase 1, three parallel forms of a 30-question general weather knowledge test were constructed and matched on item difficulty. Administration order of the parallel forms was counterbalanced across pilots, controlling for the event that the three forms might not be exactly equivalent in difficulty. Each test was administered on a laptop computer using software written by the experimenters in Microsoft *Visual Studio 2005*™.

In Phase 1, one of the 3 forms had been given as a pre-test, followed by the video training product, followed by a second form as the post-test. Now, in Phase 2, the remaining form was given to assess any groupwise change in weather knowledge scores.

Advanced General Aviation Research Simulator (AGARS). Again, the CAMI AGARS constituted the flight simulator platform for assessing flight behavior. Details of AGARS can be found in the Phase 1 report.

Procedure

The Phase 2 simulator mission was engineered to be highly similar to Phase 1. To recap briefly, pilots were asked

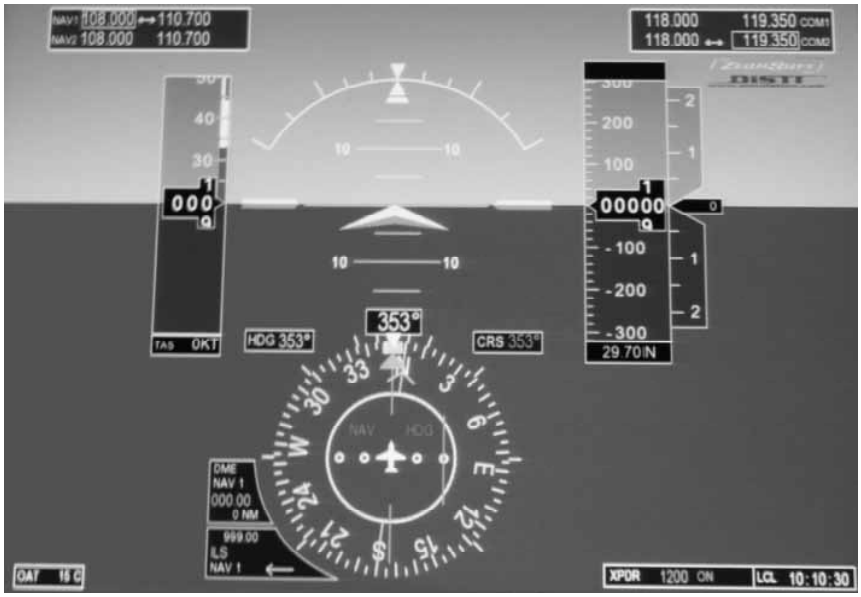

Figure 1. AGARS primary flight display.

to plan an east-to-west, visual-flight-rules (VFR) flight from Amarillo, TX (AMA) to Albuquerque, NM (ABQ). This route takes approximately 90 minutes to fly in the Piper Malibu configuration at high-speed cruise. Pilots had 2 cockpit VORs (VHF OmniRange Navigation System) and an ADF (Automatic Direction Finder). Again, they had the ability to access a Web-based weather emulation on a stand-alone PC during preflight. Upon finishing their flight planning, they took the final version of the weather knowledge test as dictated by the counterbalance order. Next, a 15-minute convenience break was given to each pilot. Following the break, an abbreviated refamiliarization session with AGARS was given. Again, pilots were allowed to ask for assistance with flight settings at any time during the course of the flight scenario.

As before, the route consisted of gradually rising terrain during the first two-thirds of the flight, followed by a dramatic elevation change during the last third. During the course of the flight, pilots were exposed to deteriorating VFR weather conditions. Again, visibility was initially set at 8 nautical miles and gradually decreased to 5 miles by the time the pilots had flown approximately two-thirds of the route. Cloud ceilings were lowered from 4500 feet AGL to 3500 AGL across the same terrain. As a result, the ceilings again gradually squeezed the pilots closer to the ground.

Shortly into the flight, the barometric pressure dropped from the preflight planning level of 30.10 to 29.98. This afforded a potential error between actual and intended altitude for pilots not receiving a barometric update (either from AFSS or AWOS) after departure. Specifically, pilots failing to update their Kollsman setting would fly an actual altitude approximately 120' *below* their intended altitude. The authors acknowledge that, given the aircraft's blind transponder encoder plus the departure airspace, in real life an air traffic controller would have normally detected the altitude discrepancy and issued a correction to the pilot. However, since the study's purpose was specifically to study both errors of commission and omission, this correction was purposely skipped to study the consequences.

Because the flight situation essentially did not change from Phase 1 to 2, the issue of learning effect had to be addressed.[4] Pilots might fly better in Phase 2 simply because they would now know better the airframe and the physical terrain.

Two methods were therefore used to distract pilots from the similarities between Phases 1 and 2. First, the direction of approach for Phase 2 weather was made symmetrical to, and counterbalanced with, whatever each pilot had experienced during Phase 1. For instance, if a pilot had experienced Phase 1 weather approaching from 45° (their 4 o'clock) on this east-to-west flight, the Phase 2 weather approached symmetrically from 135° (their 8 o'clock). This symmetry was purely a distractor, not a variable of interest.

The second distractor was the introduction of a primary flight display (PFD). The PFD incorporated the basic flight instruments that provided aircraft attitude, airspeed, altitude, vertical speed, heading, rate-of-turn, slip-and-skid, navigation, transponder, and systems annunciation data (See Figure 1). The inset map view (containing map, traffic, and terrain information) was removed

[4]Two reasons dictated not using 2 different, counterbalanced routes. First, the AGARS scenery database did not encompass the entire continental U.S. Second, even if it had, pilots would have easily spotted the essential flight features (i.e., slowly rising terrain terminated by hills, with deteriorating visibility and ceiling along the way).

for consistency with the Phase 1 display information. Navigational information was coupled to the autopilot and the NAV/COM radio heads. Temperature and time were also presented on the display in the lower left and right corners respectively. Pilots had the ability to adjust the barometric pressure, course direction indicator, and heading bug. Each pilot received a 20-30 minute training session using the display. During this training, pilots were asked to fly a short flight from an airport out to a VOR at a requested altitude and then return back to the departure airport to land. They were asked to use the autopilot to fly a direct course to the station and then reverse their course and fly an outbound leg back to the airport. This gave the pilots time to utilize and familiarize themselves with all of the display's functionality.

RESULTS

Data Normality, Phase 2

Normality of data frequency distributions is a prerequisite for the use of parametric statistics. Therefore, a preliminary check of all data was made.

Normality of weather knowledge test scores. Acceptable normality of weather knowledge test scores was supported by 2-tailed Shapiro-Wilk tests ($p_{pre\text{-}test}$=.297, $p_{post\text{-}test}$=.786, p_{final}= .653, all non-significant (NS)).

Normality of Web-based weather information data. In Phase 1, the Web-emulation page view duration data presented considerable departure from normality. Many pages received very little viewing time while some received a great deal. Moreover, some of the longest viewing times reflected mistakes (pilots forgetting to close out a particular page). Therefore, medians and quartiles were used in Phase 1 to more accurately reflect group usage patterns. That approach was continued in Phase 2.

Normality of flight simulator data. Appendix A shows Phase 1 AGARS flight data displayed in the left column, with Phase 2 data in the right column. Phase 1 flight duration was the only distribution to pass the Wilk-Shapiro test of normality.

Most of these data were simply too extreme to be corrected by any standard mathematical transform such as a log or square root. Therefore, most analysis was subsequently done using distribution-free (non-parametric) statistics.

Equivalency of weather knowledge test forms A, B, and C. With Phase 2 complete, we can now analyze the 3 test forms for equivalency of difficulty. As usual, we first visually inspected the data frequency distributions. Figure 2 shows frequencies collapsed over all 3 forms for percent correct and elapsed time (time spent taking the test). Elapsed time appeared unexpectedly non-normal. This will be elaborated upon shortly.

Table 3 shows the knowledge test means and standard errors of the mean for the N=42 surviving score triplets (2 of the 44 returning pilots had originally failed to take the Phase 1 knowledge pre-test).

A Shapiro-Wilk test supported normality across forms A,B,C for percent correct (N=42, p_A=.313, p_B =.715, p_C=.359). However, the distributions failed Mauchy's Test of Sphericity (p=.043), so a corrected repeated-measures ANOVA (Greenhouse-Geisser) was used to examine mean differences.

The results implied that Forms A, B, and C were not equivalent in difficulty (p =.001). Examination of the .95 confidence intervals revealed that Form A scores were significantly higher than both B and C, with B and C otherwise being about equal.

We therefore concluded that Form A was a somewhat easier test. Fortunately, the treatment order counterbalance statistically controlled for this, so it presented no fatal flaw to the overall analysis.

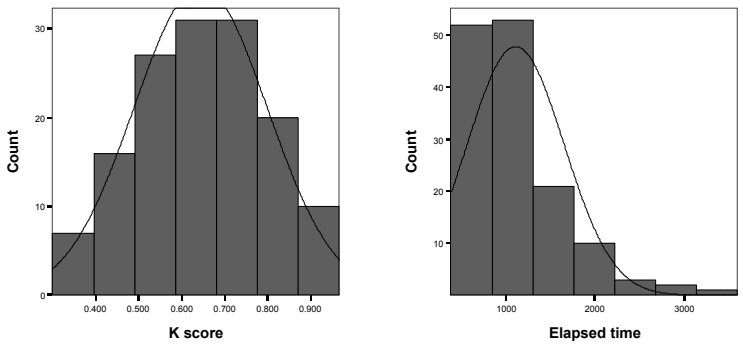

Figure 2. Score frequency distributions for a) the knowledge test, and b) time spent taking the test. Histograms represent combined data from all three administrations (Forms A, B, and C).

Table 3. Knowledge-score means, collapsed across all 3 test forms.

	Mean Correct	Standard error of the mean
Form A	70.7%	2.0
Form B	62.9%	2.0
Form C	64.2%	2.7

Table 4. Elapsed time means, collapsed across all 3 test forms.

	Mean ET (seconds)	Standard error of the mean
Form A	1077	73.8
Form B	1193	85.1
Form C	958	82.6

Table 5. Weather knowledge means--% correct, N=42—and (.95 CI).

	Pre-treatment	Post-treatment	Final
Trg Prod 1	66.7 (± 10.2)	68.1 (± 9.2)	63.3 (± 10.1)
Trg Prod 2	65.0 (± 7.8)	67.4 (± 6.3)	66.4 (± 6.3)
Control	66.2 (± 7.9)	66.9 (± 6.0)	63.6 (± 5.6)

Table 6. Elapsed time (ET) means (seconds, N=42).

	Pre-treatment	Post-treatment	Final	Row Ave.
Trg Prod 1	1486	1034	982	1167
Trg Prod 2	1414	1021	924	1120
Control	1117	831	873	940
Column Ave.	1339	962	927	1076

Similar to Table 3, Table 4 shows means for elapsed time (ET), collapsed across Phases 1+2. This is the time each pilot spent taking each of their 3 tests, with results grouped by test form A, B, or C.

Here, Shapiro-Wilk rejected normality on all 3 test forms for elapsed time (N=42, p_A=.015, p_B <.001, p_C <.001). Therefore, a nonparametric Friedman test was used to test differences between group medians.

Results implied that pilots spent more time taking some forms of the test than others (p =.017). Followup Wilcoxon pairwise difference tests pinpointed that pilots spent significantly more time taking form B than C ([$p_{A\text{-}B}$=.130, $p_{B\text{-}C}$=.006, $p_{A\text{-}C}$=.103). Fortunately, the treatment order counterbalance again protected the analysis from fatal harm.

At this time, other than sampling error, we have no explanation for why form A should be easiest, or for why form C should be fastest.

Specific Effects

Effect of the weather training products on GA pilot weather knowledge. Stepping back to view the big picture, one of the most interesting questions is whether the weather training product had any effect on pilots' weather knowledge. After Phase 1, we saw no significant change. Now, what about over the entire time course of the experiment?

Reanalysis across all 3 groups (pre-treatment, post-treatment, and final) found acceptable Shapiro-Wilk normality for weather knowledge scores (all p >.297), and no trouble with sphericity (p_{Mauchy} = .079). But, repeated measures ANOVA showed neither significant weather knowledge effects across time (p_F = .396) nor effect of training product (p = .908). Adding instrument rating and pilot's locality of residence to the analysis failed to uncover any interaction effects (smallest p = .389). Table 5 shows the means (with 95% confidence interval in parentheses).

There might be more than meets the eye here, however. When the data were collapsed across training products, it was clear that pilots spent, on average, significantly *less time* taking their knowledge tests *after* the pre-test than before it. These elapsed time distributions failed Shapiro-Wilk (all ps < .005), so the Friedman test was used, which revealed a significant overall decrease in ET over the course of the experiment (refer to column averages, $p_{Friedman}$ = .000005). Table 6 (bottom row, light gray) shows this trend.

Could the fact that pilots were simply spending less time on the later tests be the reason we failed to see any significant knowledge boost for the weather training groups? In Phase 1, we discovered that the elapsed time spent taking the test (ET) did not predict weather knowledge score. On average, pilots who spent *more* time taking the test actually got *lower* average scores ($r_{Spearman}$ = -.187), although the effect was not statistically significant. Therefore, ET was useless as a covariate in ANOVA and could not be used to enhance the power of our analysis.

Table 6 shows that the largest average time drop occurred between the first and second administrations. This "pre-post drop" in ET can probably be explained simply by 1) on first testing, all pilots were unfamiliar with the software and 2) over half the pilots took their first test in the undisturbed privacy of their hotel rooms or homes, with little time pressure, whereas all took the post-tests and finals in the laboratory, where they were under supervision, often with a plane to catch that afternoon, *and* none could fly the simulator until after their knowledge test was finished. This does not necessarily mean they were less attentive or serious in the later sessions. It may merely mean they were highly motivated to get the test over with and get into the simulator, and likely concentrated more intently on the last 2 tests.

Conclusion. In Phase 1, we interpreted the overall results as implying that weather is simply too complex a subject for a single, brief weather training product to have much effect on general weather knowledge. Now, we can restate that finding to say that individual test score variability—"noise," if you will—seems just too large to detect whatever "signal" may have been generated by the training products. Methodologically, the way to get around that problem would be to develop a product-specific knowledge test designed to better assess the specific details of each training product. Of course, that would require establishing internal and external validity and reliability of the individual test items, which is an entire project in its own right (and the very reason we used pre-validated FAA items to begin with). In other words, the test would have to be more specific and sensitive to be able to detect any effect of a 90-minute training video on weather knowledge.

Comparison of Web preflight briefings. Figure 3 compares Phase 1 page view durations (top) with Phase 2 (bottom). Figure 3's logarithmic y-axis makes it easier to see the various small values, yet still represent large values. By comparing medians (rather than means), we expect more stable viewing estimates for each individual page since the effect of extreme values will be reduced. However, Figure 3 is sufficiently complex to make meaningful visual inspection difficult. Some statistical analysis and visual reinterpretation of the data may help make more sense of what they imply.

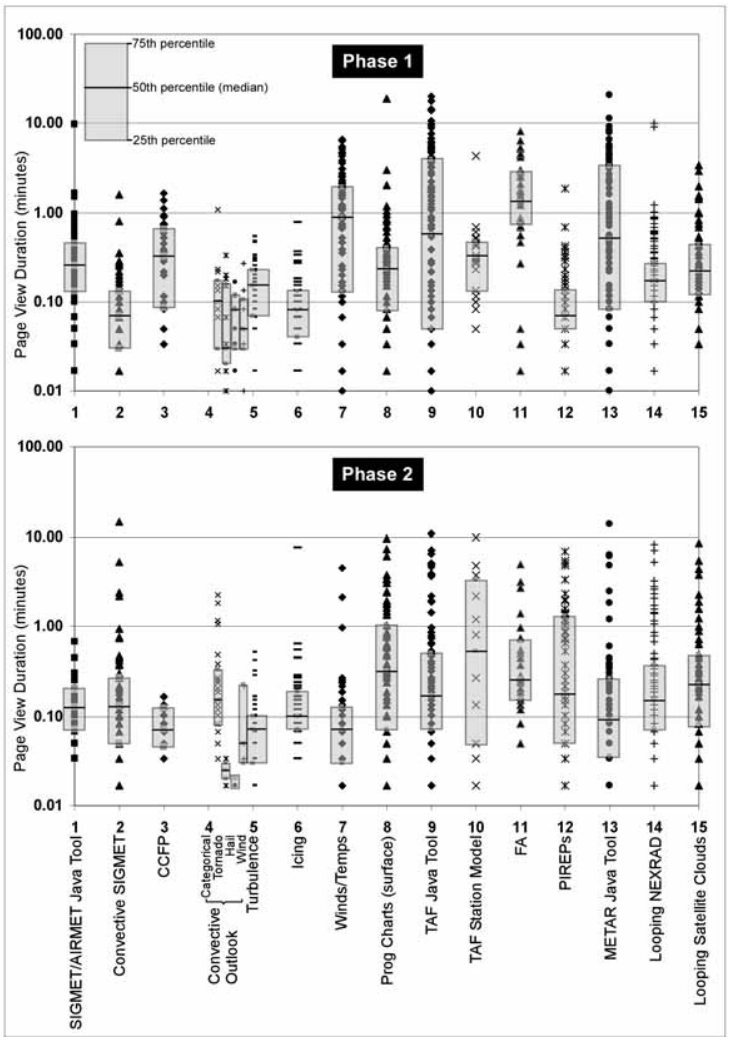

Figure 3. Whole-group page view durations for the part-task emulation of www.aviationweather.gov. Top: Phase 1 (N=50). Bottom: Phase 2 (N=44). Partial box plots show 25th, 50th (median), and 75th percentiles. Note that the y-axis is logarithmic.

There were 15 main Web pages, plus an additional 3 sub-pages for page 4, making 18 Phase 1-2 pairs we can correlate. This should say something about *relative viewing time* and *consistency of preferred pages* between Phase 1 and Phase 2.

In Figure 4 (left), Web pages below the 45° dashed "identity line" showed an *in*crease in groupwise median page view duration from Phase 1 to Phase 2; pages above the identity line showed a *de*crease.[5] In comparing all 18 page-pairs, the pattern of median view durations showed no significant change between Phases 1 and 2 ($p_{Wilcoxon}$ = .14, NS).[6] Stated another way, the group behaved only a bit differently the first time compared to the second. Some pages were viewed longer, others less.

In contrast, Figure 4 (right) shows how the raw *number of page views* decreased significantly (> 20%) from Phase 1 to Phase 2 ($p_{Wilcoxon}$ = .0002). In fact, the best-fit (solid) line of Figure 4 (right) shows that the pattern of this decrease was remarkably uniform across pages (r_s = .972, p <.000001).[7] Moreover, the statistical significance was not just trivially due to the loss of 6 pilots in Phase 2 because the horizontal (x) axis of Figure 4 (right) reflects

[5]The low Phase 1-2 correlation (embodied by the linear fit line) was partly due to large decreases in Phase 2 median view duration for 2 outliers—pages 7 and 11—the only two text-based pages (all others were graphical). Page 7 was winds/temps as text; page 11 was the area forecast (FA) as text. From this, we might be tempted to argue that text-based products are less efficient than graphical products, but that would not take into account differences in the amount of information per page. Plus, when pages 7 and 11 are ignored, $r_s \rightarrow .399$—less of a change than might be expected.

[6]An alternate way to look at the pattern was to note a correlation between each page's Phase 1 and 2 median views, but not a significant one (r_s = .374, NS).

[7]Almost all the data points fall above the 45° (dashed) identity line, indicating more time spent in Phase 1 than in Phase 2.

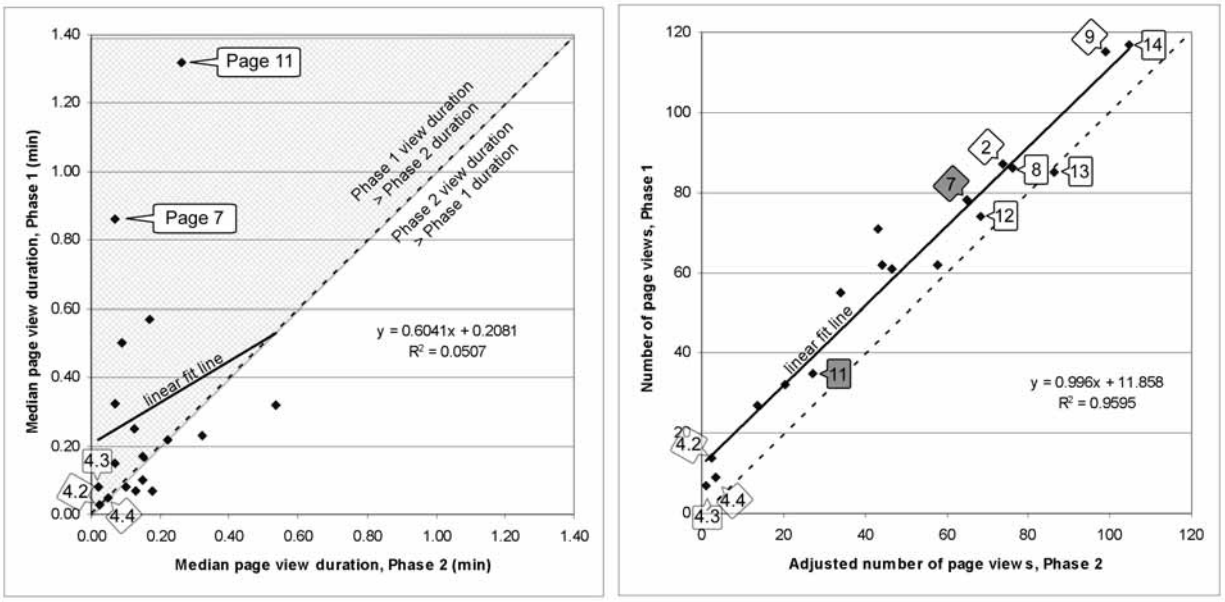

Figure 4. Left: Scatterplot of each page's median view duration for Phase 1 (y-axis) vs. Phase 2 (x-axis). Right: Scatterplot of numbers of times each page was viewed by all pilots considered as a single group (page views).

correction for that attrition (all Phase 2 page views were multiplied by 50/44).

Did the overall decrease in page views mean that, by Phase 2, pilots started disregarding their preflight weather briefings? Not necessarily. The simplest explanation for the drop may be just that pilots had already used the computerized system once. If some remembered which pages they preferred (or did not prefer), that would predictably decrease the overall Phase 2 page views.

In a sense, some pages did seem "more popular" than others. In Figure 4 (left and right) the "least popular" pages are clustered near the lower left, and the "most popular" near the upper right. For instance, the page 4 variants (Convective Outlook) uniformly received very few views. In contrast, pages 9 (TAF Java Tool) and 14 (Looping NEXRAD) received the highest number of views.

Examining this issue of "popularity" more deeply, we can assume that people tend to dwell on things they find either

1. highly informative or
2. hard to understand

Unfortunately, explanation 1 implies something good, explanation 2 not so good. Moreover, either explanation (or both) may be working in any given pilot for any given page. This confound would have to be untangled in future studies using more sophisticated methods.

For now, let us just imagine a simple "group dwell index," calculated by multiplying each page's median number of *page views* times its median view *duration*. Dwell can then represent "groupwise amount of attention paid to each page.

$$Dwell = N_I * T_I \qquad (1)$$

where N = the *i*th page's number of views and T = view duration (~ is the symbol for "median"). Figure 5 shows the result.

For example, Figure 5 (right) shows that the page 4 variants (convective outlook) were low-dwell pages in both Phases 1 and 2. Again, pages below the 45° dashed identity line (e.g., 8, 12, 15) gained dwell in Phase 2.

Higher-dwell pages of interest (labeled in Figure 5) were

page	content	Dwell increased during Phase 2?
07	Winds/temps (text)	No
09	TAF Java tool	No
11	FA (area forecast, text)	No
13	METAR Java tool	No
12	PIREPs	Yes
08	Prog charts (surface)	Yes

Interestingly, PIREPs gained considerable dwell in Phase 2—even though there were no PIREPs. The reason is unclear. METARs and TAFs may not truly have lost value in Phase 2—just dwell. These pages were based on a mouse float-over system (floating the mouse cursor over a small black "station-reporting" rectangle caused METAR or TAF text to pop up). So, their decrease in dwell perhaps only reflected the fact that it took less time for pilots to figure them out in Phase 2.

Flight behavior. Figure 6 shows the Phase 1 flight paths (left column) and Phase 2 flight paths (right column), grouped by training product (rows 1-3). Non-instrument-rated pilots' paths are shown as black lines; instrument-rated pilots' are white lines.

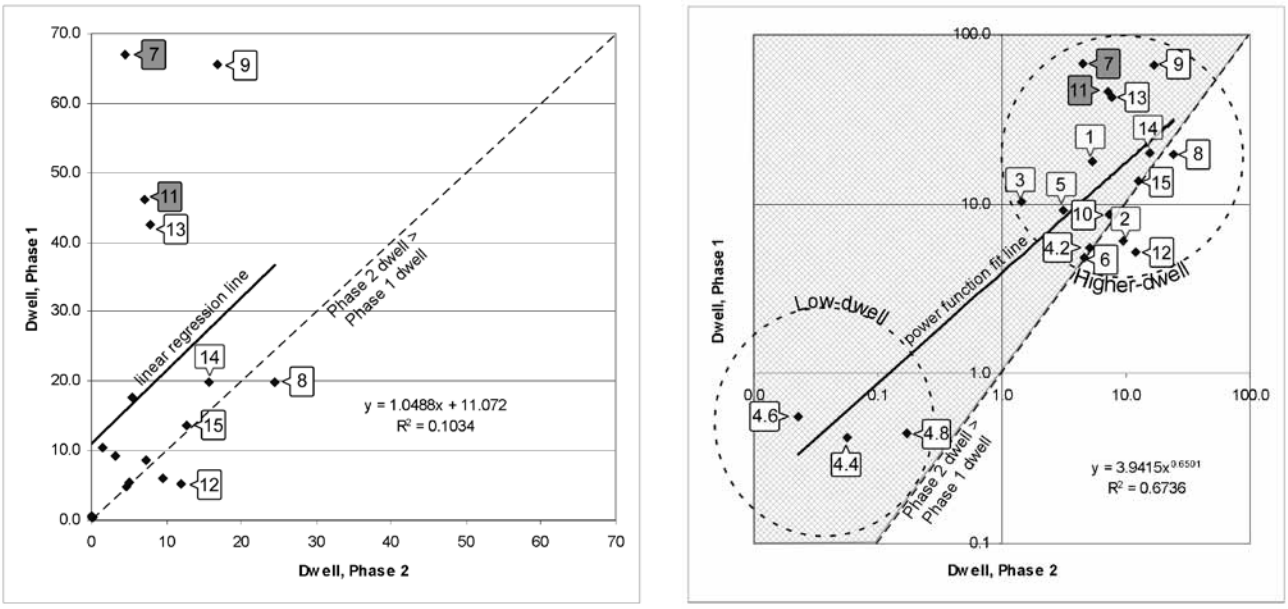

Figure 5. Scatterplot of our value metric (group dwell) for each Web page (left), plotted as a linear-linear, with linear best-fit line (right) plotted as a log-log, with power function best-fit line. Again, Phase 2 values are adjusted for plot attrition. The log-log plot (right) makes it easier to see values near (0,0), which are obscured in the linear plot (left).

Figure 6 shows us mainly how similar the flight patterns were, across training groups, instrument rating, and Phases 1 versus 2. Most pilots who did complete the flight picked their way relatively directly, straight through the mountain passes east of ABQ, even though this required great care to simultaneously maintain adequate cloud and ground clearance. Interestingly, the one Phase 1 Control group instrument-rated pilot (lower-left box, white path) seen to flank the weather by flying north and later south in Phase 2 was the same pilot. He reported that, after being successful the first time, he was simply trying the same strategy, just going the opposite direction.

We can check flight pattern consistency—namely, if a pilot did or did not make it all the way to ABQ during Phase 1, did he or she do the same thing in Phase 2? Table 7 shows the 2x2 consistency matrix for 42 pilots.[8]

Most pilots behaved more consistently than expected by chance (gray-highlighted cells, 22+11=33 of 42 = 78.6%, $p_{Fisher's\ Exact\ Test}$ = .0007). Pilots tended to repeat whatever flight decision they made the first time (e.g., if they flew all the way to ABQ in Phase 1, they generally did so in Phase 2 as well).

Next, looking only at Table 7's 9 *in*consistent pilots, we might ask whether the 6 Phase 2 "new risk-takers" significantly outnumber the 3 "new diverters." However, odds-ratio analysis disconfirms this (p_z=.32, NS), meaning there was no statistically greater tendency for risk-taking in Phase 2 than in Phase 1.

Finally, we logically tried to see if some of our independent variables might correlate with consistency. The 3 "new diverters" were coded as -1 (signifying a decrease in risk-taking), consistent pilots as 0, and "new risk-takers" as 1 (signifying an increase in risk-taking). However, subsequent Spearman correlations between consistency and major independent variables (pilot age, flight hours, instrument rating, average weather knowledge test score and elapsed time [[pre+post+final]/3], and pilot's locality of residence) were all non-significant.[9]

Thus, we can offer no explanation for this consistency, other than to note the universal human tendency for many of us behave in similar ways over time when faced with similar circumstances.

AGARS intercorrelations. Table 8 shows nonparametric Spearman correlations between key variables. Statistically significant correlations are highlighted in gray. Point-biserial correlations (r_{pb}) are used when one variable is dichotomous, the other continuous.

Trivial relations. Four of the largest correlations are statistically significant, but trivial. These cells are marked in a light shade of gray.

1-2. Instrument rating x pilot age (r_{pb} = .382) and instrument rating x flight hours (r_{pb} = .379) simply mean that instrument-rated pilots tend to be older and have more flight hours.

[8] 44 completed both Phases 1 and 2, but 2 experienced controlled flight into terrain and so were excluded from this analysis

[9] Of course, the -1,0,1 coding scheme was logical, but purely speculative and arbitrary. The Spearman correlation was also not the method of choice for some of these tests. However, given that none of the resulting 2-tailed *p*-values was smaller than .10, readers may excuse us for not pursuing the matter.

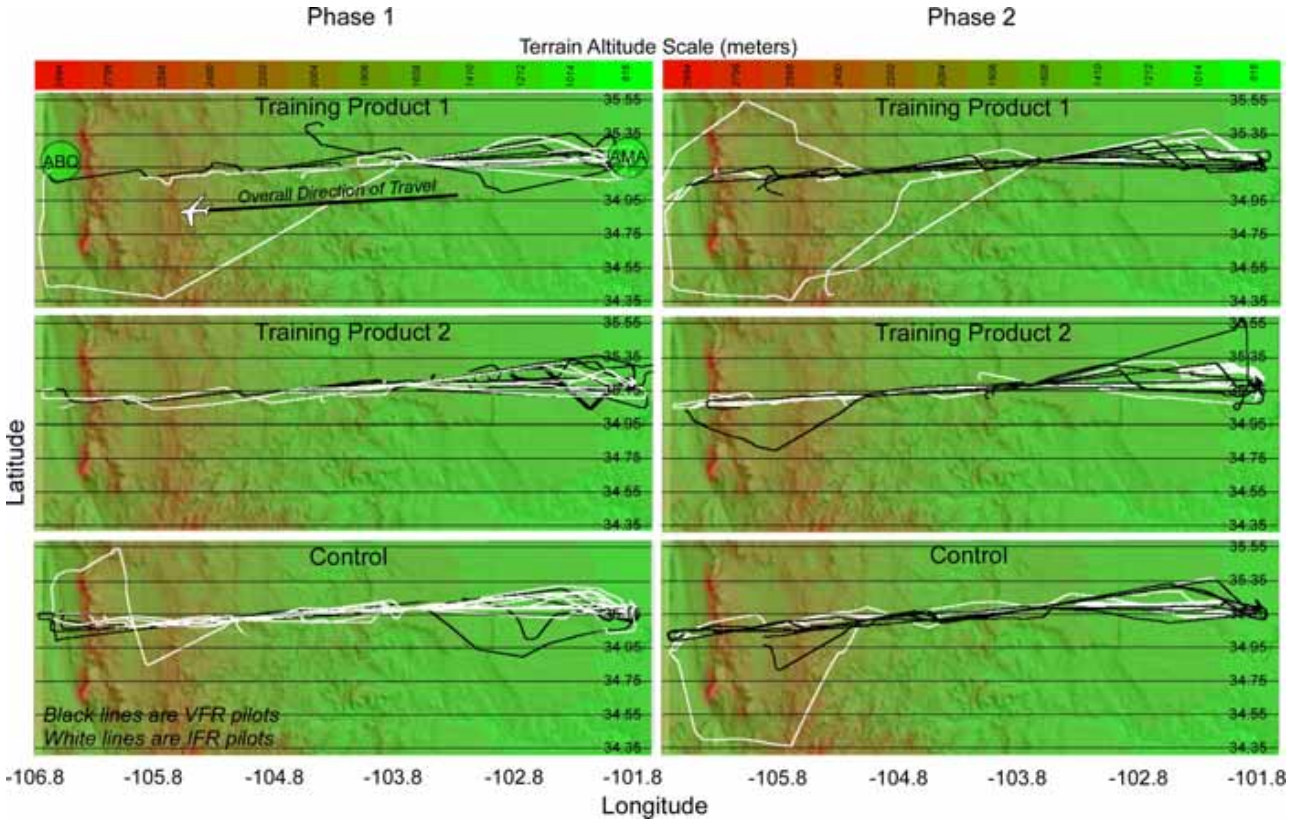

Figure 6. Top-down, flight-profile views for Phases 1 and 2, laid over the terrain map. Terrain slowly rose as pilots flew east to west, squeezing them between clouds and ground, especially near the north-south ridge just before ABQ. Digital elevation data were obtained from National Geophysical Data Center (2008) and drawn by Mathematica (2008).

Table 7. Consistency of flight decisions from Phase 1 to 2.

		Phase 2	
Phase 1		No	Yes
Made it to ABQ	No	22 (16.7)	6 (11.3)
	Yes	3 (8.3)	11 (5.7)

1st number is actual n (2nd is expected n). p_{Fisher} = .0007 (2-tailed)

Table 8. Correlations between key Phase 2 variables.

	Variable 2									
Variable 1	Instrument Rating (1=instrument-rated)[1]	Locality of Residence (1=Local)[1]	Pilot Age[2]	Pilot Flight Hours[2]	Ave. Wx Knowledge2, [4]	Web Pre-flight duration[2]	Flight Duration[2]	Minimum Dist to ABQ[2]	Minutes scud running[2]	Minutes in IMC[2]
Instrument Rating	1.0									
State of Residence	[3]	1.0								
Pilot Age	.382 (.010)	-.275	1.0							
Pilot Flight Hours	.379 (.011)	.060	.736 (<.001)	1.0						
Ave. Wx Knowledge[4]	.247	-.255	-.066	.154	1.0					
Web Preflight Duration	-.058	**-.421 (.004)**	**.475 (.001)**	**.371 (.013)**	.225	1.0				
Flight Duration	.130	-.281	-.107	.043	.070	.109	1.0			
Minimum Dist to ABQ	-.083	.215	.231	.030	-.064	-.023	-.908 (<.001)	1.0		
Minutes scud running	.077	.206	-.049	-.037	-.036	.038	.090	-.194	1.0	
Minutes in IMC	.088	-.202	-.148	-.235	-.260	-.076	.138	-.248	.234	1.0
Minutes < 500' AGL	-.147	.034	-.139	-.053	-.175	.049	**.373 (.013)**	**-.422 (.004)**	-.033	**.446 (.002)**

[1] r_{pb} = Point-biserial. [2] r_s = Spearman rho. Low p-values are in parentheses (all others are non-significant (NS)); [3]No correlation run because sample had been partitioned for these factors. [4](pre-test+post-test+final test)/3

Table 9. Durable, non-trivial relations between Phases 1 and 2 variables.

	Variable 2									
Variable 1	Instrument Rating (1=inst rated)[1]	Locality of Residence (1=Local)[1]	Pilot Age[2]	Pilot Flight Hours[2]	Ave. Wx Knowledge[2]	Web Pre-flight duration[2]	Flight Duration[2]	Minimum Dist to ABQ[2]	Minutes scud running[2]	Minutes in IMC[2]
Instrument Rating	1.0									
State of Residence		1.0								
Pilot Age			1.0							
Pilot Flight Hours				1.0						
Ave. Wx Knowledge[4]					1.0					
Web Preflight Duration		-.348 (.013) Ph1 -.421 (.004) Ph2	.417 (.003) Ph1 .475 (.001) Ph2			1.0				
Flight Duration							1.0			
Minimum Dist to ABQ								1.0		
Minutes scud running									1.0	
Minutes in IMC										1.0
Minutes < 500' AGL										

[1] r_{pb} = Point-biserial. [2] r_s = Spearman rho. *P*-values are in parentheses.

3. Pilot age x flight hours (r_s = .736) merely means that older pilots tend to accumulate more flight time.
4. Flight duration x minimum distance to ABQ (r_s = -.908) only means that the longer pilots fly, the closer they tend to get to the destination (ABQ).

Non-trivial relations. Discounting trivial relations, a few interesting ones remain. In Table 8, these are boldfaced and marked in a darker shade of gray. They are:

1. Web preflight duration x Locality of residence(r_{pb}=-.421)
2. Web preflight duration x Pilot age(r_s=.475)
3. Web preflight duration x Pilot flight hours(r_s=.368)
4. Minutes < 500' AGL x Flight duration(r_s=.373)
5. Minutes < 500' AGL x Minimum distance to ABQ(r_s=-.422)
6. Minutes < 500' AGL x Minutes in IMC(r_s=.446)

Correlations 1-3 imply that non-Oklahoma pilots, older pilots, and higher flight-hour pilots tended to spend slightly more time using the Web preflight briefing tool. The effect sizes were no more than modest, accounting for r^2_{pb} = 18, 23, and 14% of the variance, respectively.

Correlations 4-6 imply that pilots who flew longer, got closer to ABQ, and those who spent more time in IMC tended to spend slightly more time < 500' AGL. Effect sizes were also modest, accounting for r^2_{pb} = 14, 18, and 20% of the variance, respectively.

Durability of non-trivial relations. A "durable" relation can be defined as one remaining statistically significant across both Phases 1 and 2. Table 9 shows the only two durable relations.

1. Web preflight duration x Locality of residence (r_{pb}=-.348/-.421)
2. Web preflight duration x Pilot age (r_s=.417/.475)

In other words, both local pilots and younger pilots spent slightly less time on their Web weather preflight briefing. Arguably, local pilots were more familiar with local terrain and weather patterns. And, older pilots may have been either slightly more careful briefers, or might have simply been a bit less familiar with Web-based briefing, especially *www.aviationweather.gov*.

Effect of pilot weather knowledge on subsequent flight safety. As in Phase 1, having higher weather knowledge (as measured by these test questions) did not significantly predict flight safety. As Table 8 showed, average weather

knowledge[10] did not correlate significantly with any Phase 2 flight behavior variables.[11] Also as in Phase 1, neither did it seem related to age, flight hours, locality of residence, or instrument rating (r_s = -.066, .154, r_{pb} = -.255, .247 respectively, all NS).

Effect of Web preflight briefing time on subsequent flight safety. Were pilots who spent more time on their Phase 2 Web-based weather briefing safer pilots? As in Phase 1, not significantly. Table 8 shows all non-significant Spearman correlations of Web preflight duration with flight duration, minimum distance to ABQ, minutes scud running, minutes in IMC, and minutes < 500' AGL, with correlation values ranging from $-.076 \leq r_s \leq .109$.

We now have a bit more sophisticated sense of how these pilots used the Web-emulation to brief themselves on the weather. It might seem logical to explore relations between, say, individual page view durations and our flight safety variables. However, there is arguably far too much variability in the data to be able to do this confidently.[12]

Effect of the weather training products on takeoff hesitancy. In Phase 1, we saw evidence that the weather training product may have induced takeoff hesitancy. Did this hesitancy persist over time? In Phase 1, 12 of 50 pilots initially stated that having to fly this mission VFR, they would choose not to even take off. In Phase 2, 7 of the 44 returning pilots made the same decision. Overall, the degree of change was not significant (p_{X2} = .330, NS). Stated the opposite way, Phase 1 and Phase 2 decisions were significantly correlated[13] r_ϕ = .323 (p = .032).

Table 10 shows that 4 + 30 = 34 pilots (the 2 cells highlighted gray) repeated their Phase 1 takeoff decision in Phase 2, while 3 + 7 = 10 (23%) reversed their decision—7 of those 10 (70%) being pilots who formerly did not want to take off, who now did want to take off, even though the flight situation was essentially identical to Phase 1.

Table 10. Takeoff decision across phase for 44 pilots who participated in both Phase 1 and 2 ("Yes" means "Would take off").

		Phase 2	
		No	Yes
Phase 1	No	4 (1.8)	7 (9.3)
	Yes	3 (5.3)	30 (27.8)

1st number is actual n (2nd is expected n)

Table 11 shows numbers of Phase 1 and Phase 2 pilots who initially hesitated, versus the values expected by chance (in parentheses). In Phase 1, we saw significant effect (lack of hesitancy) centered in the Control group. In Phase 2, this effect was not significant. The Yates-corrected Phase 2 p_{X2} is .554, implying that the training groups did not differ significantly.

This implies that training product-induced takeoff hesitancy did *not* persist over time. This is important because Phase 1 evidence for effect of training products was sparse and rested on the effect of a 3-variable model (consisting of training product + pilot age + takeoff hesitancy) to predict whether or not pilots would complete the entire flight to ABQ. Were hesitancy to cease to exert an effect, that would probably obviate the model.

In Phase 2, this hesitancy did appear to diminish and 2 changes seemed to drive it. First, as Table 11 shows, the Yes/No ratio for training product 2 changed from 9/7 to 12/3. Second, the Control group's Yes/No ratio went from 17/1 to 13/1 (due to the serendipitous attrition of 4 Phase 1 non-hesitating controls).

Serendipity or not, unlike Phase 1, no assertion can now be made that the training product induced takeoff hesitancy in Phase 2. This finding is critical, because it destroys a putative causal chain, namely that:
training product → takeoff hesitancy → degree of penetration into adverse weather → degree of risk

This causal chain was Phase 1's only viable candidate model for asserting that a 90-minute weather training product might have an demonstrable effect on pilot behavior—the primary motivation for doing this study. Now, the Phase 2 analysis implies that, if a "Takeoff Hesitancy Effect" ever existed, it was not durable.

Effect of takeoff hesitancy on subsequent flight safety. The above argument reduces to the following question: Did the 7 Phase 2 hesitators perhaps still end up somehow flying more safely than the remaining 37 pilots?

Not appreciably. There were no significant differences between hesitators and non-hesitators for minutes spent in IMC, minutes scud running, or minutes < 500' AGL (2-tailed Mann-Whitney p_U = .268, .089, .950 respectively, all NS). In Phase 1, hesitators seemed to continue their conservatism into their flight, making significantly briefer

[10](pre+posttest+final score)/3.
[11]Spearman correlations with flight duration, minimum distance to ABQ, minutes scud running, minutes in IMC, and minutes < 500' AGL ranged from $-.260 \leq r_s \leq .07$, all NS.
[12]First, the variation in numbers of page views-per-dependent variable (DV) was enormous (range 1-92, mean 44.8, SD 28.0), meaning that correlations and models would either be based on wildly different numbers of cases or would be saddled with huge numbers of zero values. Second, the frequency distributions for the 18 Web pages' durations were, without exception, unacceptably non-normal for parametric techniques. Even excluding non-zero values, all Shapiro-Wilk *p*s were ≤ .001 except Collaborative Convective Forecast Product (CCFP) = .011 and Convective Outlook-Wind = .183 (but, which was based only on n=3). Currently, there is no widely accepted method of nonparametric multiple regression. So, in short, we would mistrust the results.
[13]The correlation used here was ϕ (phi), which measures the relation between 2 dichotomous variables (in this case, Phase 1 hesitancy yes/no vs. Phase 2 hesitancy yes/no).

Table 11. Takeoff hesitancy, Phase 1 vs. 2.

		Phase 1			Phase 2		
		Trg Prod 1	Trg Prod 2	Control	Trg Prod 1	Trg Prod 2	Control
Initial takeoff decision	Yes	12 (12.2)	9 (12.2)	17 (13.7)	12 (12.6)	12 (12.6)	13 (11.8)
	No	4 (3.8)	7 (3.8)	1 (4.3)	3 (2.4)	3 (2.4)	1 (2.2)
Pairwise odds-ratios, 1-tailed *p*		← .152 →			NS		
			← .004 →				
		← .037 →					

flights, with consequently less penetration into the marginal weather close to ABQ. This was not true in Phase 2 (p_U = .550, .450, respectively, NS).

Effect of the weather training products on subsequent flight safety. When all was said and done, did viewing a weather training product significantly affect flight safety? In Phase 1, there was *indirect* indication of this. Seeing a weather training video related to takeoff hesitancy, which related to flight duration, which related to minutes spent < 500' AGL—although seeing the weather video did not significantly *directly* relate to minutes spent < 500' AGL (nor to scud running or time spent in IMC).

However, as we saw, the same indirect correlational chain did not seem at work in Phase 2. Nor could any direct relation be seen between training product and subsequent flight safety, as measured by flight duration, minimum distance to ABQ, minutes in IMC, minutes scud running, or minutes < 500' AGL (.154 < $p_{Kruskal\text{-}Wallis}$ < .768, NS).

IMC penetration. Despite highly emphasized instructions to fly VFR-only, a small number of Phase 1 and 2 pilots still penetrated IMC. Phase 2 penetration ranged in duration from 0.03-84.7 min. Most pilots avoided IMC altogether, and the number of penetrations decreased from Phase 1, although not technically significant when duration was analyzed.[14]

We can argue that those who spent less than a minute in IMC probably did so inadvertently. Seven Phase 2 pilots spent more than 1 minute in IMC (versus 16 in Phase 1). Five spent more than 4 minutes (versus 10 in Phase 1). Figure 7 illustrates there were only 3 "repeat offenders" for long-duration penetration (>4 min) across both studies—2 of these were VFR pilots, all resided outside Oklahoma[15]—providing no solid evidence why these 3 particular individuals behaved as they did. It may prove useful to explore this in a follow-up report.

Modeling Flight Behavior

One of the questions we want to answer is, *"What differentiated pilots who chose to complete the flight through deteriorating weather from pilots who chose not to complete the flight?"* To investigate this question meant constructing models—simplifications of the situation that still captured its essential features.

Correlations are, at heart, the simplest possible kinds of model—a set of relations between single variables. Table 8 was a good place to start investigating these relations. However, it is rare that real-world events are ever completely explained by one, single factor. More often, multiple causative (or facilitative) factors are at work simultaneously. To try to get at such multi-factor relations, we turned to multivariate modeling.

Cluster analysis. In Phase 1, we saw how a subset of the candidate variables formed 2 significant similarity clusters (Table 12 was derived from Table 6 of the Phase 1 report).

In Phase 2, however, a repeat cluster analysis failed to find any variables related sufficiently to sort the pilots into even 2 clusters. The logical significance of this will become apparent shortly.

Binary logistic regression analysis. In Phase 1, stepwise forward likelihood-ratio binary logistic regression analysis produced the 3-variable model referred to earlier. However, when this was again attempted, it was unproductive. The candidate variables from Table 12 not only failed to cluster but also failed to significantly predict the binary outcome variable *ToABQ* (meaning whether or not a given pilot made it all the way through the deteriorating weather to the destination ABQ).

We next retested the Phase 1 3-variable logistic model for durability. In Phase 1, a model based on training product, age, and takeoff decision[16] was able to predict 64% of the explainable (Nagelkerke) variance in *ToABQ* (p=.000004) and correctly predicted 83.3% of the cases

[14]Wilcoxon signed-rank Phase 1-2 difference test results for a) all IMC penetrations, b) > 0.5 min, c) > 4 min = p = .614, .140, 158, respectively, (all NS).

[15]Given that about half the pilots were Oklahomans, the odds of 3 randomly selected pilots all being non-Oklahomans is still roughly $1/2^3$, or .125 (NS), meaning it can be considered a chance occurrence.

[16]Note that takeoff decision reflects "takeoff hesitancy" as discussed earlier.

Table 12. Phase 1 variables contributing significantly to clustering.

Continuous	Categorical
Age	Training product
Flight hours	Takeoff hesitancy
Final minimum distance to ABQ	Wx recheck just before takeoff
Minutes flying < 500' AGL	Flew all the way to ABQ
Clustering	
Cluster 1 tendencies	Cluster 2 tendencies
Younger	Older
Lower flight hr	Higher flight hr
Closer final minimum dist to ABQ	Farther final minimum dist to ABQ
More minutes at < 500' AGL	Fewer minutes at < 500' AGL
Control group (no wx trg product)	Received a wx training product
Initial takeoff response was to fly	Initial takeoff response was to not fly
No wx check just before takeoff	Wx check just before takeoff
Flew all the way to ABQ	Diverted before ABQ

(diversion vs. continuation on to ABQ), compared to a baseline prediction rate of 62.5%.[17]

In retesting this model with Phase 2 data, however, the identical model predicted only 19.9% of the Nagelkerke variance and 72.1% of the cases, compared to its baseline rate of 60.5%. This was not a significant improvement (p=.145, NS) over an educated guess (that is, the baseline, "constant-only" model).

This performance degradation of the Phase 1 model was certainly not simply due to the raw number of pilots who actually made it all the way to ABQ (18 of 48[18] in Phase 1 vs. 17 of 43 in Phase 2). Nor was it due to pilot age (because pilots were only 3-4 months older than they were during Phase 1).

By default, the Phase 1 model seemingly collapsed due to inconsistencies in takeoff decision from Phase 1 to Phase 2. In other words, whatever coherent effect, or "signal," the weather training products may have engendered in Phase 1 dissolved amongst the "noise" of individual variation in Phase 2. Table 10 showed the 2x2 consistency matrix. Ten pilots (3+7) reversed their Phase 1 takeoff decisions in Phase 2. While Fisher's Exact Test gives this probability at p=.054—technically non-significant—it is arguably close enough to suspect that we have the culprit that disabled the Phase 1 3-variable model.

[17]The baseline rate is the model's ability to predict an outcome by chance alone, given only knowledge of average group behavior. For instance, if 50% of Americans voted Democratic, that would be the baseline rate, assuming we had no other knowledge about individual voters. However, if we knew individual voters' incomes, educational levels, ethnicities, genders, religious preferences, and job categories, we might expect to predict individual votes at greater than a 50% success rate. The primary purpose of multivariate modeling is to maximize that kind of additional predictability.

[18]The original Phase 1 N=50, with 1 missing data, 1 eliminated for CFIT ® 48. The original Phase 2 N=44, with 1 eliminated for CFIT ® 43.

DISCUSSION

The purpose of this research was to investigate the following questions:

1. Video weather training products
   a. What are the immediate effects on pilot weather knowledge and flight behavior in the face of potential instrument meteorological conditions?
   b. Do these effects persist over time?
2. How are modern Web-based weather products used during preflight briefing?
3. Do Oklahoma pilots differ appreciably from non-Oklahoma pilots in either weather knowledge or weather-related flight behavior?

Phase 1, Brief Summary

Main findings, Question 1. In Phase 1 of this project, 50 GA pilots participated in a study designed to collect data on both pilot weather knowledge and flight behavior. Pilots took a weather knowledge pre-test, followed by exposure either to 1 of 2 weather training videos (the Experimental groups), or to a video having nothing to do with weather (the Control group). They then took a knowledge post-test to measure knowledge gain induced by the training product. Next, they planned for, and flew, a simulated flight mission through deteriorating weather from Amarillo, TX, to Albuquerque, NM (ABQ). Numerical flight data were collected and flight behaviors noted.

A limited number of significant effects were seen in Phase 1. For one, there was a tendency for pilots who viewed either of the 2 weather videos to hesitate taking off into the marginal weather. These "hesitators" flew only after encouragement. In contrast, 17 of 18 control group pilots took off without any encouragement.

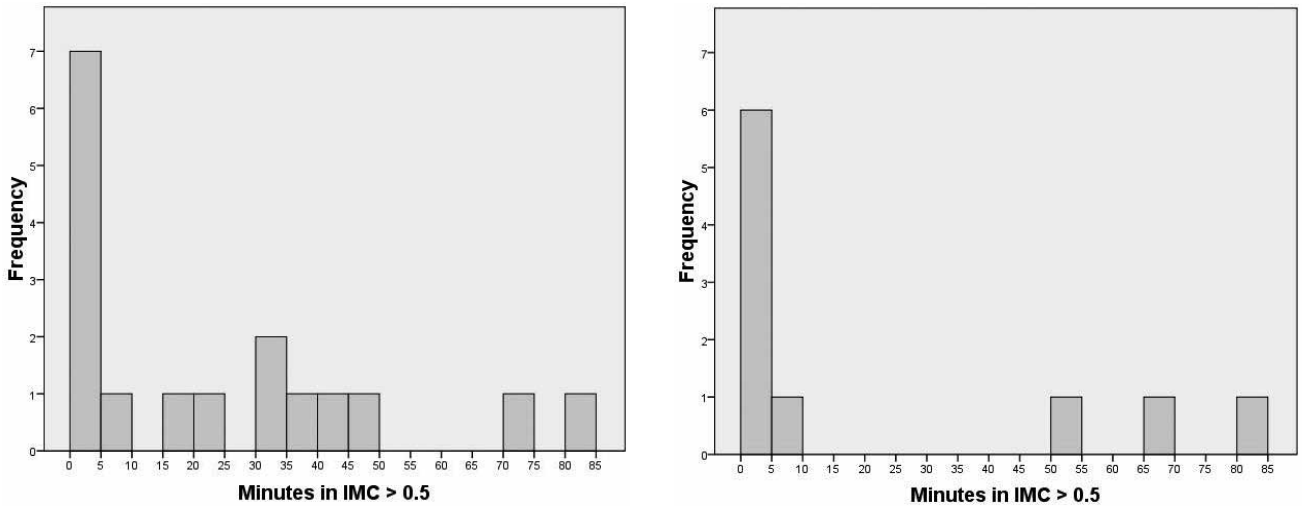

Figure 7. (Left) Number of Phase 1 pilots (y-axis) by time-spent-in-IMC (x-axis) for the time range 0.5 – 82.5 minutes. (Right) Phase 2 pilots, showing drop-off in long-term IMC dwell time.

Subsequently, the hesitators continued their conservatism, tending to make shorter flights than non-hesitators. Since the bulk of the scenario's danger lay near the flight's destination, we might conclude that watching a weather training video induced *takeoff hesitancy*, which then induced shorter flights, which then led to lower group-wise risk exposure. There was correlational evidence to support each individual link of this chain.

However, the overall chain of logic was not that simple. Ultimately, none of the beginning-state variables (e.g., pilot age, flight hours) ended up *directly* correlating with the end-state flight-risk variables (scud running, time spent in IMC, or time spent at < 500' AGL). Therefore, no simple model based on video training product alone could be ultimately shown to modulate flight risk.

Therefore, to try to explain weather-related risk-taking, we next turned to multivariate modeling. Non-normalities in the data frequency distributions[19] precluded the use of more-powerful parametric multiple regression statistical techniques. So, a binary logistic regression model was used, which took multiple candidate variables (continuous and discrete) to find the best combination capable of explaining the variation in a single, discrete dependent variable—namely, whether or not a pilot completed the entire flight to ABQ. This model was based on the assumption that the farther one flew into the deteriorating weather, the greater the overall risk.

In Phase 1, that analysis led to a reasonably well-performing, 3-variable model of weather-related risk taking based on

- Video training product
- Pilot age
- Takeoff hesitancy

[19]Key data frequency distributions are shown in Appendix A.

We can interpret this to mean that—if a brief video weather training product did exert a measurable effect—it probably did so less by directly influencing pilots' weather knowledge or preflight briefing habits and more by appealing to a subset of somewhat older pilots with latent conservative behavioral (flying) tendencies to begin with. Those pilots may arguably have been sensitized, less by specific training product content than by the fact of having been exposed to "something safety-related," which subsequently made them more cautious about the specific flight scenario to which they were exposed.

Secondary findings. As Figure 10 (Appendix B) shows, in Phase 1, there were slight-but-significant tendencies for older pilots to spend a bit more time on their Web-based preflight briefing, whereas local pilots spent less time on it. Instrument-rated pilots spent slightly less time too close to the ground (< 500' AGL). So did higher flight hour pilots. However, instrument-rated pilots also tended to be older, *with* higher flight hours. So, it was difficult to pinpoint whether rating, age, or flight hours was most related to ground clearance.

There was also a slight tendency for older pilots and higher flight hour pilots to fly shorter flights (meaning they penetrated the weather slightly less). Age and experience may engender some risk aversion. Alternatively, younger pilots might have been merely "gaming the system," treating the flight more like a game than a real flight. It is difficult to say.

As an incidental side effect noticed during this research, these pilots' long-term retention of weather knowledge proved somewhat lower (19%) than national norms on FAA certification exam scores obtained by freshly licensed pilots. This may or may not be an important issue having bearing on how the FAA constructs test questions and assesses weather competency.

However, it is sufficiently complex to merit review in a separate "Phase 3" report.

Finally, one instance of actual controlled flight into terrain (CFIT) was seen in Phase 1. This underscored the humbling fact that genuine accidents rarely follow the exact pattern implied by group statistics. This incident happened because of nothing more elaborate than momentary in-flight attentional lapse while the pilot was studying the sectional. It was a Control group pilot, not younger or lower flight hour, as our model would have led us to believe.

Phase 2, Brief Summary

Main findings, Question 1. As Figure 11 (Appendix B) shows, the pattern of non-trivial relations between variables differed substantially from Phase 1. Most relations did not replicate, other than trivial ones.

As previously mentioned, particularly salient was the failure of the 3-variable Phase 1 model to again significantly predict risk-taking behavior in Phase 2. In Phase 2, not only was there no significant evidence of a direct effect of the training products on anything, but even the partial (or, perhaps, indirect) effect represented by the Phase 1 3-variable model disappeared.

The most straightforward explanation for the 3-variable model's collapse may simply be what behavioral psychologists call *desensitization*. In Phase 2, pilots were under less pressure: They had fewer tests to take; they were already familiar with equipment and procedure; they now fully understood that the FAA researchers were benign. They could relax and act naturally. And, when they did, it became hard to discern any great differences between the weather training groups and the Control group.

In the Phase 1 report, we discussed a possible *cognitive priming hypothesis*. First-time exposure to the weather training product could have "tipped off the participants" that the study was to be about weather. Given the context of FAA officials conducting an experiment within an FAA facility, a devil's advocate could argue that any Phase 1 effect of training product might owe more to priming about weather and risk than to any permanent cognitive or behavioral change the products themselves might have induced. Most likely, any effect was affective (emotional), and merely decayed/desensitized over time.

Secondary findings. As in Phase 1, a single episode of controlled flight into terrain also occurred in Phase 2. Figure 8 shows the flight profile for this individual, an older, instrument-rated, high flight hour pilot. As in Phase 1, the explanation was not complex—the pilot appeared to suffer a lapse of attention and ran straight into the rising terrain (Figure 8, left). The flight notes recorded that the individual had unusually great initial trouble mastering the controls of the Malibu. Shortly after takeoff, there was direct ascent into brief IMC. Recovery from this led to direct, steady descent, which nearly resulted in impact with the ground (Figure 8, right, minimum AGL = 9.78'). Many would agree that such an near-impact could be disorienting and could lead to subsequent "tunnel vision," with difficulty attending to normal multiple piloting tasks, including awareness of terrain clearance.

Findings Common to Both Studies

Durable relations. As Table 9 and Appendix B show, a few non-trivial durable relations persisted from Phase 1 to 2. There was a slight tendency for older pilots, and for Oklahoma pilots, to spend a bit more time on their Web-based preflight briefing. This is readily explained if we assume that older pilots were likely to be slightly less familiar with Web-based preflight briefing, and that local pilots tended to know the terrain better.

Consistency of flight behavior. Table 7 showed that in Phase 2, more than 78% of pilots made the same ultimate choice about either diverting or continuing on to the

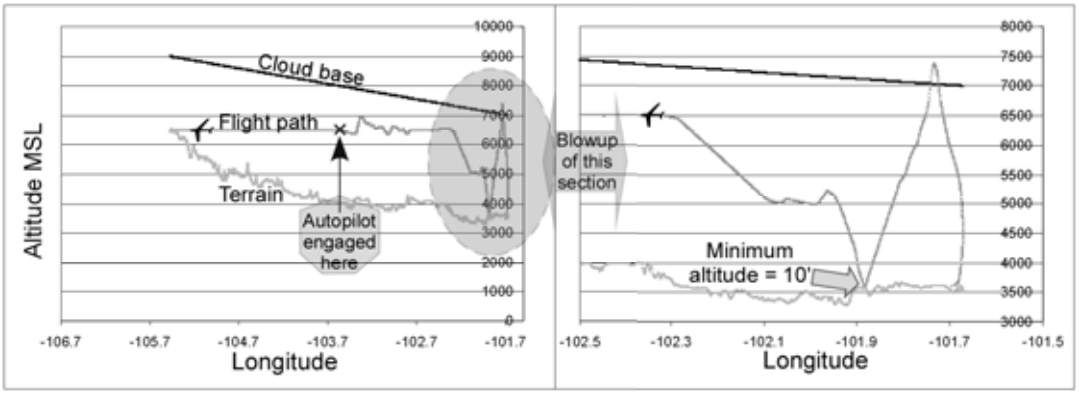

Figure 8. The one Phase 2 CFIT. (Left) Horizontal view of the pilot's flight profile. (Right) Expanded view of the takeoff and subsequent events, including a near-collision with terrain.

destination that they first made in Phase 1. Only 6 pilots grew more daring from Phase 1 to Phase 2, while 3 grew more conservative—not a significant groupwise change. This consistency was remarkable, and what inconsistency there was could not be explained by any of our major independent variables.

Rule violations. Despite this being stressed as a VFR-only flight, in Phase 1, 10 pilots ascended immediately into IMC and spent at least 4 minutes there. In Phase 2, the number dropped to 5, although that decrease was not statistically significant. Three pilots were "repeat offenders" across both phases. This is particularly perplexing, given that these pilots were obviously well aware of being involved in a study sponsored by the aviation regulator, and yet still violated cloud clearances and minimum visibility requirements. At this time, we have no explanation for this behavior but intend to pursue it in a "Phase 3" report.

Question 2: Use of Web-based preflight weather briefing materials. Both Phase 1 and 2 studies show the study of Web-based preflight briefing to be more complex than anticipated. However, we note that the present studies are among the first human factors studies of NWS Web-based preflight briefing to be done (if not the first). Therefore, methodological imperfection is unsurprising.

First, there was an issue with page view duration outliers, as detailed in the Phase 1 report. Second, users take time to explore a new system, so "first impressions" of usage are bound to differ from those of persons trained to an asymptotic skill level. Third, the number and distribution of page views can be inflated by users merely quickly going to a page, finding out there is nothing of value there, and moving on.

Finally, the amount of time spent per page (view duration) confounds two completely different concepts, namely *information importance* and *information difficulty*. In other words, long dwell times do not necessarily mean a page is chock full of information. It can be, but it also can mean that the information on that page is simply hard to access or to understand. Untangling this confound will require more sophisticated methodology, possibly eye-tracking or self-report surveys, and possibly even physiological measurement of workload.

Resolution of these issues will be fundamental to continued human factors study of Web-based preflight briefing. However, such study is necessary, since the Web is so obviously the future of GA preflight weather briefing.

Question 3: Equivalency of local Oklahoma pilots to non-locals. We did not see significant differences between local pilots and non-local pilots. The only significant finding was that locals took slightly less time to brief for this relatively local flight. But, Oklahoma pilots are arguably more familiar with their own local terrain and weather patterns, and need less briefing time for a flight such as this, so that issue is trivial.

Importantly, this addresses the issue of whether CAMI studies are generalizable to the national population of U.S. GA pilots. To recap the Phase 1 report, the fact is that U.S. pilots study a fairly uniform curriculum (largely driven by the licensing exams). This guarantees a measure of pilot uniformity. What is certainly far more important to research planning is the individual variation in knowledge and skill present between one pilot and another—not where a particular pilot happens to live. Yes, there are specific regions where certain flying skills are more called-upon than others. The high winds in the Midwestern U.S. are a good example. But—unless the task to which a given group of pilots is put depends critically on some small, specific set of skills—geographical region-of-residence probably will not matter a great deal.

What this means is that researchers simply need to adhere to good practice in selecting pilots and assigning them to treatment conditions. As long as designs are counterbalanced, and pilots are reasonably well-matched for age, flight hours, and instrument rating over treatment cells, there is probably only occasional need to recruit non-locally. Our final cost figures put the human effort and dollar cost of testing a non-local pilot at approximately 5-10 times the expense of recruiting a local pilot. Therefore, what non-local pilots are probably best used for is precisely when an elite sample is required but not locally available. For instance, if high-hour, young, VFR pilots were needed for some reason, then we would probably want to consider recruiting non-locally.

REFERENCES

National Geophysical Data Center (2008). *NOAA-NGDC-MGG-GLOBE Custom Data Selection Page.* Digital elevation data downloaded Oct. 23, 2008 from www.ngdc.noaa.gov/cgi-bin/mgg/ff/nph-newform.pl/mgg/topo/customdatacd.

Wolfram Research (2008). *Mathematica V7.0.*

APPENDIX A

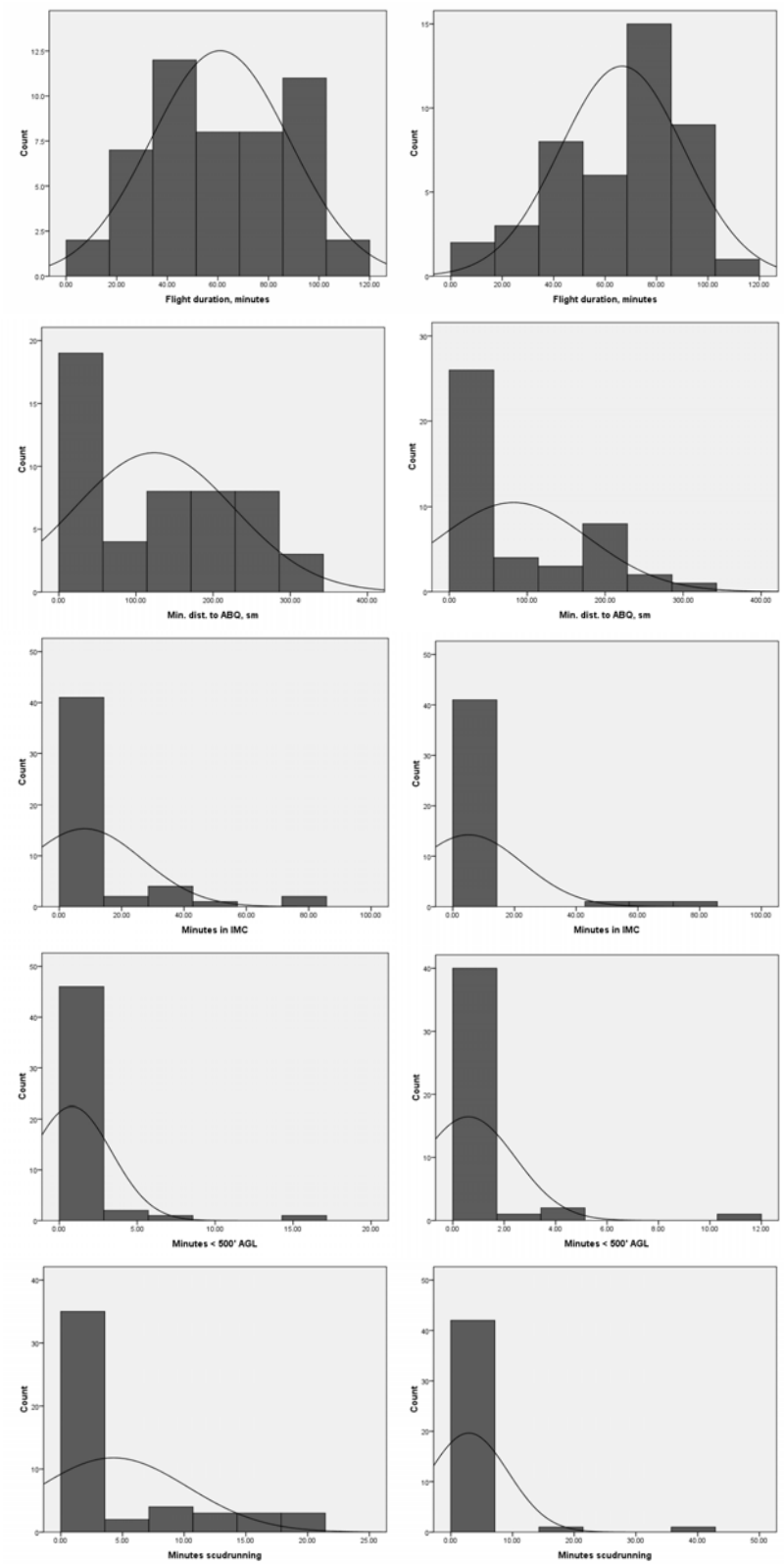

Figure 9. Frequency histograms for numbers of pilots (y-axis) by flight duration, minimum distance to Albuquerque, minutes in IMC, minutes below 500' ground clearance, and minutes scud running (x-axes). Phase 1 data are at left, Phase 2 at right.

APPENDIX B

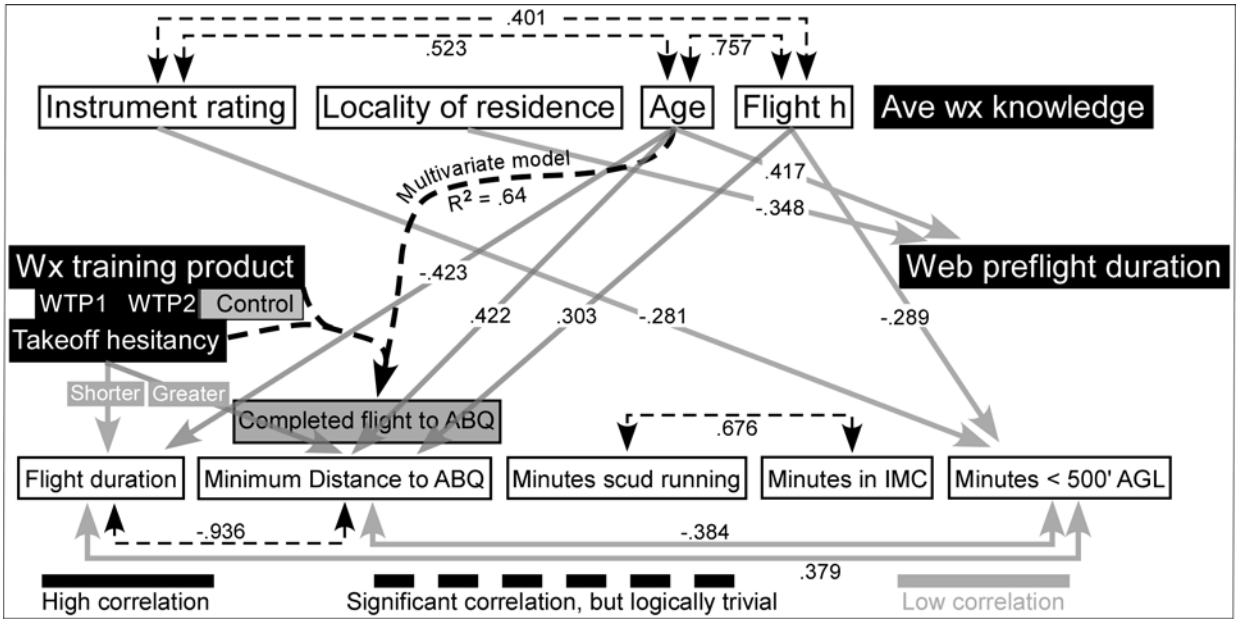

Figure 10. Phase 1 univariate and multivariate correlational structure.

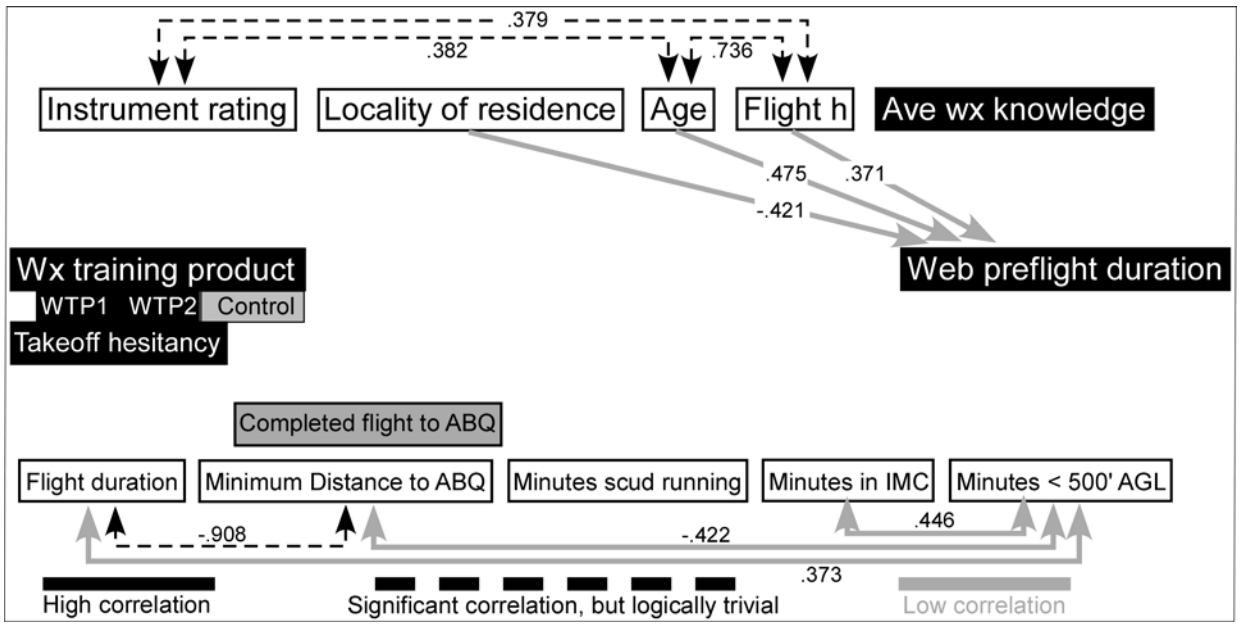

Figure 11. Phase 2 correlational structure.