Federal Aviation Administration

DOT/FAA/AM-10/17
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 3

William R. Knecht
Michael Lenz

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

November 2010

Final Report

OK-11-0024-JAH

NOTICE

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-10/17		

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 3	November 2010
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Knecht WR, 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

The primary purpose of Phases 1 and 2 of this research was to test the effects of video weather training products on weather-related risk-taking. During the investigation, two unexpected observations were made: (1) Despite specific instructions to fly visual-flight-rules-only (VFR), nine of 50 Phase 1 pilots spent more than 10 min in simulated instrument meteorological conditions (IMC), plus three of those nine repeated that behavior in Phase 2; (2) Whole-group (N=50) weather knowledge test scores were significantly lower (19%, $p<.001$) than average FAA certification exam scores obtained by freshly licensed pilots, implying knowledge decay over time.

To assess if any of the IMC violations were willful (rather than inadvertent), we sent a brief questionnaire to the nine pilots of interest. Five responded. After analysis, the leading explanation seemed that their flight profiles were consistent with preflight terrain avoidance planning (TAP). These pilots seemed determined to fly straight and level above the highest known obstacle, even if that obstacle was distant and TAP altitude meant flying initial VFR-into-IMC.

The average group decline in certification exam scores was equally significant from a logical standpoint. Since knowledge *retention* tends to be a function of knowledge *relevancy*, if FAA test questions were uniformly relevant to real-world weather encounters, we would expect pilots' scores to increase with experience, not decrease. Since experience tends to increase with time, this should offset the normal decay process of forgetting.

However, this study shows that it did not. This was consistent with pilot anecdotes that FAA test questions often seemed, to them, "trick questions," or otherwise based on tasks that pilots rarely do and conditions rarely encountered.

This suggests ways to improve FAA exams: (1) Screen existing questions for real-world relevancy, eliminating those based solely on rote learning; (2) Scale the relative number of weather-test items to the relative hazard and/or encounter frequency of real-world weather types (dangerous and common weather types deserve relatively more test questions); (3) Computerize the testing procedure; (4) Require pilots to pass a certain percentage of weather questions.

Critics may argue that the relatively small percentage of weather-related questions on any given exam could make the test hard to pass for some individuals due to sampling error. However, this could be addressed by *computerized adaptive testing*, which presents harder questions to a candidate after correct answers, and easier questions after incorrect answers. Computerized adaptive testing quickly homes in on a candidate's native ability level and self-terminates after reaching a preset reliability (e.g., 95%). Because adaptive tests tend to be more efficient (shorter and more reliable) than fixed-item tests, this would free up more testing time for weather-related items, making sampling error much less of a problem.

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	16	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

CONTENTS

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 3

INTRODUCTION

This constitutes the final report of a three-part series. The Phase 1 and 2 studies are similarly named (Knecht, Ball, & Lenz, 2010). The overall purpose of this research was to investigate three major questions regarding general aviation (GA) pilots:

1) Do video weather training products significantly affect GA pilot weather knowledge and flight behavior into instrument meteorological conditions (IMC)?
2) How are modern Web-based weather products used during preflight briefing?
3) Do local Oklahoma pilots differ appreciably from non-local pilots in either weather knowledge or weather-related flight behavior?

Brief Review of Phases 1 and 2

In Phase 1 of this project, 50 GA pilots were tested for pilot weather knowledge and flight behavior. Pilots took a general weather knowledge *pre-test*, followed by exposure either to one of two 90-minute 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 visual flight-rules mission through marginal weather from Amarillo, TX (AMA) to Albuquerque, NM (ABQ). Specifically, the scenario was constructed to "squeeze" pilots between slowly rising terrain and the cloud bases. Numerical flight data were collected and flight behaviors noted.

In Phase 2, after a time lapse of 3-4 months, pilots returned for a weather knowledge *follow-up test* plus a second simulator flight under similar conditions. Phase 2 was designed to test persistence-of-effect for any gains the initial weather training products may have induced during Phase 1.

Findings

Readers are directed to the Phase 1 and 2 technical reports for a detailed discussion of research Questions 2 and 3. To very briefly recap research Question 1, the weather training products seemed to produce no statistically significant increase in weather *knowledge*. However, while they did initially seem to exert a significant effect on *flight behavior*, this behavioral effect did not persist into Phase 2. The easiest explanation is that weather is a complex subject, and 90 min of training is simply too little to produce much measurable effect, let alone a long-lasting one.

Two unexpected results of the Phase 1 and 2 studies remained that were not discussed in those reports. These form the subject of the current, final report.

1) Compared to national averages, our pilots seemed to show lower-than-expected scores on their weather knowledge exam questions.
2) A number of pilots ascended into IMC and remained there for a significant amount of time, despite instructions that this was to be a visual flight rules-only (VFR) flight.

PHASES 1 AND 2 RESULTS REVISITED

Unexpected Result 1—Decrease In GA Pilots' Weather Knowledge Over Time

Does the average GA pilot's weather knowledge increase or decrease over time? There are logical arguments on both sides of this important question. Pilots study hard to pass their certification exams. Hence, we might expect them to be "at the top of their game" at test time and to subsequently forget material as time passes. On the other hand, they gain real-life experience as time passes. So, weather knowledge might actually increase as years go by.

One way to test our pilots for such change would have been to obtain their original certification tests, re-administer the weather items, and then directly analyze score changes, factoring in the number of years that had passed for each pilot. However elegant this approach, practicality and privacy made it impossible.

We therefore took a less elegant, but more practical, approach to seeing if pilots' weather knowledge had increased or decreased over time. We administered a knowledge test comprised of FAA test questions and then compared our pilots' current test scores to national averages to look for differences.

This is a non-standard approach that requires some justification. First, our test would certainly not be an exact duplicate of any FAA test ever administered. Therefore, we set up our analysis up to compare

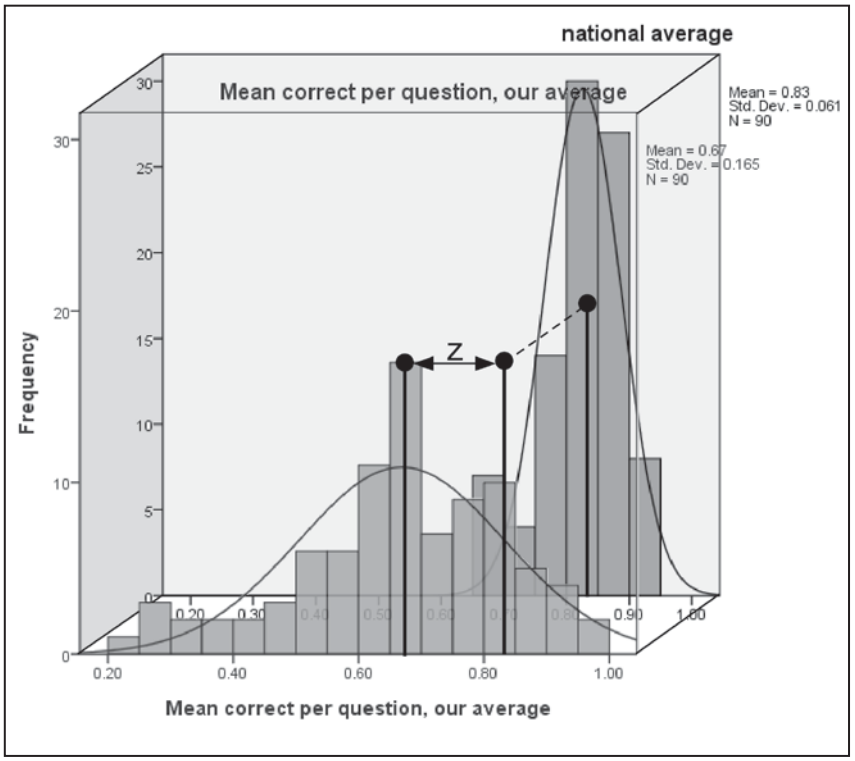

Figure 1. Frequency distributions for the 90 test items used in our study. The x-axis represents proportion-correct bins; the y-axis represents frequency counts (numbers of items falling with the x-range of each bin).

proportion-correct data for the 90 *specific questions we administered*. Each question had its own known item difficulty.[1] This resulted in two separate frequency distributions of items answered correctly (Figure 1)—one theoretical distribution (based on the national data) and one empirical distribution (based on our data). These two distributions could then be legitimately compared statistically because they involved the same test items.

A second potential issue involved annual fluctuations in score values for those individual test items. Lack of stability-over-time could introduce error into our analysis. Fortunately, the known average scores on individual items showed remarkable consistency from year to year.

A third issue involved statistical power. We had administered three alternate forms of the test. To increase statistical power, we aggregated our data. In other words, each of the 90 test items' empirical proportion-correct score was based on the average of all appropriate pilots in our experiment.

Finally, that term "appropriate pilot" took on a special meaning here. The overall experiment had involved both private[2] and instrument-rated pilots. To avoid having to construct separate tests for private and instrument-rated pilots, we instead used the same 90 items to measure both private pilots and instrument-rated pilots. This was done by a) incorporating one-third private pilot questions and two-thirds instrument-rating questions into each alternate form, and b) primarily examining only change scores.[3]

Nonetheless, in order to fairly compare our pilots with national averages for the analysis we are now discussing, the data had to be culled to leave *only private-pilot questions answered by private pilots and instrument-rating questions answered by instrument-rated pilots*. Therefore, Figure 1 reflects that culling process.

Figure 1's x-axis bins now represent two frequency distributions. The broader, flatter "our average" distribution with mean .67 represents our pilots' test scores. The more peaked "national average" distribution with mean .83 represents FAA national scores for the same items.

[1]Data were provided by AFS-630, the group that develops the test questions. These national averages are running averages based on large samples, which enhances statistical stability. For instance, in 2008, more than 15,000 pilots took the certification exam for instrument rating, while more than 28,000 took the private pilot exam (FAA, 2008).

[2]The term "private pilot" refers to the FAA rating standing above sport pilot and below instrument-rated pilot. Private pilots are not licensed to fly into areas of known adverse weather and are, instead, trained to recognize adverse weather and avoid it.

[3]Mixing questions in this way presented no particular statistical problem because the subsequent statistical analyses were primarily either correlations or change score comparisons. Both methods compared each pilot only to his or her own scores over time. So, as long as all three test versions were reasonably equivalent in expected difficulty, contained the same *proportions* of private versus instrument-rated questions, and test administration order was counterbalanced, correlations would test for *uniformity* over time, whereas change score tests would (obviously) test for *changes* over time.

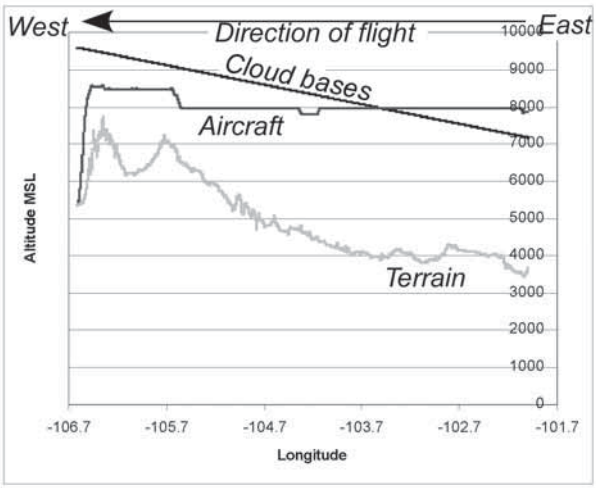

Figure 2. Typical altitude profile of a significant IMC penetration. The pilot immediately climbed into IMC and maintained level flight for most of the flight thereafter.

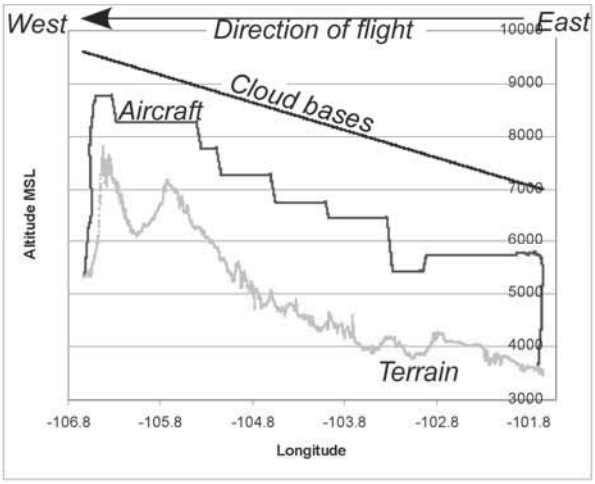

Figure 3. Typical altitude profile, showing maintenance of both ground and cloud clearance.

And, the distance between the means of these distributions represents an "estimate of forgetting" between national averages and our experimental data.

This estimate of forgetting can be expressed as a raw "percentage of forgetting," 100*(.83-.67)/.83 = 19.1%. Alternately, it can be expressed as a parametric *z*-score,[4] which can be tested for significance. Using a conservative estimate of the standard error of the mean (SEM=.0173) derived from our pilots' data,[5] $z \approx (.83-.67)/.0173 = 9.1$, $p_z < .00000000001$. As a cross-check, a far more conservative nonparametric, median-based Mann-Whitney U-test can be applied, yielding $p_U < .001$. By either standard, this shows that as a group, our pilots answered these particular questions significantly less correctly than did the national sample. The issue then becomes how to interpret that finding. We will return to this later in the Discussion section.

Unexpected Result 2—Long-duration penetration into IMC

Background

Despite this being stressed to pilots as a VFR-only flight, 9 of 50 (18%) in Phase 1 ascended immediately into IMC and spent at least 10 min there. Figure 2 shows a typical IMC penetration profile.

Ruling out the trivial

We first needed to rule out the simplest possible explanation for this behavior, namely some kind of problem with the way the simulator depicted clouds. Because, if the cloud bases were too hard to detect, pilots could have unintentionally ascended into them without realizing it.

Several factors argued against that trivial explanation. First was the fact that we had flown the simulator ourselves many times and had not found it too difficult to avoid cloud bases. There are obvious clues when one goes into IMC—severely restricted visibility both straight ahead and toward the ground. When the preflight briefing tells you there are clouds and you climb and see gray skies overhead and suddenly lose sight of the ground and cannot see in front of you—you are unmistakably in IMC. Moreover, 82% of Phase 1 and 93% of Phase 2 pilots were able to avoid the clouds completely, which is unlikely if clouds were difficult to perceive. Figure 3 shows a typical pilot's altitude profile, with step-climbs that clearly denote awareness of both terrain and cloud bases.

Examining other hypotheses for the current flight into IMC

One thing we could determine was whether flight-into-IMC decreased from Phase 1 to Phase 2. A statistically significant decrease might suggest that some pilots had realized and corrected their original error. Table 1 presents the data.

In Phase 1, 9 of 50 pilots (18%) were classified as "long-term IMC penetrators" spending more than 10 min in IMC. In Phase 2, this decreased to 3 of 44 (7%).[6] This Phase 1 → 2 decrease was, indeed, nominally significant

[4] *z = raw distance/standard error of the mean.*

[5] Our pilot sample SEM (.0173) was based on a pooled estimate of score variance (Ostle & Mensing, 1975). It was predictably larger (thus, more conservative) than that based on the national sample (.0064) because our sample size was smaller.

[6] In the remaining discussion, "penetrator" will mean "long-duration IMC penetrator," defined as 10 min or more in IMC.

Table 1. Features of long-term IMC penetrators.

		Instrument-rated?	Min in IMC		Returned questionnaire?	Plausible explanation found for this pilot
Pilot	Wx Trg Prod		Phase 1	Phase 2		
1	Trg Prod 1	Y	15.63	0	Y	N
2	Trg Prod 1	N	31.76	52.19	N	N
3	Trg Prod 1	N	82.61	**Absent**	N	N
4	Control	Y	41.1	**Absent**	Y	N
5	Control	Y	71.97	67.08	Y	Y
6	Control	Y	47.52	0.88	Y	Y
7	Control	Y	30.03	84.70	Y	N
8	Control	N	38.36	**Absent**	N	N
9	Control	N	23.67	0	N	N

Table 2. Numbers of pilots showing long-duration IMC penetration in Phase 1—*actual* (in **bold**) v. *expected* (in parentheses).

		Trg Prod 1	Trg Prod 2	Control
Ph1 IMC penetration > 10 min	Yes	**3** (2.9)	**0** (2.9)	**6** (3.2)
	No	**13** (13.1)	**16** (13.1)	**12** (14.8)
		.23		.02
		.45		

(1-tailed $p_{odds\ ratio}$ = .049). However, note that much of that decrease was spurious, being due to three Phase 1 penetrators subsequently dropping out of the study (Table 1, pilots 3, 4, and 8 labeled "Absent"). That fact nullifies any claim of an actual meaningful decrease because we cannot know how much time those absent pilots would have spent in Phase-2 IMC, had they returned.

All this does arouse some suspicion, of course. Perhaps the very reason pilots 3, 4, and 8 failed to return was *because* they had realized their mistake and that provoked a strong emotional reaction.

But, this is speculation. To try to test that speculation, we can first estimate the probability (as actually happened) that exactly three of the Phase 1 penetrators would go on to become *repeaters* in Phase 2 (Table 1's gray highlighted rows, pilots 2,5,7). The analysis two paragraphs above did not calculate the chance of *specific pilots* repeating their behavior—only of there being *any* nine penetrators in Phase 1 versus *any* three in Phase 2. Fine-tuning the analysis to calculate the chance of having exactly three repeaters gives us a cleaner way to assess the Phase 1→2 decrease because dropouts no longer matter.

Unfortunately, that result is inconclusive ($p_{binomial}$ = .117, NS).[7] Nonetheless, one other suspicious fact remains. Six Phase 1 pilots never returned for Phase 2. Suspiciously, half of those six were also Phase 1 penetrators. If that proved to be statistically unlikely, it could support a hypothesis that these three "dropout penetrators" had left for a common reason—and a very likely reason would be that their time in IMC had provoked some kind of emotional reaction, as we mentioned earlier.

Unfortunately, that result is also inconclusive ($p_{binomial}$ = .071, NS),[8] although it does trend[9] in the predicted direction.

Other patterns in these data were similarly frustrating to assess. For instance, of the six Phase 1 penetrators who did return for Phase 2, 3 subsequently spent less actual time in Phase 2 IMC. Yet, two spent *more* time in Phase 2 IMC, rendering that analysis inconclusive as well.

Instrument rating, pilot age and flight hours

We might suspect that instrument rating, age, and flight hours would exert some effect on IMC penetration. However, in examining instrument rating, five of the nine pilots were instrument-rated; four were not. Figure 4 shows graphically that instrument-rated pilots' altitude profiles do not look radically different from those of non-instrument-rated pilots.

In examining the 50 Phase 1 pilots, neither age nor flight hours correlates closely with time spent in Phase 1 IMC [2-tailed $r_{Spearman}$ = -.09 (age), -.08 (flight h), NS]. Neither did the nine Phase 1 penetrators differ from the remaining 41 pilots on either variable [2-tailed $p_{Mann\text{-}Whitney\ U}$ = .487 (age), .553 (flight h), NS].

Weather training product

Did the weather training product affect long-term IMC penetration? Table 2 addresses this.

Training Product 2 seemed to exert an effect on long-term penetration (2-tailed $p_{X^2 exact}$ = .036).[10] Training

[7] See Appendix B for derivation.

[8] See Appendix B for derivation.

[9] In this study, "significance" will be defined as probability of occurrence (p) <.05, whereas "trend" will mean .05<p<.10.

[10] Fisher's Exact Test is used here because 3 of 6 expected cell frequencies are < 5.

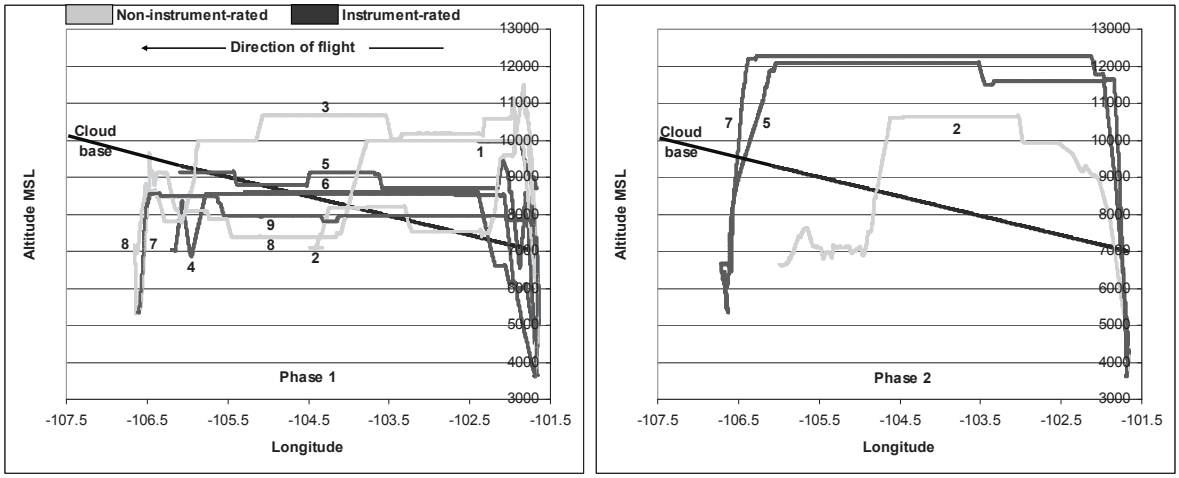

Figure 4. Altitude profiles of all IMC penetrations > 10 min. These also show the aforementioned tendency of most penetrators to immediately ascend into the clouds and stay there in relatively level flight.

Product 2 pilots had significantly fewer long-duration penetrations than Controls (0 v. 16, 2-tailed $p_{X^2 exact}$ = .02). The other two pairwise contrasts were nonsignificant (2-tailed $p_{X^2 exact}$ = .23, .45).

Why would Training Product 2 stand out? Perhaps because it particularly emphasized visual recognition of hazardous weather types.[11] Given how difficult visual hazard recognition can be, immediate exposure to video training may have engendered greater caution during the subsequent simulator flight.

Questionnaire

To probe the possible reasons for this long-duration IMC penetration, a brief follow-up questionnaire was developed (see Appendix A) and sent to the nine penetrators along with an addressed business-reply envelope.[12] This questionnaire was designed to avoid judgmental or pejorative overtones and could be filled out in just a few minutes. The intent was to explore simple, plausible reasons why a pilot may have strayed from the instructions, for instance:

- Being distracted due to the difficulty of trying to control the flight simulator
- Being distracted due to extreme anxiety
- Being distracted due to feeling out of control
- Simply misunderstanding the instructions

A free-response item was also included to give pilots the opportunity to state whether there was anything they might have done differently in Phase 1, if given the chance. For example, a pilot could state he was trying to break out above the clouds, or say that this was just a simulation and not to be taken seriously. Any or all such factors might explain why a person might ascend into virtual clouds in a flight simulator, without implying conscious, willful rule violation during real flights.

Given the good chance that most pilots would forget at least part of their flight profiles, each questionnaire was accompanied by a picture of that pilot's exact Phase 1 flight profile to refresh their memories. Drawn using *Mathematica* (Wolfram Research, 2008), it depicted U.S. Geological Survey terrain data (NGDC, 2008) with that individual pilot's 4D flight profile overlaid (latitude/longitude/altitude at time *t*). Figure 5 shows a representative 50%-scale image.[13] The cloud base appears as a translucent blue "glass sheet." Droplines from the flight track convey both ground track and altitude information. Yellow droplines indicate flight below the cloud base (visual meteorological conditions, VMC). Red droplines indicate flight into IMC.

Questionnaire data

Only five of the nine (56%) questionnaires were returned. For non-respondents, a total of four reminders were made via e-mail, regular mail, and/or telephone over a 4-month period, including a second mailing of the questionnaire and flight profile.

Table 3 shows frequency counts for the five returned questionnaires. Paired questions compared whether pilots

[11]Product 2 showed pilots still pictures of marginal weather and asked them if such weather fit the definition of VMC. This turned out to be a difficult and rather humbling task.

[12] By way of caution, we need to be conservative in making categorical assertions on the basis of questionnaires (such as our own) that have not been subjected to test-retest reliability assessment. Most authors of such instruments routinely fail to state this caveat. In test construction, *reliability* captures the degree to which subjects re-taking the test would give the same answers they did the previous time. Like correlation, reliability can theoretically go from -1 (retest scores are opposite of the original test; all items answered correctly the first time are now answered incorrectly and vice versa) to +1 (retest is identical).

[13]These figures were quite distinctive-looking. Hence, for the sake of confidentiality, Figure 5 is not a profile of any of the nine pilots currently under discussion.

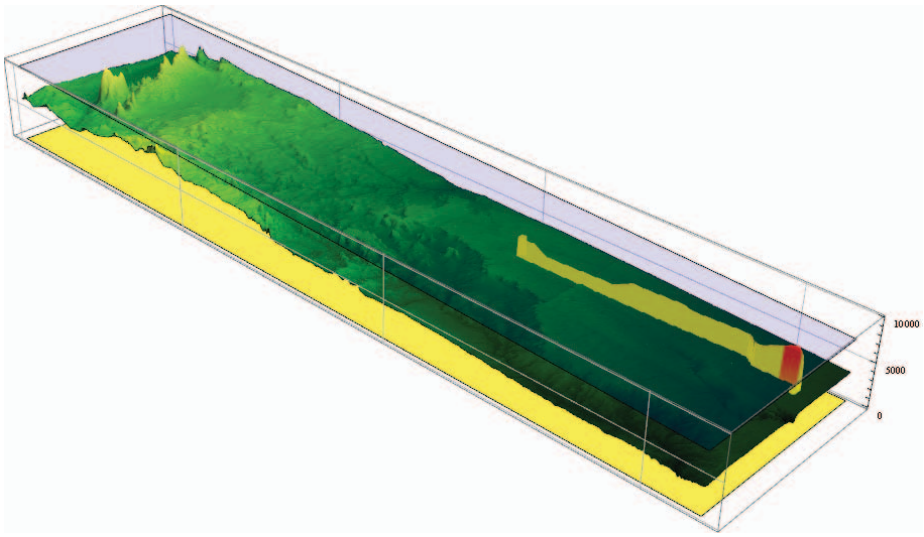

Figure 5. Sample 4D flight profile for a flight from AMA (lower right) to ABQ (upper left). Here, the pilot briefly ascended into IMC right after takeoff (red droplines). He quickly recognized this and descended back into VMC (yellow droplines), continuing in VMC until diverting about 2/5 of the way to ABQ.

Table 3. Frequency counts from pilot questionnaires (5 of 9 total possible returned) for pilots who spent ≥ 10 min in IMC during Phase 1.

		Not at all	Just slightly	Slightly below average for this type of flight	About average for this type of flight	Slightly above average for this type of flight	Very much so	Extremely so
1	How challenging... **during the first 10-15 minutes?**				2	1	2	
2	How challenging... **during the next 10-15 minutes?**		2		1	1	1	
3	anxiety level ...**during the first 10-15 minutes.**		1		1	2	1	
4	anxiety level... **during the next 10-15 minutes.**	1	1	1		2		
5	confident, focused, in command **during the first 10-15 minutes.**		1	3	1			
6	confident, focused, in command **during the next 10-15 minutes.**		1	1	3			
7	understanding the flight mission **during the first 10-15 minutes.**		2	1	2			
8	understanding the flight mission **during the next 10-15 minutes.**			2	2	1		
	Change-of-affect from the 1st 10-15 min to the 2nd 10-15 min of flight	-3	-2	-1	0	1	2	3
2-1	How challenging... **next 10-15 minutes – first 10-15 minutes**		2	2	1			
4-3	anxiety level..... **next 10-15 minutes – first 10-15 minutes**	1		3	1			
6-5	confident..... **next 10-15 minutes – first 10-15 minutes**				3	2		
8-7	understanding..... **next 10-15 minutes – first 10-15 minutes**				1	4		

felt better/the same/worse between the *first* 10-15 min of the flight versus the *second* 10-15 min regarding:

- how challenging the simulator was to fly
- their overall anxiety level
- their overall feeling of confidence
- their understanding of the flight mission

The intent was to see whether each pilot felt increasingly stressed versus increasingly relaxed as time went by.

In Table 3, four rows up from the bottom (white text on gray background), the term "change-of-affect" refers to how pilots reported their emotional situation *changed* between the first 10-15 min and the second 10-15 min. For example, if an individual pilot felt challenged "about average" on question #1 and "slightly below average" on question #2, then the change-of-affect score would be "-1" for that pair of questions, because he or she felt "1 unit less-challenged" during the second 10-15 min compared to the first 10-15. This was how we estimated changes in emotional stress with time.

As previously shown in Figure 4, the first 10-15 min represented time when these nine pilots were flying in the clouds. Table 2 shows (in the gray highlighted cells) that during that time, two pilots found the flight "very much" challenging, one felt "very much" anxious, one felt "just slightly" confident, and two felt they understood the flight mission "just slightly."

Superficially, that looks impressive. However, looks are deceiving. A single instrument-rated pilot (Table 1, Pilot 5) contributed over half that most-extreme data, reporting extreme scores in all four categories. Therefore, only that one pilot showed a potentially simple explanation for going IMC (i.e., it may have happened inadvertently while he was feeling overwhelmed).

Subtracting out that one pilot leaves little information about the remaining four. According to their own reports, those four felt that they mainly understood the mission, felt relatively confident, and not terribly challenged nor anxious.

Turning to the verbal explanations,[14] further explanation was found for only one additional pilot (Pilot 6). This instrument-rated pilot (different from the one mentioned two paragraphs above) wrote about being aware of the clouds and wanting to stay beneath them but felt that clearing terrain was more important than avoiding clouds. Of all explanations given, that was arguably the most straightforward.

So, contrary to expectations, the change-of-affect scores in Table 3 indicates that most pilots reported feeling *less* challenged, *less* anxious, *more* confident, and understood the mission *better* as time went by. For Pilot 9, that just about corresponds to the time of breaking out of the clouds, since level flight + a steadily rising cloud base = eventual spontaneous breakout. But, as Figure 4 showed earlier, the remaining pilots were *still in the clouds during the second 10-15 minutes.*[15]

DISCUSSION

Despite specific instructions to fly VFR in both phases of this study, nine of 50 Phase 1 pilots spent more than 10 min in IMC, while three of 44 did so in Phase 2. This "long-term IMC penetration" was not supposed to happen. Moreover, all three Phase 2 long-duration IMC "penetrators" had also been penetrators in Phase 1, which was unlikely, although not technically significant (p = .117).

One major question was, "Were any of these violations willful?" To find out, we sent a questionnaire to these nine pilots (Appendix A). Five returned the questionnaire.

We hypothesized at least four charitable reasons that pilots might have behaved against instructions:

- Being distracted due to the difficulty of trying to control the flight simulator
- Being distracted due to extreme anxiety
- Being distracted due to feeling out of control
- Simply misunderstanding the instructions

After analysis, none of these hypotheses appeared common to a significant majority of respondents. Individual questionnaires revealed only that a few pilots responded consistently with a few of the hypotheses; in particular, Pilot #5 reported all four. In the free-response section, Pilot #6 reported being more concerned with terrain avoidance than cloud clearance. Yet, no clear, simple, reliable explanation emerged common to the majority of the respondents.

A small number of technically nonsignificant-but-suggestively-close results were seen:

- Phase 1's *ratio of penetrators* (9/50) suspiciously decreased to 3/44 in Phase 2 (1-tailed $p_{odds\text{-}ratio}$ = .049).
  - » However, this result was contaminated because three Phase 1 penetrators dropped out of the study before Phase 2. Had this Phase 1→2 decrease been significant, it might have implied that Phase 1 long-term penetrations were merely mistakes that most pilots learned to correct.
- Six Phase 1 pilots failed to return for Phase 2. Of these six, as just stated, suspiciously, half were also Phase 1 long-term penetrators ($p_{binomial}$ = .071).

[14]These are also omitted for reasons of privacy.

[15]This was about a 90-min flight. So, the second 10-15 min would be the time from 10-15 min (start) to 20-30 min (finish)—less than 1/3 of the way to the destination.

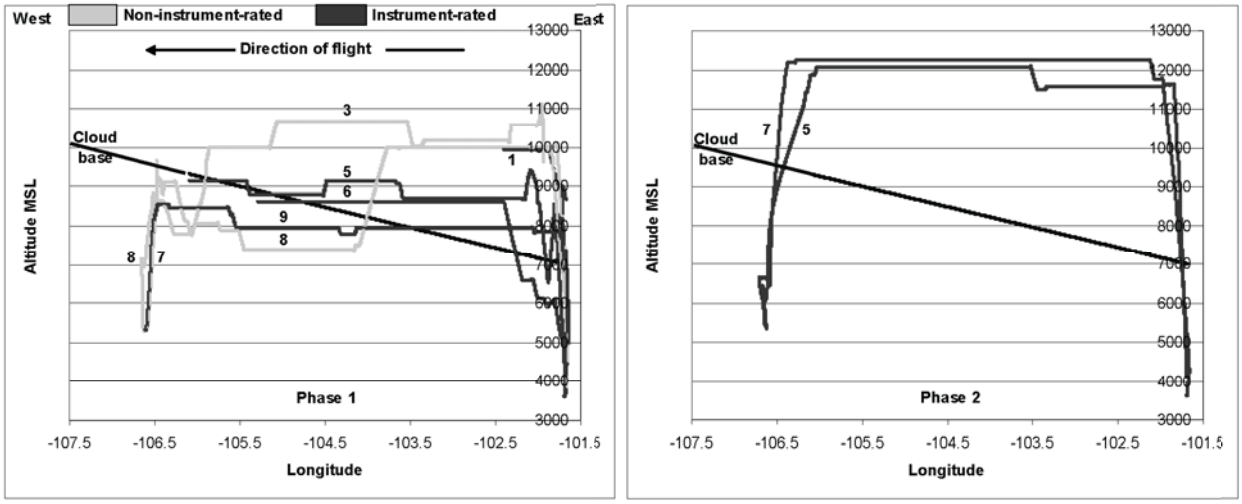

Figure 6. Altitude profiles of Figure 5, with 2 pilots removed who diverted before reaching the destination (ABQ).

» Had this been significant, it might have implied that dropouts recognized their error and became deeply emotional over it.

These last two results are consistent with something being afoot—perhaps some pilots being aware they had done something wrong and either self-correcting or leaving the study—but the sample sizes were simply too small to say with great confidence.

So, what probably really happened? From a purely logical perspective, the front-running explanation for this long-term IMC penetration may be Pilot 6's comment about being more worried about terrain avoidance than cloud clearance. Refer now to Figure 6.

Figure 6 shows the altitude profiles of all pilots except for #2 and #4 (who diverted before reaching ABQ). Keep in mind that this is only supposition, but *in every case, these profiles are consistent with preflight terrain avoidance planning (TAP).* Note how all pilots initially climbed to an altitude nearly as great or greater than the highest peak over which they needed to fly (about 8700 ft). In Phase 1, Pilot 9 climbed to "only" 8000 ft, but even that exceeded the height of the mountain passes east of ABQ and only needed an additional 700 ft final boost to clear the peaks themselves. Note that Pilots 3 and 8 eventually descended out of Phase 1 IMC, but they *started out* behaving like TAP. The point is that—logically speaking—each one of these pilots seemed to be executing TAP, at least initially. Sadly, what appears evident by visual inspection turns out to be difficult to support statistically.

Implications of This Research for Pilot Training

The original purpose of Phases 1 and 2 was to test the effect of video weather training products on weather-related risk-taking. That was the primary point of the entire experiment. However, little effect of those training products was seen prior to the current analysis.

The current analysis suggests that *certain selected kinds of potentially hazardous weather situations may be learned by video instruction*. Specifically, training product 2 concentrated on visual perception of hazardous weather. In Table 2 we saw that pilots who saw training product 2 subsequently showed *no long-duration IMC penetration* during either phase of the experiment.

A compelling logic supports this kind of instruction, namely that pilots *have to experience adverse weather somehow*, despite its risk. Perhaps perceptual-based video instruction can fill that need. VFR flight instructors purposely exclude bad-weather encounters because they are too risky. Even IFR training typically avoids known extreme weather hazards. Still, pilots have to learn to respect bad weather somehow. Why leave that learning process *ab tempestas*—to "encounter with storm"—when technology exists to teach pilots what they need to know before a real-life encounter, rather than after?

The one caveat concerning video instruction is simply that *learning good weather-flying practices takes time*. The current study shows that one 90-minute training video may change pilot behavior to a small degree, but that profound, long-lasting change requires a bigger investment. Just as we cannot build an entire house from a single brick—no matter how well-crafted the brick—there is much to learn about weather, and we tend to forget over time. Therefore, learning has to be initially sufficient *and* ongoing in order to become and remain effective.

Another implication for pilot training involves preflight and in-flight weather briefing. During the original experiment, many pilots informally expressed opinions that *the future of weather briefing looks increasingly Internet-based*, as opposed to coming solely from the Flight Service

Station (FSS). Both types of briefing will require training as self-briefing increasingly augments the FSS and GA aircraft begin to acquire lower-cost equipment capable of receiving and displaying integrated weather information.

Implications of Weather Knowledge Forgetting for Pilot Testing

The current study also has a number of implications for pilot testing. Most broadly, it shows that these pilots' weather knowledge test scores revealed a significant decline (19%) compared to average FAA certification exam scores obtained by freshly licensed pilots. This is a statistically significant finding ($p<.001$).

This is equally significant from a logical standpoint. Since knowledge retention tends to be a function of knowledge relevancy. If FAA test questions were uniformly highly relevant to real-world weather encounters, we would expect pilots' scores to increase with experience, not decrease. And, since experience tends to increase with time, this should offset the normal decay process of forgetting.

However, this study shows that it did not. This finding is consistent with common anecdotal complaints of these pilots, namely that FAA test questions often seemed, in their words, "trick questions," or ones otherwise based on tasks that pilots rarely do and conditions they rarely encounter.[16]

This suggests ways to improve the FAA exams. The most obvious response is to *screen existing questions for real-world relevancy, eliminating those based solely on rote learning*.

A second suggestion is to *gear the relative number of weather-test items to the relative hazard and encounter frequency of different weather types*. Dangerous, common weather types deserve relatively more questions. This would "factor-weight" the tests to reflect the actual danger and chances of encountering each specific real-world weather threat.

A third suggestion is to *computerize the licensure-testing procedure* to allow questions concerning topics heretofore untestable. Because the future looks Web-based and graphical (as opposed to text-based), FAA testing can now adapt to address these new and powerful technologies. For instance, a computer-based test would allow display of videos of actual weather situations such as marginal visibility, followed by technical and practical questions relevant to those situations. Such a test could also address Web-based preflight weather briefing, giving applicants a chance to demonstrate actual hands-on proficiency using such tools to find desired kinds of information.

A fourth possible improvement was suggested by Wiegmann et al. (2008) after they noted that pilots could pass their certification exams despite failing all the weather items. Pilots could be required to pass a certain percentage of the weather questions, as well.

Of course, critics can counter-argue that the relatively small percentage of weather-related questions on any given exam could make the test hard to pass for some individuals due to sampling error.[17] However, this sampling error itself could be addressed by *computerized adaptive testing* (CAT, van der Linden & Glas, 2000). CAT presents harder questions to a candidate after correct answers, and easier questions after incorrect answers. CAT typically quickly homes in on a candidate's ability level and self-terminates after reaching a preset reliability level (e.g., 95%).[18] Because adaptive tests tend to be more efficient (shorter and statistically reliable) than fixed-item tests, this would free up more testing time for weather-related items, making sampling error much less of a problem.

[16]This should cast no aspersions on AFS-630, which inherited a test bank containing a substantial number of items written before current staff's time. Replacing old, known items with new ones of unknown performance is extremely time-consuming.

[17]Sampling error occurs when the number of test items is small compared to the total body of knowledge. Since each test is only a sample of *what could be tested*, a few unlucky individuals, who may otherwise know quite a lot, may face a test consisting—purely by chance—mainly of things they do not know.

[18]A reliability level (a.k.a. *confidence level*) of 95% means that, given an infinite number of tests of equivalent difficulty, 95% of the time the candidate should test within ±1.96 standard error of the mean from their theoretical "true native ability score."

REFERENCES

Federal Aviation Administration (2008). Downloaded June 1, 2009 from http://www.faa.gov/data_research/aviation_data_statistics/test_statistics/index.cfm.

Knecht, W.R., Ball, J., & Lenz, M. (2010). *Effects of video weather training products, Web-based preflight weather briefing, and local versus non-local pilots on G.A. pilot weather knowledge and flight behavior, Phase 1.* (Technical Report DOT/FAA/AM-10/1). Washington, DC: Federal Aviation Administration.

Knecht, W.R., Ball, J., & Lenz, M. (2010). *Effects of video weather training products, Web-based preflight weather briefing, and local versus non-local pilots on G.A. pilot weather knowledge and flight behavior, Phase 2.* (Technical Report DOT/FAA/AM-10/6). Washington, DC: Federal Aviation Administration.

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

Ostle, B, & Mensing, R. W. (1975). Statistics in research (3rd ed.). Ames: Iowa State University Press. p. 97.

van der Linden, W.J., & Glas, C.A.W. (2000). *Computerized adaptive testing: Theory and practice.* Boston: Kluwer.

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

Wiegmann, D.A., Talleur, D.A., & Johnson, C.M. (2008). Redesigning weather-related training and testing of general aviation pilots: Applying traditional curriculum evaluation and advanced simulation-based methods. (Technical Report FAA-08-1). Retrieved November 19, 2008 from http://www.humanfactors.uiuc.edu/ Reports&PapersPDFs/TechReport/08-01.pdf

APPENDIX A
Questions Asked of Pilots Who Spent More Than 10 Min in IMC

As you'll see below, these questions all compare your thoughts and emotions during the very first part of the flight, as compared to a little while later (after you'd had a chance to get a little more used to the Malibu). This tells us something about the Malibu's learning curve and will help us plan future experiments, so we appreciate your expertise here.		Not at all	Just slightly	Slightly below average for this type of flight	About average for this type of flight	Slightly above average for this type of flight	Very much so	Extremely so
1	How challenging was the CAMI flight simulator to fly **during the first 10-15 minutes?**							
2	How challenging was it **during the next 10-15 minutes after that?**							
3	Describe your overall anxiety level in the simulator **during the first 10-15 minutes.**							
4	Describe your overall anxiety level **during the next 10-15 minutes after that.**							
5	Describe your overall feeling of being confident, focused, and in command **during the first 10-15 minutes.**							
6	Describe your overall feeling of being confident, focused and in command **during the next 10-15 minutes after that.**							
7	Describe your overall feeling of understanding the flight mission **during the first 10-15 minutes.**							
8	Describe your overall feeling of understanding the flight mission **during the next 10-15 minutes after that.**							
9	Looking back, is there anything you would have done differently in the **Phase 1** experiment? If so, briefly describe it.							

APPENDIX B

Regarding footnote 7: Suspiciously, three of the nine Phase 1 long-term penetrators were also Phase 2 penetrators. Intuitively, that seems unlikely, so we want to find the probability of this happening by chance. The problem decomposes like so:

Number of long-term Phase 1 IMC penetrators	9
Number of Phase 1 pilots	50
Each Phase 1 pilot's chance of being a long-term IMC penetrator	9/50
(9/50 now becomes our estimate of the base rate for any single pilot's "propensity to penetrate")	
Each Phase 1 pilot's chance of being both a Ph 1 AND a Ph 2 penetrator	(9/50)*(9/50) = 81/2500
Chance of exactly k Phase 1 pilots also being penetrators in Phase 2	$p = \frac{n!}{k!(n-k)!}\left(a^k b^{n-k}\right)$

where n=44 (because there were only 44 Ph 2 pilots), and k=3 (describing the (n - k+*1*)th—here, the 42nd—coefficient c_{n-k+1} in Pascal's triangle = $13244a^3\ b^{41}$, representing the expansion of the binomial $(a+b)^{44}$, where a=81/2500=the chance of any single pilot being a "repeater", and b=1-(81/2500)=the chance of that pilot *not* being a repeater.

In other words, we solve for $p=c_{42}(a^3b^{41})$, where a^3b^{41} represents the unique condition where 3 pilots are repeaters and 41 are not, and c_{42} is the coefficient representing the relative number of times we would expect to see 3 repeaters, given an infinite number of Bernoulli trials.

Solving, $p = \frac{44!}{3!(44-3)!}\left(a^3b^{41}\right)$

$$= \frac{44*43*42*41!}{(3*2)*41!}\left((81/2500)^3(1-(81/2500))^{41}\right)$$

$$= .117$$

Regarding footnote 8: There were nine long-term IMC penetrators out of 50 pilots in Phase 1. In Phase 2, there were three penetrators out of 44 pilots. This means that six pilots dropped out of the study. Suspiciously, of these six dropouts, three (half) had also been penetrators in Phase 1. We want to find the probability of this happening by chance. The problem decomposes like so:

Number of pilots failing to return for Phase 2	6
Number of pilots in Phase 1	50
Each Phase 1 pilot's chance of failing to return for Phase 2	6/50
Number of long-term Phase 1 IMC penetrators	9
Each Phase 1 pilot's chance of being a long-term IMC penetrator	9/50
Each Phase 1 pilot's chance of being a penetrator AND failing to return	(9/50)*(6/50) = 54/2500
Chance of exactly k Phase 1 pilots being penetrators AND failing to return	$p = \frac{n!}{k!(n-k)!}\left(a^k b^{n-k}\right)$

where n=50, k=3 (describing the (n – k +*1*)th—here, the 48th—coefficient c_{n-k+1} in Pascal's triangle, representing the expansion of the binomial $(a+b)^{50}$, where a=54/2500, b=1-(54/2500).

Here, we solve for $p=c_{48}(a^3b^{47})$.

Solving, $p = \frac{50!}{3!(50-3)!}\left(a^3b^{47}\right)$

$$= \frac{50*49*48*47!}{(3*2)*47!}\left((54/2500)^3(1-(54/2500))^{47}\right)$$

$$= .071.$$