Federal Aviation
Administration

DOT/FAA/AM-19/1
Office of Aerospace Medicine
Washington, DC 20591

Mobile Meteorological Information Tailored to Landing Phase of Flight. Part II: Refinement

William R. Knecht[1]
Padhrig McCarthy[2]

[1]FAA Civil Aerospace Medical Institute
Oklahoma City, OK 73125

[2]Research Applications Laboratory (RAL)
National Center for Atmospheric Research (NCAR)

March 2018

Final Report

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents thereof.

This publication and all Office of Aerospace Medicine technical reports are available in full-text from the Civil Aerospace Medical Institute's publications website: (http://www.faa.gov/go/oamtechreports)

Technical Report Documentation Page

1. Report No. DOT/FAA/AM-19/1	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Mobile Meteorological Information Tailored to Landing Phase of Flight. Part II: Refinement		5. Report Date September 2019 6. Performing Organization Code
7. Author(s) Knecht, W[1] McCarthy, P[2]		8. Performing Organization Report No.
9. Performing Organization Name and Address [1]FAA Civil Aerospace Medical Institute P.O. Box 25082 Oklahoma City, OK 73125 [2]Research Applications Laboratory (RAL) National Center for Atmospheric Research (NCAR)		10. Work Unit No. (TRAIS) 11. Contract or Grant No.
12. Sponsoring Agency name and AddReress Office of Aerospace Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, DC 20591		13. Type of Report and Period Covered 14. Sponsoring Agency Code
15. Supplemental Notes		

16. Abstract

This study represents the fourth in a series of tests of a mobile meteorological application intended for aircraft pilots, and designed to run on a tablet computer. The current study focuses on features that the third study (Knecht & Dumont, 2017) proved useful in assessing the risk of runway winds during the landing phase of flight. Specifically, graphical depictions of runway winds were compared with textual descriptions of the same Meteorological Terminal Aviation Routine (METAR) report-like information.

Significant findings emerged. First, graphical depiction was significantly more efficient than textual depiction, taking only 70% as much cognitive processing time, with no penalty in accuracy. Moreover, when available viewing time was severely constrained to just 5 seconds, pilots timed out significantly fewer times with graphical depiction.

Second, graphical depiction produced significantly fewer misclassifications of landing difficulty than textual depiction.

Third, graphical depiction ultimately produced fewer mistakes in deciding whether or not to land.

Fourth, graphical depiction produced significantly higher pilots' confidence in their landing decisions, particularly when available viewing time was severely constrained. And, because those decisions were demonstrably better, the higher confidence appeared warranted.

Fifth, pilots appeared to employ heuristics (simplifying rules) when estimating risk due to runway winds. In textual depiction, especially when time is limited, pilots appear to ignore trigonometry and instead base their risk estimates as if the stated wind speed is the crosswind component. Similarly, in graphical depiction, pilots appear to focus on the more severe wind component as the limiting risk factor for that landing. Interestingly, both heuristics lead to overestimation of risk, and more conservative landing behavior.

Finally, pilots unanimously preferred the graphical depiction, both in this study, and the previous one. Unanimity of preference is quite rare in product development.

17. Key Words aviation weather, runway wind, crosswind, cockpit display, human factors		18. Distribution Statement Document is available to the public through the Internet: (http://www.faa.gov/go/oamtechreports/)	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 36	22. Price

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

Mobile meteorological information tailored to landing phase of flight. Part II: Refinement

William R. Knecht
Civil Aerospace Medical Institute, FAA, AAM-510

Padhrig McCarthy
Research Applications Laboratory (RAL)
National Center for Atmospheric Research (NCAR)

EXECUTIVE SUMMARY

This study represents the fourth in a series of tests of a mobile meteorological application intended for aircraft pilots, and designed to run on a tablet computer. The current study focuses on features that the third study (Knecht & Dumont, 2017) proved useful in assessing the risk of runway winds during the landing phase of flight. Specifically, graphical depictions of runway winds were compared with textual descriptions of the same Meteorological Terminal Aviation Routine (METAR) report-like information.

Significant findings emerged. First, graphical depiction was significantly more efficient than textual depiction, taking only 70% as much cognitive processing time, with no penalty in accuracy. Moreover, when available viewing time was severely constrained to just 5 seconds, pilots timed out significantly fewer times with graphical depiction.

Second, graphical depiction produced significantly fewer misclassifications of landing difficulty than textual depiction.

Third, graphical depiction ultimately produced fewer mistakes in deciding whether or not to land.

Fourth, graphical depiction produced significantly higher pilots' confidence in their landing decisions, particularly when available viewing time was severely constrained. And, because those decisions were demonstrably better, the higher confidence appeared warranted.

Fifth, pilots appeared to employ heuristics (simplifying rules) when estimating risk due to runway winds. In textual depiction, especially when time is limited, pilots appear to ignore trigonometry and instead base their risk estimates as if the stated wind speed is the crosswind component. Similarly, in graphical depiction, pilots appear to focus on the more severe wind component as the limiting risk factor for that landing. Interestingly, both heuristics lead to overestimation of risk, and more conservative landing behavior.

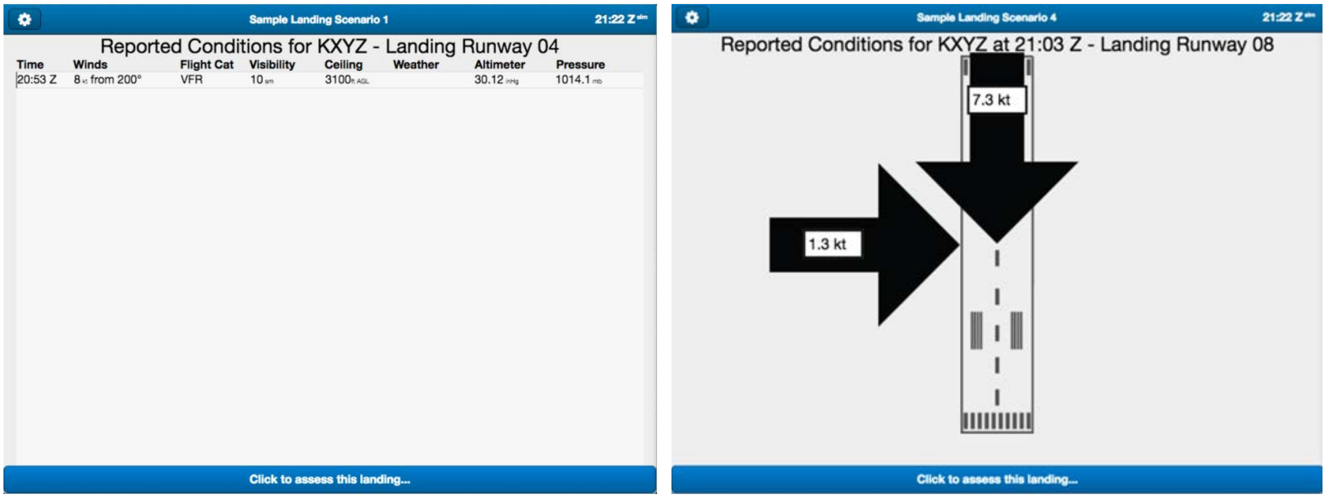
Finally, pilots unanimously preferred the graphical depiction, both in this study, and the previous one. Unanimity of preference is quite rare in product development.

INTRODUCTION

This report summarizes continued empirical testing of a low-cost, portable device designed to deliver timely weather information to the general aviation (GA) flightdeck. This device is a mobile meteorological application (MMET) that runs on a tablet computer, and is currently under development by the Research Applications Laboratory (RAL) of the National Center for Atmospheric Research (NCAR).

This MMET has so far been tested in multiple phases. Phase 1 evaluation began at the FAA's William J. Hughes Technical Center's Aviation Weather Demonstration and Evaluation (AWDE) Services branch as a scenario-based cognitive walkthrough (AWDE DOC150). Phase 2 testing was performed at the Technical Center's Human Factors Branch Cockpit Simulator Facility (Ahlstrom, Caddigan, Schulz, Ohneiser, Bastholm & Dworsky, 2015). That study focused on pilot separation from weather in the cruise phase of flight. Phase 3 testing was conducted at the FAA's Civil Aerospace Medical Center (CAMI), and focused on runway winds during the landing phase of flight (Knecht & Dumont 2017).

The current (Phase 4) study refines the Phase 3 methodology by testing an even more-advanced concept for displaying METAR information. Figure 1 illustrates the starting point—two wind-information depictions previously tested in Phase 3.



Sample Landing Scenario 1

21:22 Z sim

Reported Conditions for KXYZ - Landing Runway 04

Time	Winds	Flight Cat	Visibility	Ceiling	Weather	Altimeter	Pressure
20:53 Z	8 kt from 200°	VFR	10 sm	3100 ft AGL		30.12 inHg	1014.1 mb

Click to assess this landing...

Sample Landing Scenario 4

21:22 Z sim

Reported Conditions for KXYZ at 21:03 Z - Landing Runway 08

Click to assess this landing...

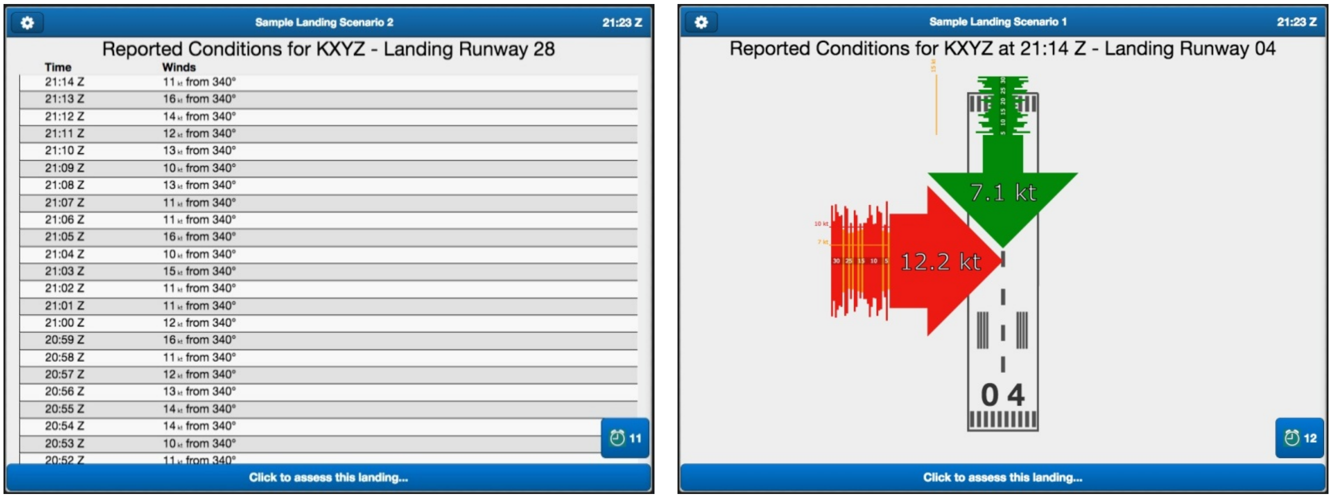
a b

Figure 1. Two wind-information depictions previously tested in Phase 3, a) the traditional textual METAR used as an experimental baseline, or control condition, b) the experimental "two-arrow" screen, showing runway-relative wind components.

In Phase 3, the two-arrow graphical runway wind depiction (Fig. 1b) greatly outperformed its textual comparator (Fig. 1a). In Figure 1b, notice how the most-current observation's ground-level wind speed is represented as separate headwind/tailwind and crosswind components. This method of depicting the winds resulted in pilots spending significantly less mean viewing time (8.9 sec, $p = 1.5\times10^{-8}$), compared to 17.4 seconds for the equivalent, traditional text-based METAR (Fig. 1b), with no measurable loss in the *quality* of landing-difficulty judgements.

This was a significant finding, one suggesting immediate, useful improvement in the way runway wind information could be displayed to pilots during landings.

The current (Phase 4) study now builds upon Phase 3 to investigate additional refinements. The Figure 1b depiction is enhanced with color coding, plus a proto-representation of wind component-speed variability built into the tail of each arrow (the color-coding will be the main emphasis of the current study). Figure 2b illustrates, with 2a showing the same information in textual format (to be used as the statistical control condition).



Sample Landing Scenario 2

21:23 Z

Reported Conditions for KXYZ - Landing Runway 28

Time	Winds
21:14 Z	11 kt from 340°
21:13 Z	16 kt from 340°
21:12 Z	14 kt from 340°
21:11 Z	12 kt from 340°
21:10 Z	13 kt from 340°
21:09 Z	10 kt from 340°
21:08 Z	13 kt from 340°
21:07 Z	11 kt from 340°
21:06 Z	11 kt from 340°
21:05 Z	16 kt from 340°
21:04 Z	10 kt from 340°
21:03 Z	15 kt from 340°
21:02 Z	11 kt from 340°
21:01 Z	11 kt from 340°
21:00 Z	12 kt from 340°
20:59 Z	16 kt from 340°
20:58 Z	11 kt from 340°
20:57 Z	12 kt from 340°
20:56 Z	13 kt from 340°
20:55 Z	14 kt from 340°
20:54 Z	14 kt from 340°
20:53 Z	10 kt from 340°
20:52 Z	11 kt from 340°

Click to assess this landing...

Sample Landing Scenario 1

21:23 Z

Reported Conditions for KXYZ at 21:14 Z - Landing Runway 04

Click to assess this landing...

a b

Figure 2. a) the control condition—textual information describing wind speed, direction, and variability, b) the experimental condition—the same information in graphical form (red color indicates crosswind exceeds pilot-specified thresholds). These two depictions will be tested against each other.

METHODS

Experimental Questions

Our primary goal was to answer specific questions about this latest version of low-altitude runway wind information depiction.

1. Would this enhanced graphical depiction be at least as accurate in determining landing difficulty as -textual depictions of runway wind information currently in common use?
2. Would pilots again find the graphical depiction faster to use, as they had in the previous experiment?
3. Would there again be a speed-accuracy tradeoff, as we found in our prior experiment?
4. How would greatly restricting the amount of time pilots had to view the runway wind information depiction affect their judgments about the landing difficulty posed by those winds?
5. How would pilots' ultimate decisions whether or not to land be affected by the manner in which runway wind information was depicted?
6. What relative levels of confidence would pilots express in the textual and graphical depictions, when compared with one another?

Experimental Method

Hardware and Software

The original Phase 3 MMET software was written by Dumont (2017), and was enhanced for this Phase 4 research by author McCarthy. It was written in JavaScript and packaged with Sencha Command, a tool for developing platform-independent JavaScript deployments. The web application bundle ran in the WebKit browser on an Apple tablet computer (iPad). The application loaded all scenario wind data from static files deployed with the application, and perturbed the data, adding a small amount of "speed noise" before showing it to pilots for evaluation. This ensured that not all scenarios of the same difficulty level would end up suspiciously having the exact same wind speed. Pilot interactions with the application were logged via WiFi connection to the internet, and the resulting records post-processed with a perl script to extract the experimental variables used to establish results presented here.

Measuring Quality of Information Depiction

This research explores methods of improving the display (depiction) of runway wind information. In order to claim that something is "improved," we have to have some reasonable way to measure display *quality*. We therefore chose to operationalize quality as:

1. Speed	of cognitive processing
2. Accuracy	of decision
3. Confidence-in-estimate	of decision accuracy

"Speed" of cognitive processing is simply how long it takes to make a decision whether to land, go around, or divert-to-alternate, given the low-altitude wind information. We used ***available viewing time*** as a proxy for speed. This was defined as the time from when the weather information first appeared to the pilot, until the pilot moved on to the Assessment screen (described below).

We defined "accuracy" as the difference between the ***perceived difficulty*** of a wind scenario and its ***objective difficulty***. The smaller the difference (defined as δ, "delta"), the better the wind display. This will be defined in detail presently. For now, Figure 3 illustrates the basic concept.

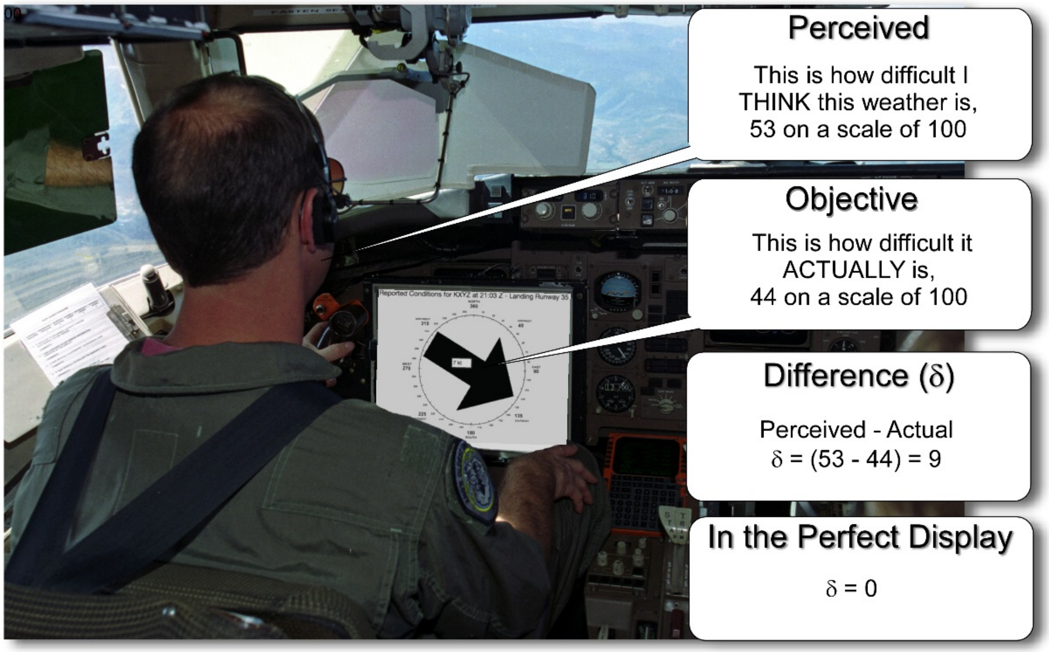

Perceived
This is how difficult I THINK this weather is, 53 on a scale of 100
Objective
This is how difficult it ACTUALLY is, 44 on a scale of 100
Difference (δ)
Perceived - Actual
δ = (53 - 44) = 9
In the Perfect Display
δ = 0
Reported Conditions for KXYZ at 21:03 Z - Landing Runway 35

Figure 3. "Display quality" was measured as the difference δ, defined as "perceived scenario difficulty minus objective scenario difficulty," both on a scale of 0-100. In a perfect display δ would equal zero.

"Confidence-in-estimate" simply meant "How confident was each pilot in their final decision about each scenario?" Now, we all know that "high confidence" in something does not guarantee "high quality." Nonetheless, confidence is useful in estimating how readily people may accept a new product, which is naturally of great concern to manufacturers.

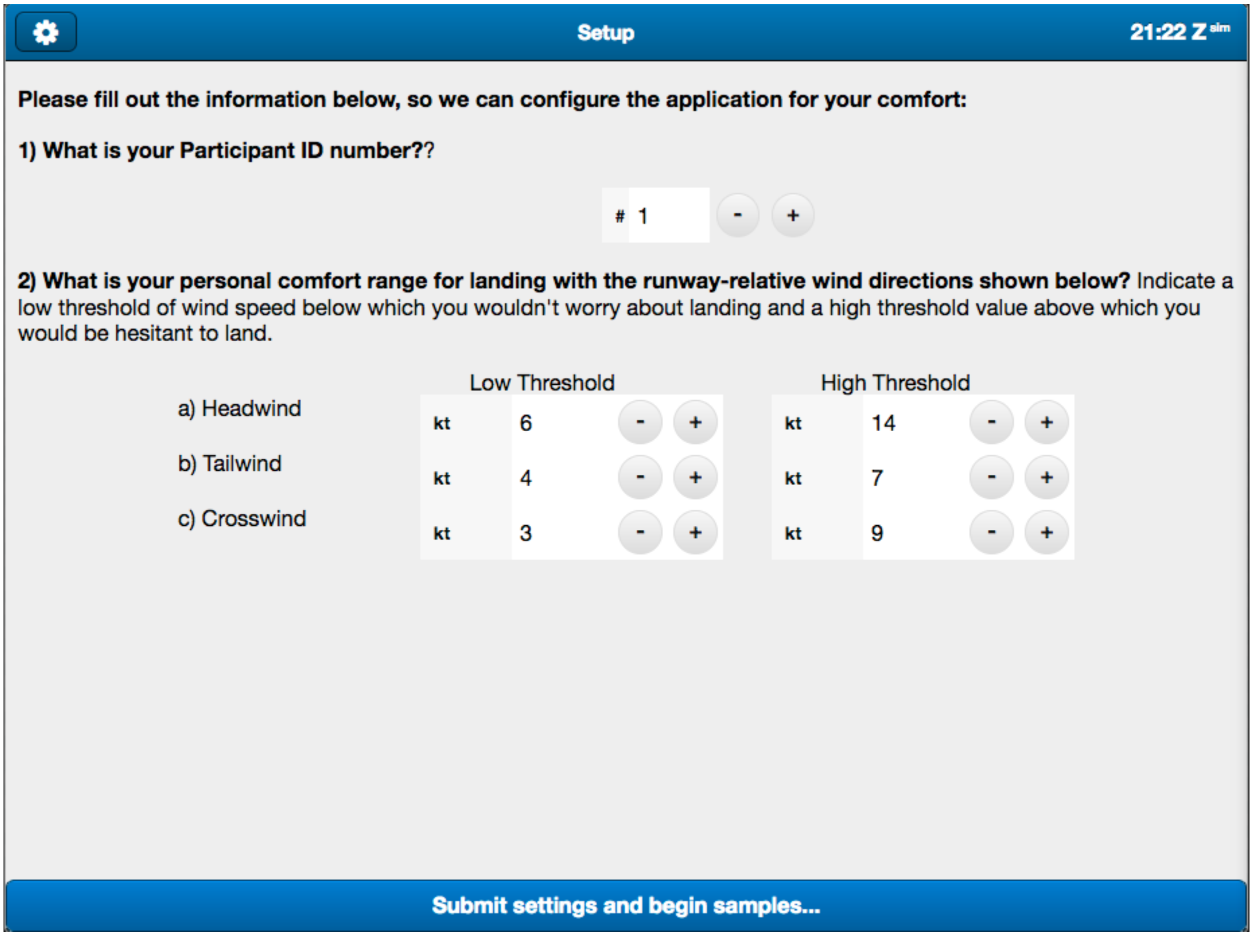
The terms "objective scenario difficulty" and "confidence-in-estimate" require detail concerning the exact method of calculation. We shall postpone discussing confidence-in-estimate since it is easier to understand after seeing the Assessment screen (Fig. 4 below). We therefore turn first to how we calculated scenario difficulty.

Calculation of Objective Scenario Difficulty.

Operationalizing our experimental method required wind scenarios with various *objective* (or "actual") levels of difficulty. However, this required controlling for each pilots' *skill* and *risk-tolerance*. For instance, if one pilot thought a 3-kt crosswind was "easy" and another thought a 5-kt crosswind was easy, then to construct an objectively "easy" scenario, we would obviously want the crosswind component to be between 0 and 3 kts for the first pilot and 0-5 kts for the second.

The mathematical term for this kind of individual tailoring process is called *normalization*, and its goal here is to create a single "normal" scale (e.g., 0-100) for "landing difficulty" that can be applied to all pilots, no matter what their skill or risk tolerance. This allows us to compute wind values to test each pilot individually, and then later compare them with one another statistically.

To create such a normal scale, during the Setup page at the very beginning of each pilot's test session, we had pilots give us their individual "thresholds" for wind-component speeds. Figure 4 shows a screenshot of how this looked.



Please fill out the information below, so we can configure the application for your comfort:

1) What is your Participant ID number??

1

2) What is your personal comfort range for landing with the runway-relative wind directions shown below? Indicate a low threshold of wind speed below which you wouldn't worry about landing and a high threshold value above which you would be hesitant to land.

	Low Threshold		High Threshold	
a) Headwind	kt	6	kt	14
b) Tailwind	kt	4	kt	7
c) Crosswind	kt	3	kt	9

Figure 4. Full-sized screenshot of the Setup page, showing, for example, a "Low Headwind Threshold" of 6 kt and a "High Crosswind Threshold of 9 kt" for a hypothetical pilot.

1. "Low Threshold" was defined as "Below that speed = I wouldn't worry about that wind component."
2. "High Threshold" was defined as "Above that speed = I would hesitate to land with that wind component."

Knowing each pilot's "easy" and "difficult" wind speeds allowed us to define wind speeds for "easy" and "difficult" scenarios for each pilot individually. Additionally, from these two values we could interpolate the remaining "moderate" level of difficulty by simply picking a wind speed halfway between the two extremes.

Appendix A details the exact algorithm used to construct the three levels of objective scenario difficulty. Essentially, what that algorithm did was to mathematically transform the range of speeds gathered from the Setup screen—something akin to "stretching a rubber ruler and then sliding it sideways"—until that old range now fit the new, "normal" 0-100 scale.

Theoretically, this individually customized method of creating scenarios should be far more objective and statistically sensitive than merely picking arbitrary wind speed values and assuming that their difficulty levels would be the same for all pilots. Because "objective difficulty" was now normalized on a standardized 0-100 scale, which controlled for each pilot's skill and risk-tolerance, we could, in theory, compute δs and thereby compare one pilot to another in terms of how closely their perceptions of a given scenario's difficulty matched its objective difficulty.

The Enhanced Graphical Depiction

Now—understanding how "landing difficulty" can be objectively defined for each pilot—one can fully appreciate the enhancements being tested in the current graphical wind depiction. The depictions we tested were precisely designed to communicate individualized landing difficulty both quickly and accurately. Figure 5 elaborates.

In Figure 5, the crosswind-component arrow depicts a most-recent reading of 12.2 kt, which greatly exceeds our hypothetical Figure 4 pilot's stated crosswind comfort-zone upper threshold of 9 kt taken from the Setup screen (Fig. 4). Hence, the crosswind's arrow tip is colored red to convey danger in the most-recent wind data available (21:14 Zulu, being reported at 21:23 = 9 minutes old). In contrast, the arrow tip of the headwind component is green, denoting "safe" in the most-recent wind data available, according to this pilot's stated standards.

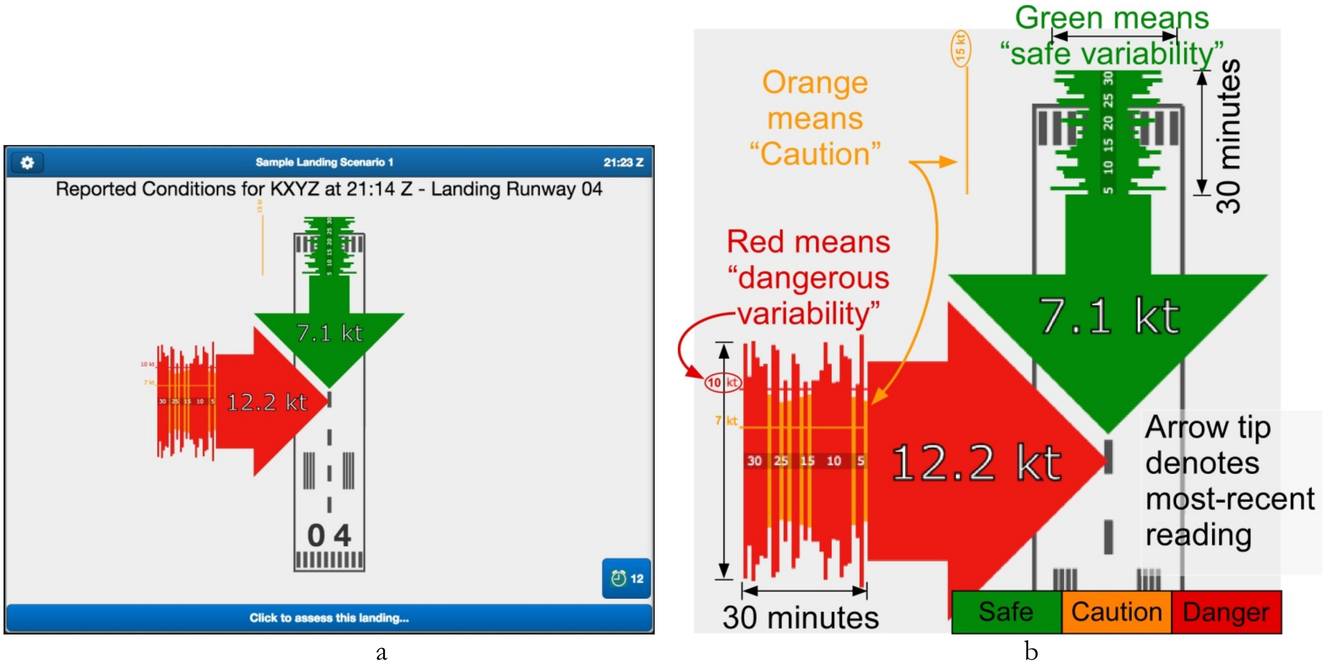


Figure 5. a) an example of the two-arrow landing-wind situation at 21:14 Z (being reported at 21:23 Z), b) enlarged, annotated view of its two wind components.

Note that the latency—the age of the data—should certainly be a factor in how much a pilot would trust these data. The older the data, the less trust we expect in them. Therefore, data latency was held constant during this experiment to negate any effect it might have on the statistical analysis (described later).

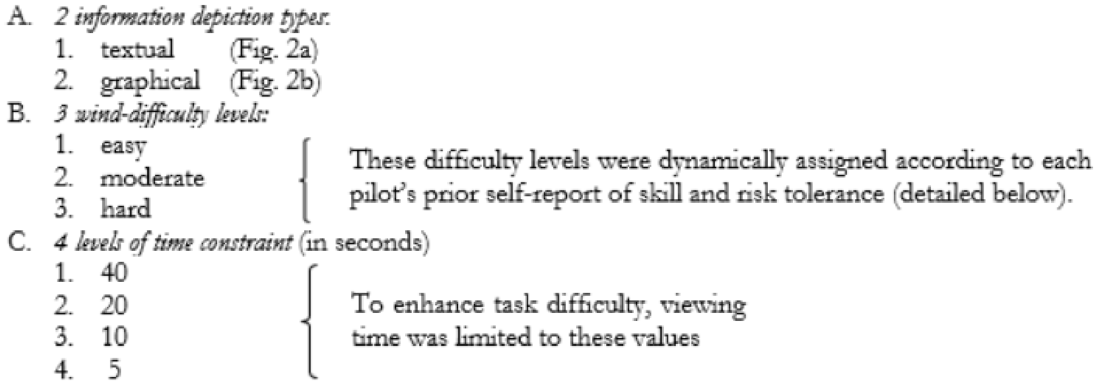
Experimental Design

In this section, we report features of the experimental design—the independent variables and dependent variables, the "look-and-feel" of the MMET assessment screen, plus how we controlled for unwanted experimental effects such as fatigue and learning effects.

Independent Variables (IV)

This study employed a *within-participants* (repeated measures) statistical design. Each pilot saw and responded to a set of 24 runway wind landing scenarios. Each scenario displayed a single page of wind information similar in appearance to Figure 2, depicting component-speed values that were customized for each pilot.

As the name suggests, independent variables are the factors we manipulate in an experiment to see how they affect designated outcomes (our dependent variables). The current design involved three IVs in a 2 × 3 × 4 = 24 total-trial design:



A. *2 information depiction types:*
   1. textual (Fig. 2a)
   2. graphical (Fig. 2b)

B. *3 wind-difficulty levels:*
   1. easy
   2. moderate
   3. hard

   These difficulty levels were dynamically assigned according to each pilot's prior self-report of skill and risk tolerance (detailed below).

C. *4 levels of time constraint* (in seconds)
   1. 40
   2. 20
   3. 10
   4. 5

   To enhance task difficulty, viewing time was limited to these values

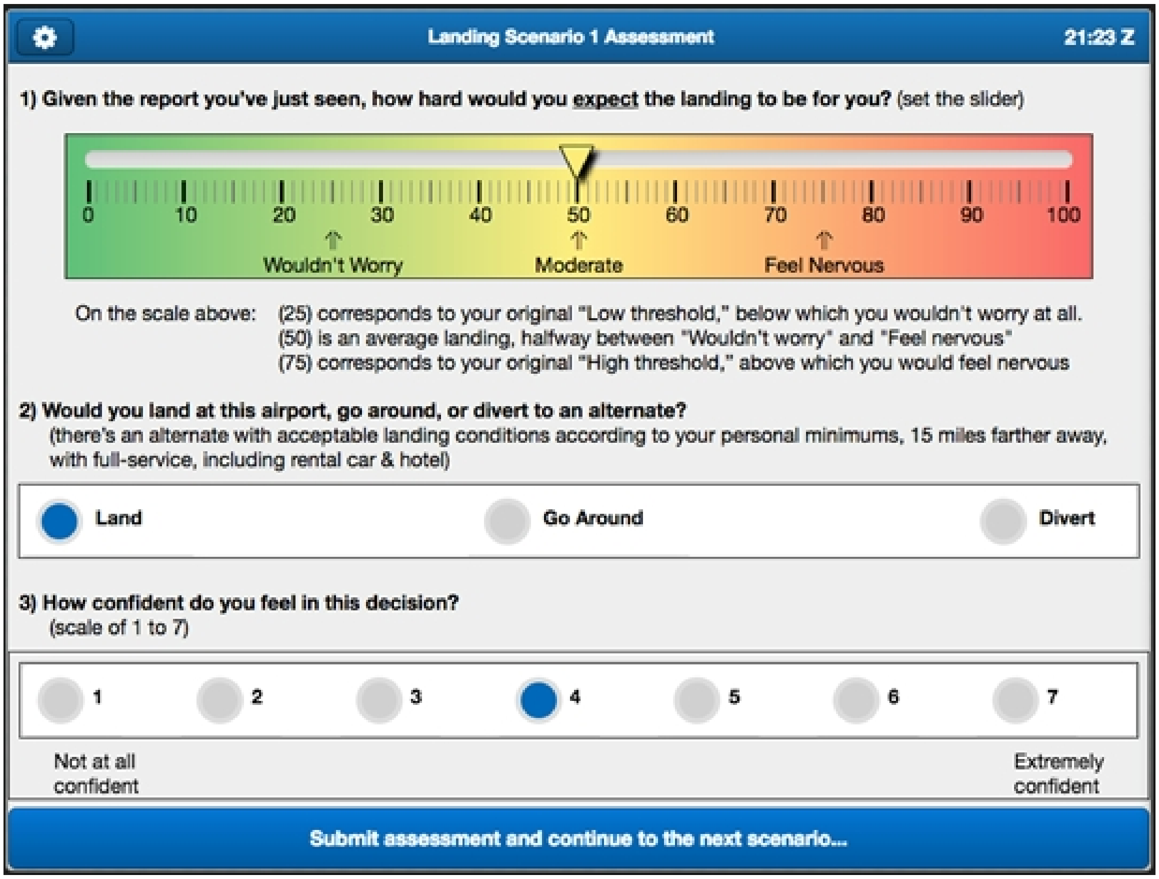
Dependent Variables (DV)

Dependent variables are experimental outcomes whose numerical value we hypothesize will depend on the numerical values we set up for our independent variables. As stated earlier, we set up three DVs:

1. Speed
2. Accuracy
3. Confidence-in-estimate

Speed. As previously mentioned, *Speed* was merely the time it took each pilot to assess the wind situation. This was used as a proxy for cognitive processing time. It was defined as the elapsed time from when the wind information page (Fig. 2a or b) was first shown to the pilot until the instant they moved on to the subsequent Assessment page (Fig. 6).

Accuracy. As also previously stated, *Accuracy* was defined as a difference score (δ, "delta"), equal to "perceived landing difficulty" minus "objective landing difficulty." For each scenario assessment, the pilot was asked to indicate *perceived* landing difficulty (PLD) by moving a slider along a scale such as shown in Figure 6. This slider showed the "normal scale" (0-100 "difficulty units"), representing how difficult the pilot *expected* the landing to be, given the wind information they had just seen, within the context of their own personal level of skill and risk tolerance, and their aircraft's capabilities (with the aircraft defined as the one they fly most often).



Landing Scenario 1 Assessment 21:23 Z

1) Given the report you've just seen, how hard would you *expect* the landing to be for you? (set the slider)

On the scale above: (25) corresponds to your original "Low threshold," below which you wouldn't worry at all.
(50) is an average landing, halfway between "Wouldn't worry" and "Feel nervous"
(75) corresponds to your original "High threshold," above which you would feel nervous

2) Would you land at this airport, go around, or divert to an alternate?
(there's an alternate with acceptable landing conditions according to your personal minimums, 15 miles farther away, with full-service, including rental car & hotel)

● Land ○ Go Around ○ Divert

3) How confident do you feel in this decision?
(scale of 1 to 7)

○ 1 ○ 2 ○ 3 ● 4 ○ 5 ○ 6 ○ 7

Not at all confident — Extremely confident

Figure 6. Full-sized screenshot of the Evaluation page.

Meanwhile, recall that each scenario's *objective* landing difficulty (OLD) had been customized for that pilot, based on her/his previously reported values for how wind speed and direction would affect landing difficulty for them, personally. Therefore, the assessment page gave everything else necessary to calculate a difference score, *perceived- objective* $=\delta$. And if, as hypothesized, one wind depiction was higher in quality than another, we would expect that either

1. most of its δ scores would be smaller, or
2. its δ scores would be similar, but pilots would take *less time* to make their risk estimates.

Independent variables B1 and B3 were used to test whether pilots could reliably discriminate between easy and difficult landing winds. Again using a 0-100 difficulty scale, the objective difficulty of B1 (Easy) landings was set at 20, and of B3 (Hard) landings was set at 90. Appendix A gives fuller detail.

In contrast, IV B2 was used to test a hypothesis of what mental model(s) pilots might use in estimating risk. Appendix B provides detail.

Confidence-of-Estimate. Figure 6 shows two putative measurements of pilot confidence in the quality of the information they had just seen. The first measurement was inferred from the pilot's choice whether to land, go around, or divert, given the wind information for that scenario. We generally expected to see landings increase as scenario difficulty decreased, with go-arounds reflecting difficulties in the middle range.

The 7-point Likert scale was expected to be a somewhat more sensitive measure of actual pilot confidence, mainly because it was a direct question about confidence rather than an inference based on the decision to land/go around/divert.

Control for Unwanted Experimental Effects

One final detail required attention: Repeated-measures designs must control for unwanted experimental effects such as fatigue and learning effects, particularly in an experiment such as this, having far too many DV combinations to counterbalance (i.e., to present every possible scenario presentation order to an equal number of participants).

Therefore, to counteract learning or fatigue over the course of each test session, the presentation order was set up to employ randomized-counterbalanced pairs. For instance, if one randomly generated presentation order with scenarios labeled A-R happened to be H J A G C B Q L F K M R N O E P I D, then a backward-pair D I P E O N R M K F L Q B C G A J H was presented to the next pilot.

No Correction for Familywise Error

Since this was an exploratory study, no correction was made for familywise error. Familywise error is the inflation of *Type I error* (a.k.a. a *false positive*—finding "significant differences" where there truly are none) due to conducting multiple statistical analyses within a single study. Logically, it is similar to reaching into a jar containing 95 white marbles and 5 black marbles, to see if a black marble might turn up. Doing this just once, one expects the probability of getting a black marble to be $p = .05$. However, if one replaces the withdrawn marble, and then repeats the entire process 100 times, the chance of getting a black marble at least once purely by chance increases greatly (to $1\text{-}.95^{100} \approx .994$), even though the underlying ratio of black to white never changes for each individual random draw.

In other words, the more times you repeat a chance process, the more times falsely "significant" results may occur purely by accident.

There are methods of correcting, or controlling, for this kind of error, to keep each individual analysis in a group of analyses "honest" at some stated value of significance (e.g., $p = .05$). However, in doing so, the *power* of the overall study—its ability to detect effects if they truly exist—decreases greatly. Therefore, it is common in broad, preliminary studies such as this one to omit the correction for familywise error, in the interest of boosting power. And, that is the approach we take here. Ideally, effects that are found preliminarily should later be replicated with a narrower study.

RESULTS

Pilot Participants

Seventeen general aviation pilots were recruited from a local flight school, and were paid $50 USD for their participation. Table 1 summarizes demographics.

Table 1. Pilot demographics (*N*=17).

Student	0	**CFII**	3	**Age-mean**	22.3	**TFH-mean**	323
Private pilot	17	**Commercial**	7	**Age-median**	22.0	**TFH-median**	200
Instrument-rated	15	**ATP**	0	**Age-SD**	3.4	**TFH-SD**	205
CFI	4	**Multi-engine**	4	**TFH^{A}-max**	800	**TFH-min**	98

^{A}Total flight hours

Pre-test Instructions and Practice

The pre-test instructions presented to pilots are shown in Appendix C. These included both text and screenshots of the application's Setup and Evaluation pages. Additionally, a sample sheet (not shown) was provided showing screenshots of the two wind depiction types in Figure 2, with text descriptions of all their important features.

Pilots were then walked through two practice pages, one for each of the two depiction types.

Preliminary Examination of Data

Recall that the research design was set up as 2 × 3 × 4 = 24 treatments of (Depiction Type (A) × Scenario Difficulty (B) × Time Constraint (C).

Before beginning intensive statistical analysis, we first checked the data to see if they satisfied the assumptions of repeated-measures analysis of variance (RM-ANOVA). If not, we tested alternate ways of analysis, and report here those found to be most satisfactory.

Perceived Landing Difficulty

The PLD Data Are Non-normal

Perceived Landing Difficulty was our proxy for accuracy—how accurately pilots could estimate the difficulty of landings that we had custom-engineered based on their personal minimums and maximums, that is the objective "easy" and "hard" wind-speed values each pilot had given us moments beforehand.

Preliminary statistical analysis of *PLD* by the Explore function of IBM SPSS revealed that non-normalities and outliers challenged analysis of the data by standard RM-ANOVA. Non-normalities are frequency distributions of IV component groups (e.g. "A1," "B2", or "C3") insufficiently resembling normal (Gaussian) bell-shaped curves. Outliers are values or scores far above or below a sub-group mean. Obviously these two conditions are related and, in fact, each can cause the other (e.g., if we add too many extreme low and high values to an otherwise-bell-shaped curve with, it will cease resembling a bell-shaped curve).

Figure 7 shows raw scores (taken right from the difficulty slider in Fig. 6, as opposed to the derived measure δ we defined earlier). Examining these raw scores, we see that the first of our three main independent variables, *Depiction Type* (with sub-groups A1-Textual vs. A2-Graphical), showed a severe dip in the middle of the both A1 and A2, and grossly failed the commonly used Shapiro-Wilk test[1] of normality ($p_{S\text{-}W\,A1}$ = 7.92*10^{-7}, $p_{S\text{-}W\,A2}$ =4.82*10^{-12}).

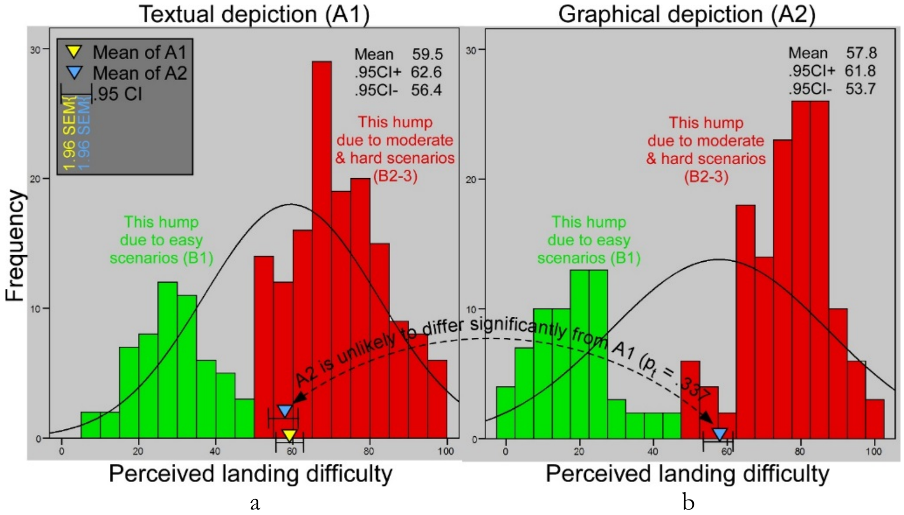


Figure 7. Frequency distribution (histograms) of *Perceived Landing Difficulty* for a) textual (A1) vs. b) graphical (A2) information depictions. The raw data fail normality testing, making RM-ANOVA inappropriate. Nonetheless, as we can easily see, the two sub-group means are so similar that their difference is unlikely to be statistically significant (p_t = .337).

Figure 8 shows that raw scores for the second main IV *Objective Landing Difficulty* (B) also displayed severe non-normality. The *PLD* frequency distribution for Easy landing scenarios (B1) had a long tail pointing rightward, while Moderate (B2) and Hard (B3) scenarios both had long tails pointing leftward. Consequently, all three sub-groups failed the Shapiro-Wilk test ($p_{S\text{-}W\,B1}$ = 2*10^{-6}, $p_{S\text{-}W\,B2}$ =4*10^{-6}, $p_{S\text{-}W\,B3}$ =.00318).

[1] In the Shapiro-Wilk test, a *p*-value < .05 is considered "failure," which means the frequency distribution in question is considered non-normal, and non-parametric tests should be used. If *p* > .05, the distribution "passes," is considered normal, and regular (parametric) tests may be used.

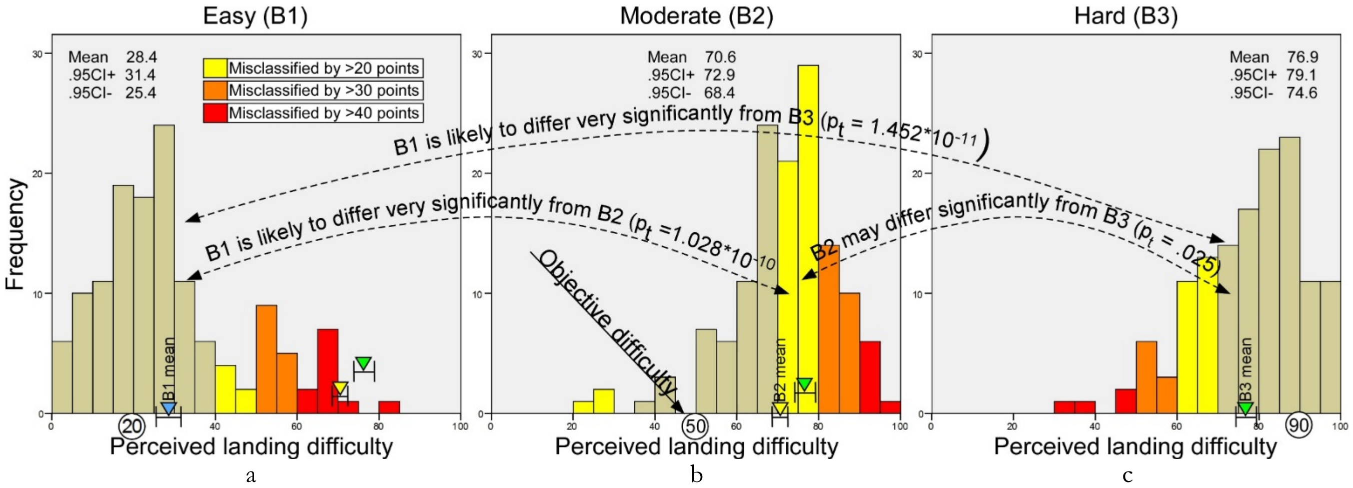


Figure 8. Frequency distribution of *Perceived Landing Difficulty* for a) Easy (B1) vs. b) Moderate (B2) vs. c) Hard wind-speed scenarios. Again, the data violate ANOVA's normality assumption, so we tested the main effect of B by averaging and paired-sample t-test.

Raw scores for third main IV, *Viewing Time* (C), fared no better. Like *Depiction Type*, the sub-groups were bimodal, Figure 9 shows the *PLD* frequency distributions. All four sub-groups failed Shapiro-Wilk ($p_{S\text{-}W\ C1}$ = 9.82*10^{-7}, $p_{S\text{-}W\ C2}$ =6*10^{-6}, $p_{S\text{-}W\ C3}$ =2*10^{-6}, $p_{S\text{-}W\ C4}$ =1.5*10^{-5}).

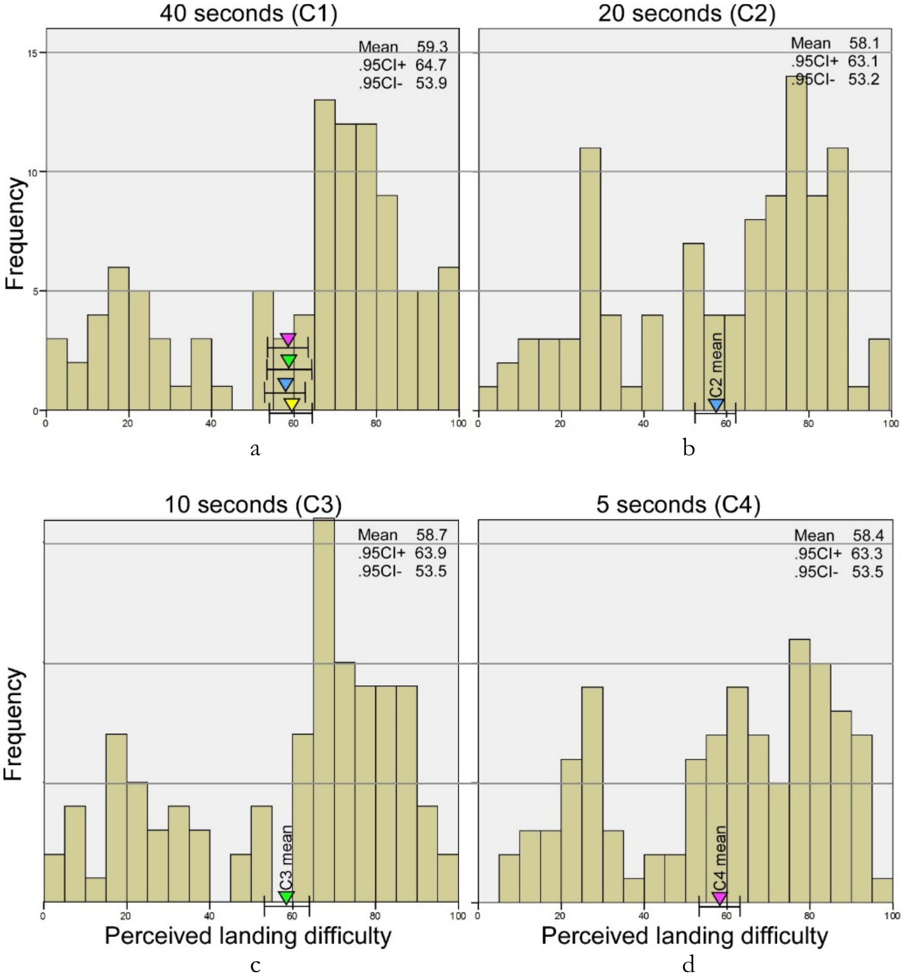


Figure 9. Frequency distribution of *Perceived Landing Difficulty* for viewing-time-constrained scenarios, a) 40-sec (C1) vs. b) 20-sec (C2) vs. c) 10-sec (C3) vs. d) 5-sec (C4). Again, the data violate ANOVA's normality assumption. As we can see, the four means are so similar that they are unlikely to be significantly different.

Results After Non-normality was Addressed.

There are methods of transforming data to make the frequency distributions more normal. We tried the standard methods on our raw scores, including logarithmic transform ($x_{new} = log(x_{old})$), power transform ($x_{new} = (x_{old})^a$), and winsorizing (replacing the lowest and highest values with copies of the next-lowest and next-highest, respectively, and then repeating the process, if necessary). We even tried deleting as many as four selected pilots' data completely[2] before and after applying various transforms. Yet, the modified distributions persistently failed normality testing, no matter what series of methods was applied.

Ultimately, alternate tactics were required. For one, we could calculate confidence intervals and visually look for main effects likely to be significant (e.g., whether A1 looked substantially different from A2). Figures 7-9 and 11-13 take

[2] Deleting a given pilot's data would be based on an assumption that many of those data were somehow faulty. For example, perhaps that pilot had not treated the experiment seriously, or had panicked after being timed-out on one or more scenarios. There should usually be some objective reason for suspecting this (e.g., if that pilot showed a large number of very low or very high scores that clearly looked suspicious).

this approach. Confidence intervals could be based on the standard error of the mean (SEM),[3] which is (rather miraculously) normal, no matter what the shape of a frequency distribution.[4] This method would, of course, not be strictly correct, since our repeated-measures data violated the confidence interval's assumption of independently sampled data. Nonetheless, we could still gain a sense about group means that happened to be either quite similar or very dissimilar.

Also, and more properly, given three IVs, we could at least test *main effects* (e.g., whether the *average* of all 12 A1 scores was reliably different from the *average* of all 12 A2 scores). Collapsing the data across two IVs to yield one average score from each pilot on the remaining IV would create a single score for each pilot for each IV. Then, a paired-sample t-test could be used, if the n=17 frequency distributions were demonstrably normal, or a Wilcoxon non-parametric test of ranks otherwise. While this lacks the elegance and richness of RM-ANOVA, the results will be more convincing to a conservative audience.

After such averaging, indeed, the n=17 main-effect frequency distributions all passed Shipiro-Wilk with significances ranging from .219-.989. Having thus addressed the non-normality problem, Table 2 contains *p*-values and effect sizes[5] for 2-tailed paired t-tests of separate main effects.

Table 2. Main effects for *Perceived Landing Difficulty* raw scores (*p*-values of 2-tailed paired t-tests. Effect size Cohen's *d* in parens)

	A2	**B2**	**B3**	**C2**	**C3**	**C4**
A1	.337 (.69)					
A2						
B1		$1.028*10^{-10}$ (15.10)	$1.452*10^{-11}$ (16.97)			
B2			0.025 (2.34)			
C1				.412 (.47)	.660 (.25)	.590 (.36)
C2					.742 (.23)	.853 (.09)
C3						.818 (.12)

Main Effect for Depiction Type (IV A). Table 2 indicates that, just like the Phase 3 experiment, the main effect of textual (A1) versus graphical (A2) depiction on resulting *PLD* was non-significant (p_t = .337), This implies that—even given constrained viewing times—*on average, textual depiction seemed to produce about as good judgment of landing difficulty as did graphical depiction* (however, be advised that the "A closer look" analysis below will tell a much richer story).

Main Effect for Objective Landing Difficulty (IV B). The ability to discriminate between Easy, Moderate, and Hard objective landing difficulties were all significant. Particularly, Easy versus Hard difficulties (B1 vs. B3) were highly significant (p_t = $1.452*10^{-11}$), with a gigantic effect size (*d* = 16.97 standard errors). This tells us that, *on the whole, pilots were able to differentiate at least between easy and hard landings* (although, again, check "A closer look" for details).

Main Effect for Level of Time Constraint (IV C). The *p*-values for the main effect *Time Constraint* on *PLD* were all non-significant, ranging from .412-.853. This implies that *on the whole, available viewing time, from 40 seconds down to 5 seconds, exerted no significant effect on perceived landing difficulty*. But—again—we need to take a closer look.

A closer look. Despite these bland results, there was more to *Perceived Landing Difficulty* than merely main effects. Main effects concern group averages. Yet, there is often great value in examining the frequency distributions themselves because those can tell us about *variation.* For instance, how often were pilots far off the mark in their judgments of landing difficulty?

A quick look back at Figures 7-9 shows there was considerable variation in these data. That means we need to realize that, despite averages, pilots are not uniform, precise computing machines when it comes to assessing windy landings.

[3] *SEM = standard deviation/sqrt(n)*, and is used to perform a statistical test (*z*-test) on independently sampled means. Any two means separated by more than 1.96 SEM are considered to be significantly different at the *p=.05* level. Thus, the area on the *x*-axis encompassing 1.96 SEM above and below a mean is called its *.95 confidence interval (CI)*. The caveat here, however, is that *z*-tests assume independent measurements, whereas our data were considered correlated since each pilot provided 24 data points.

[4] Normality of the SEM is described by the Central Limit Theorem in probability theory.

[5] Here, the effect size, Cohen's *d*, is a type of *z*-score, being the (difference between the two means) / (pooled standard error of the mean). Effect sizes of about 2 or greater indicate strong support for two groups being reliably different from each other.

Error Scores δ (Delta)

Recall our discussion of *δ*, which was defined as an error score, specifically (*Perceived Landing Difficulty - Objective Landing Difficulty*). Recall, also, the theoretical notion that, in "perfect perception" of landing difficulty, *δ* would equal zero.

Referring back to Figure 8, pilots' perceptions of landing difficulty (*PLD*) were often somewhat low or high (*δ* = ±20-30 points, colored yellow in Fig. 8). Sometimes, they were considerably high or low (*δ* = ±30-40 points, colored orange). In a few cases PLD was way off (*δ* more than ±40 points, colored red). All these cases can be called *misclassifications*, and we will later question what caused them.

To help visualize the distribution of misclassifications, we can plot frequency distributions of *δ*. In Figure 10, *δ* scores of zero denote correct judgments of landing difficulty. Scores less than zero represent *underestimates* of difficulty, scores greater than zero represent *overestimates*.

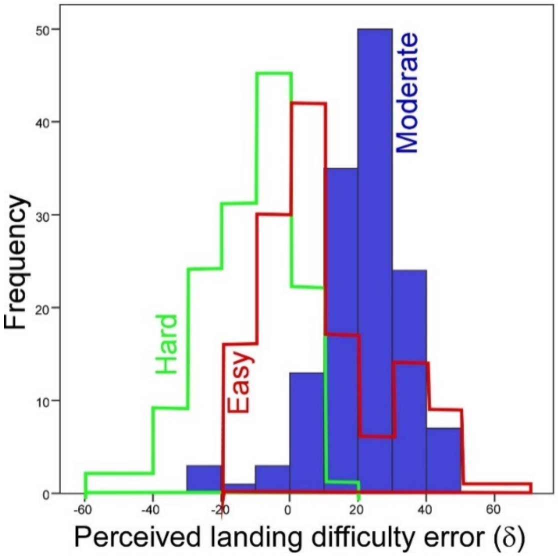


Figure 10. Perceived landing difficulty estimation error (*δ* = *Perceived Landing Difficulty* – *Objective Landing Difficulty*) for objectively Easy (B1) vs. Moderate (B2) vs. Hard (B3) scenarios. Easy scenarios are often rated more difficult than they objectively are, while Hard scenarios are often rated less-difficult than they are.

Table 3. Error (*δ*) for perceived landing difficulty (DV B) in Fig. 8.

Objective scenario difficulty	n	Mean raw score	Expected score	Mean *δ*	$p_{S\text{-}W}$
B1-Easy	136	28.38	20	8.38	.000002
B2-Moderate	136	70.63	50	**20.63**	.000004
B3-Difficult	136	76.87	90	**-13.13**	.003177

A value of $p_{S\text{-}W}$ < .05 on Shapiro-Wilk, implies significant non-normality.

Note that, in the case of Hard (difficult, B3) landings, difficulty was often underestimated (*δ* < 0, judged as *less* difficult than the landings objectively were). Symmetrically, Easy (B1) landings were often overestimated (*δ* > 0, judged as *harder* than they objectively were). The same result was also seen in our prior experiment (Knecht & Dumont, 2017), and so was not entirely unexpected.

Table 3 shows these effects using means and normality scores for *δ*. Moderate scenarios appeared to be the least-accurately judged, but let us suspend judgment on that for now. We will explore Moderate-difficulty scenarios in greater detail later, in the section titled *B2 Results and Their Significance.*

Graphical depictions produced fewer major misclassifications than did textual depictions. Table 4 numerically summarizes these major misclassifications, that is, all B1 and B3 *PLD* scores colored yellow, orange, or red in Figure 8 (i.e., all Easy scenarios rated greater than 40, and all Hard scenarios rated < 70). We see in Table 4's bottommost row that textual display produced 56 misclassifications, compared to graphical display's total of just 19. Table 4's 17 row totals fails normality testing for the graphical data, but we can still compare the two row totals with the non-parametric 2-tailed Wilcoxon test, which yields a significant p_W = .010.

We may therefore claim that—even though the <u>means</u> of the textual and graphic *δ* scores were similar—*textual depictions produced significantly more misclassifications of landing difficulty than did the graphical depictions.*

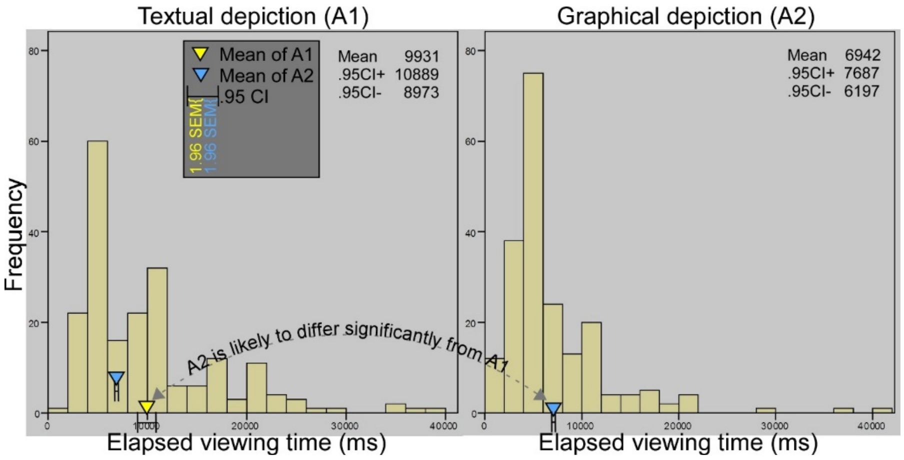
Table 4. Misclassifications of landing difficulty.

	A1 (Textual)									A2 (Graphical)								
	B1 (Easy)				B3 (Hard)					B1 (Easy)				B3 (Hard)				
S	A1B1C1	A1B1C2	A1B1C3	A1B1C4	A1B3C1	A1B3C2	A1B3C3	A1B3C4	Row Totals	A2B1C1	A2B1C2	A2B1C3	A2B1C4	A2B3C1	A2B3C2	A2B3C3	A2B3C4	Row Totals
1							65		1									0
2					69	63		63	3									0
3	42				60		67	59	4									0
4					39	52	54	47	4									0
5									0									0
6									0									0
7			45	50		70	64	59	5							65	64	2
8									0					70	56	66	60	4
9	50	59	63	67	70	69		62	7									0
10					68			50	2					50		70	69	3
11	69		65						2									0
12	65	40		55	66	68	69	34	7								63	1
13				42	70		62		3									0
14	55	54		50			50		4		52		43					2
15		55			70	60	66	45	5								54	1
16	68	80	71	58					4									0
17	50	54	67	62	63				5	50	54	45	65		68		70	6
COUNT	7	6	5	7	9	6	8	8		1	2	1	2	2	2	3	6	
Bs	*25*				*31*					*6*				*13*				
As	Total Textual Misclassifications = 56									Total Graphical Misclassifications= 19								

Elapsed Viewing Time

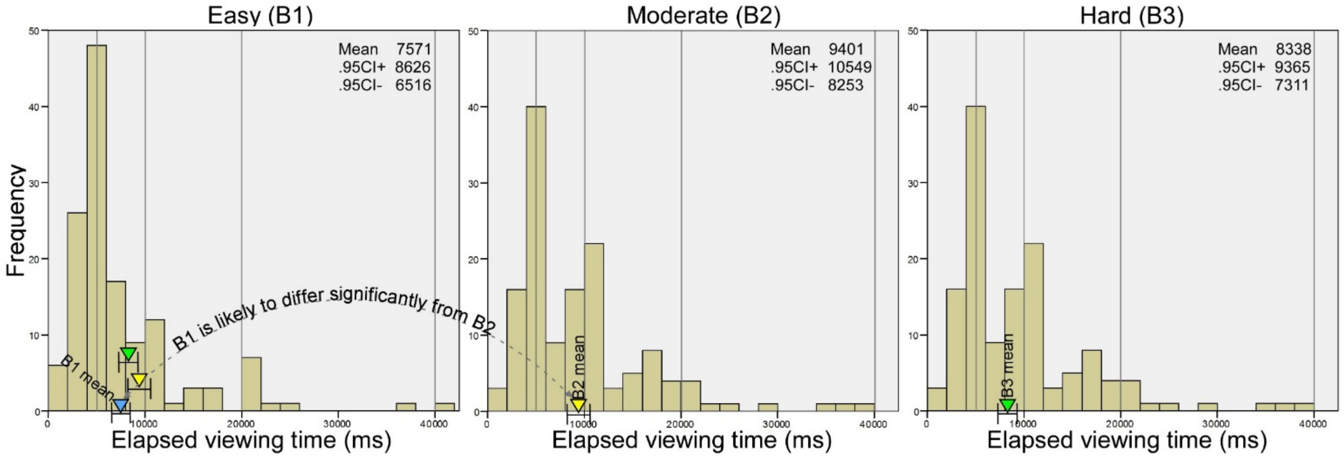
The ET Data Are Also Non-normal

Elapsed Viewing Time (ET) was our proxy for cognitive processing speed. Like the PLD data, the ET data were also non-normal. Figures 11-13 show this clearly. But, again, we can informally scan for significant group mean differences by creating confidence intervals and then collapsing each IV's data across the other two IVs, using the Wilcoxon to test for main effects.



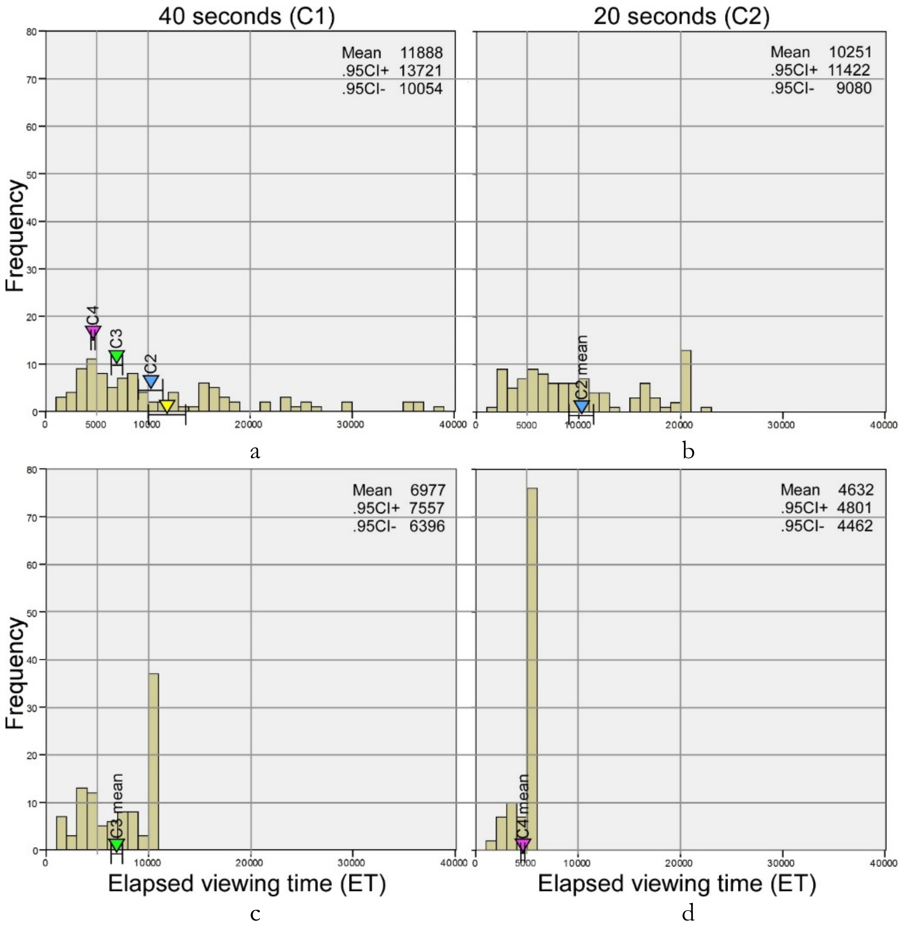
a b

Figure 11. Frequency distributions of *Elapsed Viewing Time* for a) textual (A1) vs. b) graphical (A2) information depictions. The raw data severely fail the Shapiro-Wilk normality test ($p_{S\text{-}W\ A1} = 2.30 * 10^{-14}$, $p_{S\text{-}W\ A2} = 6.09 * 10^{-18}$). Nonetheless, the two means appear likely to be significantly different if we examine the confidence intervals.



a b c

Figure 12. Frequency distributions of *Elapsed Viewing Time* for a) Easy (B1), b) Moderate (B2), and c) Hard (B3) objective landing difficulty. The raw data severely fail the normality test ($p_{S\text{-}W\,B1}$ = 7.31* 10^{-15}, $p_{S\text{-}W\,B2}$ = 5.37* 10^{-12}, $p_{S\text{-}W\,B3}$ = 8.85* 10^{-12}). Nonetheless, B1 looks likely to differ significantly from B2.



a b

c d

Figure 13. Frequency distribution of *Elapsed Viewing Time* for viewing-time-constrained scenarios, a) 40-sec (C1) vs. b) 20-sec (C2) vs. c) 10-sec (C3) vs. d) 5-sec (C4). Again, the data violate ANOVA's normality assumption ($p_{S\text{-}W\,C1}$ = 1.52* 10^{-8}, $p_{S\text{-}W\,C2}$ = 5* 10^{-6}, $p_{S\text{-}W\,C3}$ = 1.76* 10^{-8}, $p_{S\text{-}W\,C4}$ = 9.66* 10^{-15}). As we can see in panel (a), the four means look likely to be significantly different, with the exception of C1 versus C2.

Results After Non-normality is Addressed.

Similar to the approach taken with *Perceived Landing Difficulty*, we first generated confidence intervals for *Elapsed Time*, visible in Figures 11-13. While not strictly accurate, they are close, and can help us visualize what to expect from subsequent numerical testing.

We then collapsed the time data for each of the three IVs (A, B, C) across the other two to yield averages. After averaging, the three n=17 frequency distributions all passed Shipiro-Wilk with significances ranging from .207-.878, with the exceptions of B1 ($p_{S\text{-}W}$ = 043) and C4 ($p_{S\text{-}W}$ = 7.9* 10^{-5}).

Table 5 contains *p*-values and effect sizes for paired t-tests where normality permitted, with Wilcoxon *p*-values for variable pairs involving a non-normality (i.e., anything paired with B1 or C4).

Table 5. Group means and significances for *Elapsed Viewing Time* (*p*-values for 2-tailed paired t-tests with effect size *d* in parens. Wilcoxon tests highlighted yellow).

	IV group means (milliseconds)								
	A1	**A2**	**B1**	**B2**	**B3**	**C1**	**C2**	**C3**	**C4**
	9931	6942	7571	9401	8338	11888	10251	6977	4632

	Significances between IV pairs (and Cohen's *d*)					
	A2	**B2**	**B3**	**C2**	**C3**	**C4**
A1	$7.21*10^{-5}$ (2.92)					
B1		.004 (1.59)	.149 (.69)			
B2			.088 (.93)			
C1				.040 (.99)	$2.88*10^{-4}$ (3.44)	$2.93*10^{-4}$ (5.32)
C2					$1.55*10^{-4}$ (3.14)	$2.93*10^{-4}$ (5.89)
C3						$5.03*10^{-4}$ (4.89)

Results for Depiction Type (IV A). Graphical depictions produced faster judgment of landing difficulty. Table 5 indicates that, just like the prior experiment, the main effect of textual (A1) versus graphical (A2) depiction was significant (p_t = $7.21*10^{-5}$). Averaged across the entire 17 pilots, graphical depiction was an average of 2.99 seconds faster per scenario, representing just 70% as much processing time.

Results for Landing Difficulty (IV B). Pilots spent the most average time on Moderate-difficulty landings (B2, 9401 ms). *They spent about the same average time on hard landings as they did on easy ones* (B1 vs. B3). This may seem counter-intuitive, but interviews with pilots afterward revealed that they found it relatively easy to quickly scan even a long column of numbers, as long as those numbers were either very small or very large and did not vary much. This was clearly a heuristic, and we will elaborate on this in the **Discussion** section.

Results for Level of Time Constraint (IV C). Each of the four time-constraint values differed significantly from the others. Table 5 shows that all pairwise comparisons had significant *p*-values, ranging from .040-.0000721. This is not surprising, since most pilots tended to use extra time whenever it was available.

The percentage of times pilots ran out of time depended on how much time they were given. Given the maximum of 40 seconds (C1), only a single pilot timed-out on a single scenario. In contrast, over 74% of scenarios timed-out when pilots were given only 5 seconds (C4). Figure 14 presents time-outs grouped by time constraint and depiction type (graphical vs. textual).

Overall, pilots timed out on significantly more textual than graphical depictions, (75 vs 53). This is evidenced by collapsing each pilot's 24 scenarios down to one total per pilot for A1 and one total for A2. Comparison of the 17 pilots then showed acceptable Shapiro-Wilk normalities ($p_{S\text{-}W}$ = .110, .078), with a 2-tailed paired t-test significant at p_t = .010.

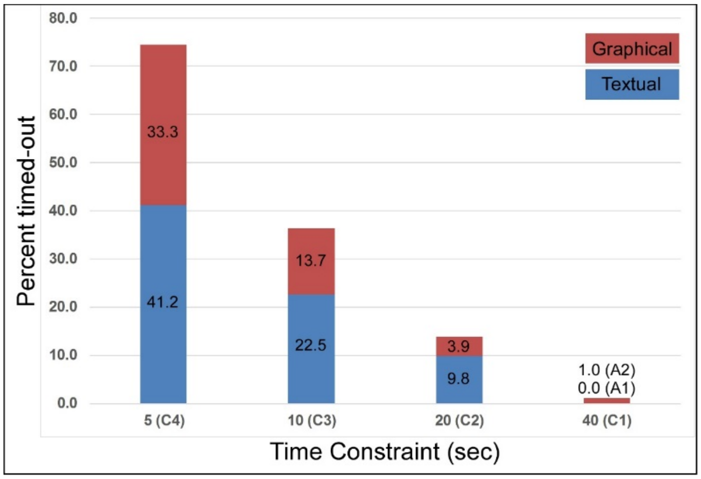


Figure 14. Percentage of scenarios that timed-out (graphical in red, textual in blue).

Relation Between Pilot Total Flight Hours and Misclassifications

It seems intuitive that more-experienced pilots might be better at classifying landing difficulty than less-experienced pilots. However, there was no significant evidence to support this. While there was a negative correlation between the number of *PLD* misclassifications and total flight hours (TFH), consistent with the notion that greater TFH might result in fewer misclassifications, that correlation was low and non-significant (*rho* = -.331, 2-tailed *p* = .195).

Speed-Accuracy Tradeoff

Ideally, we would expect to find a direct speed-accuracy tradeoff, in that "more careful" pilots might spend more time, and thus end up with lower error in their judgments of landing difficulty. In functional terms, we might expect a negative correlation between *ET* and $|\delta|$ (the absolute value of delta)—the higher the *ET*, the lower the $|\delta|$.

In fact, there was no evidence of such a direct relation. Exhaustive examination of correlations (non-parametric Spearman, *rho(A,* $|\delta|$*), rho(B,* $|\delta|$*), rho(C,* $|\delta|$*), rho(AB,* $|\delta|$*), rho(AC,* $|\delta|$*), rho(BC,* $|\delta|$*), rho(ABC,* $|\delta|$)) produced no significant results.

However, that does not erase the previous finding that graphical depictions took just 70% as much average viewing time, with no significant loss in *PLD* accuracy.

Landings

Visual Inspection of Landings

Figure 15 shows *PLD* difficulty ratings grouped by *Depiction Type* (A1-A2, by rows), *Objective Difficulty* (B1-B3, within each figure), and by *Time Constraint* (C1-C4, by columns). Figure 15 makes it easy to see individual landings, which are represented by a circle around each difficulty rating. Being able to see individual landings makes it easy to spot the tendency to land during Easy scenarios, as well as misclassifications in *PLD* ratings, and appropriate-vs.-inappropriate landings.

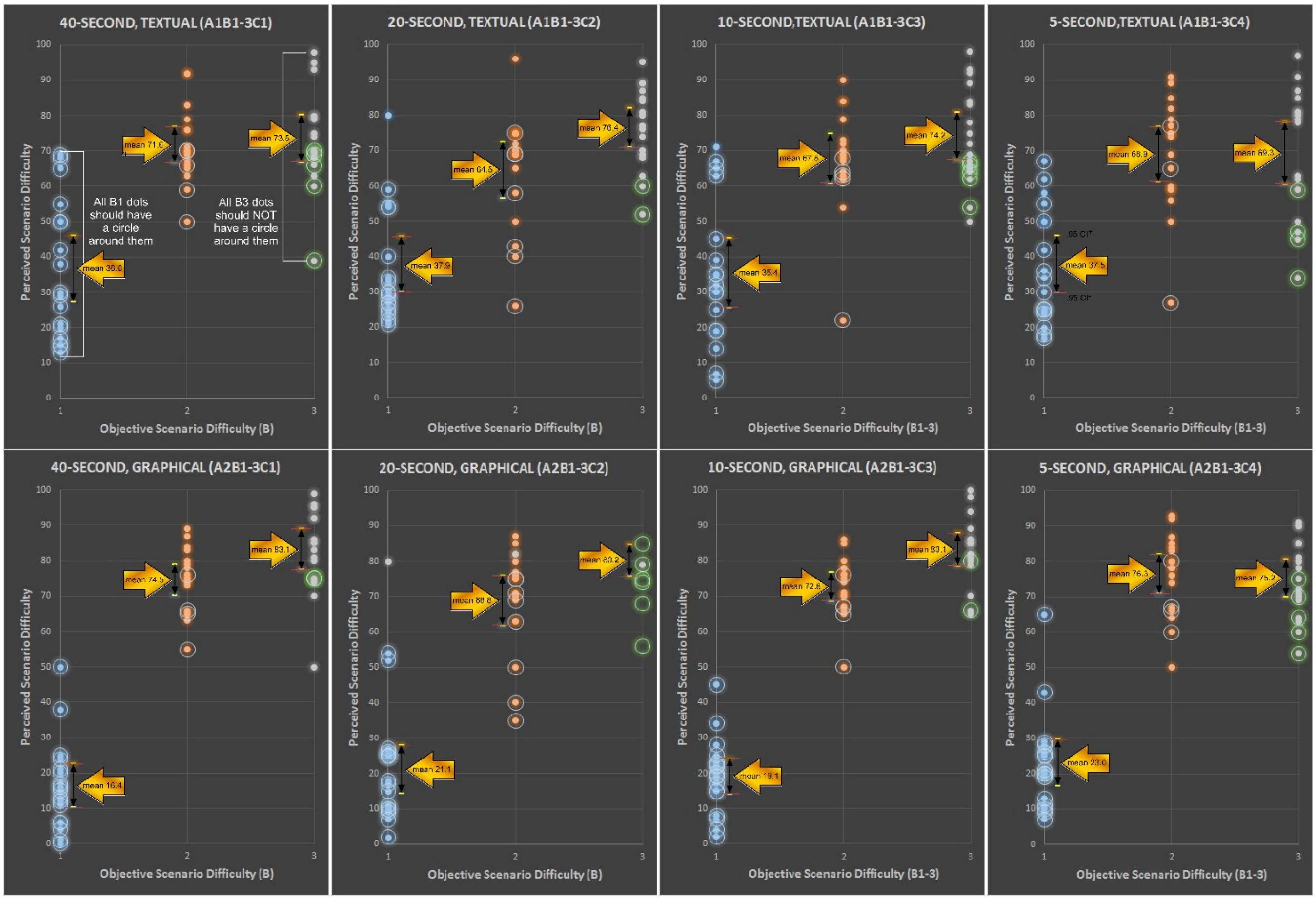


Figure 15. *Perceived Landing Difficulty* and *Landings, both* × *Objective Landing Difficulty* (B1-3). PLD is represented by a dot on the 0-100 scale. Dots with a circle represent landings. In theory, all Easy (B1) scenarios should have produced landings, while all Hard (B3) scenarios should have produced diversions or, at least, go-arounds.

Appropriateness of Landings

Discriminating between easy and difficult winds. The decision whether or not to land is the ultimate cognitive conclusion a pilot can make about a report of runway winds. Ideally, pilots will land when the winds are within their skill range, and will not, otherwise.

Figure 15 makes salient which landing decisions were appropriate. All B1 Easy scenarios should have produced landings and all B3 Hard scenarios should not. Each of Figure 15's eight panels shows objective difficulty (IV B) on its *x*-axis, versus perceived difficulty (*PLD*) on the *y*-axis. This highlights the variability in *PLD* produced by each ABC combination of IVs. The greater the range on *y*, the greater the variability of *PLD*.

Table 6 presents a summary of the same "landing appropriateness" data in "signal detection theory (SDT) format," as if unacceptably high winds were a "signal" that can be detected or missed. This tells us whether pilots landed when they should have, and whether they diverted (or at least went around) when they should have.

Table 6. Appropriateness of landings[A] for Easy (B1) vs. Hard (B3) scenarios, with data summed across level of time constraint (C).

		Textual (A!)					Graphical (A2)		
		Did land					Did land		
		Yes	No				Yes	No	
Should have landed	Yes (B1)	63[B]	5[C]	68		Yes (B1)	68	0	68
	No (B3)	16[D]	52[E]	68		No (B3)	15	53	68
		79	57	136			83	53	136

[A] Gray cells denote appropriate response, either to land or not land.
[B] Equivalent to a "Correct Rejection" in SDT, given no signal of "High Risk."
[C] Equivalent to a "False Alarm" in SDT.
[D] Equivalent to a "Miss" in SDT.
[E] Equivalent to a "Hit" in SDT. "High Risk" signal was correctly detected.

Table 6 shows us that, no matter what the data format (textual or graphical), most pilots appeared able to discriminate between easy and difficult landing conditions ($p_{Fisher's\ exact}$ = .016). They usually landed when landing winds were easy for them, and did not when winds were difficult.

The graphical depiction (A2) had slightly more correct responses than the textual depiction (viz., 68 graphical Correct Rejections[6] vs. 63 textual, 53 graphical Hits vs. 52 textual). Unfortunately, we know of no appropriate statistic to analyze this exact format, given that each pilot's 24 separate decisions were correlated. Therefore, we cannot formally state that the graphical format was *significantly* better than the textual format. Moreover, since we can see that the differences were relatively small, we should prudently conclude that the textual and graphical display were indistinguishable in "signal-detection ability" of easy-versus-difficult landing winds.

The challenge of intermediate-difficulty winds. If there is a challenge in displaying runway wind information, it seems that will involve trying to depict intermediate-difficulty winds. Watch for this theme to progress momentarily.

Pilots' Confidence in Their Decisions

Pilot Confidence in the quality of their decisions was measured on a 1-7 Likert scale after each scenario (see Fig. 7). Overall, the data were heavily skewed toward the high end of the scale. So, similarly to the *Elapsed Time* data, we collapsed the *Pilot Confidence* data by categories to yield averages-by-subject main effects for IVs A, B, and C. Even after such averaging the n=17 frequency distributions all failed Shipiro-Wilk with powerful significances ranging from 1.01* 10^{-7} to 4.42* 10^{-14}.

We therefore ran pairwise Wilcoxon tests. Table 7 shows group means and follow-up *p*-values and effect sizes (Cohen's *d*).

Table 7. Group means and significances, *p*-values for 2-tailed tests of correlated pairs. Wilcoxon tests highlighted yellow. All others are t-tests.

	IV group means (Confidence-in-decision, 1-7 scale)								
	A1	**A2**	**B1**	**B2**	**B3**	**C1**	**C2**	**C3**	**C4**
	5.13	5.99	5.88	5.26	5.53	5.75	5.64	5.60	5.25

	Significances between IV pairs (and Cohen's *d*)					
	A2	**B2**	**B3**	**C2**	**C3**	**C4**
A1	1.74*10^{-13} (3.49)					
B1		8.41*10^{-8} (3.43)	.001 (1.45)			
B2			.014 (1.10)			
C1				.251 (0.60)	.265 (0.77)	1.25*10^{-4} (2.20)
C2					.874 (0.19)	.015 (1.68)
C3						.011 (1.48)

Results for Depiction Type (IV A). Pilots had significantly more confidence in their landing decisions when using the graphical depiction than the textual depiction. Table 7 indicates a large main effect of textual (A1) versus graphical (A2) depiction, significant at p_W= 1.74*10^{-5}. Averaged across the entire 17 pilots, graphical depiction was rated as producing 0.86 points higher confidence-in-landing decision on a scale of 1-7.

[6] If the signal is defined as "high risk due to winds", then a Correct Rejection says "I detect no 'signal.' I therefore reject the notion of great risk. I can land safely," (and all that is correct because there truly is no signal).

Results for Landing Difficulty (IV B). Pilots had the highest average confidence in their landing decisions during Easy landings (B1, 5.88). *They had the least confidence during Moderate landings* (B2). This makes sense, since interviews with pilots afterward revealed that they found it relatively easy to quickly scan even a long column of numbers, as long as those numbers were either very small or very large. This was clearly a heuristic. And, *intermediate wind values required the most mental effort because the heuristic could not be applied.*

Results for Level of Time Constraint (IV C). As Table 7 and Figure 16 show, all pairwise comparisons showed *the more time they had, the more confidence pilots had in their landing decisions.*

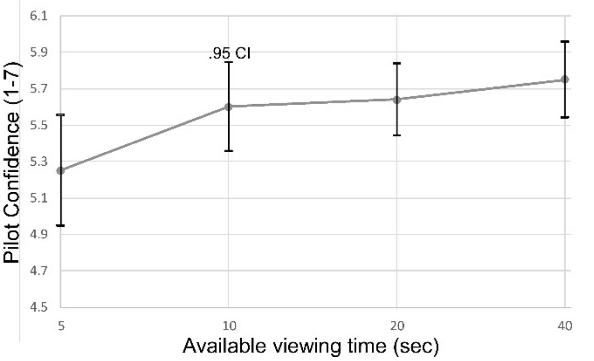


Figure 16. Pilots' confidence in their decisions to land, go-around, or divert-to-alternate (*y*-axis) as a function of available viewing time (*x*-axis, $\log_2$ scale). Error bars show the 95% confidence interval.

Higher confidence was weakly-but-positively associated with correct landing decision. Despite the absence of any feedback given about their decisions (or, perhaps *because* of it), pilots' confidence in their decision whether or not to land tended to match the eventual correctness of that decision, but not strongly so. Although we have no precise technique to measure these correlated data (because each pilot's scores will tend to correlate with themselves, and are therefore not strictly independent), we can get a rough estimate by calculating a Spearman rank-order correlation, which yields $r = .331$ ($p = .00000023$) between pilot decision-confidence level (scaled 1-7) of each scenario and that scenario's subsequent landing-decision correctness (scaled 0-1). Only Easy and Hard scenarios are included in that estimate, since they were obviously designed to have correct outcomes ("should land" and "should not land," respectively), whereas intermediate-difficulty scenarios were purposely designed to be equivocal.

Mindful of the limitations of that method, that correlation estimate would be classified as "modest" because the variance-explained ($r^2 = .11$) is small, and the high degree of significance (p) is mainly due to the large number of measurements ($n = 272$).

Interestingly, if we then group the data by *Depiction Type*, $p_{Spearman,\ Textual} = .252$, while $p_{Spearman,\ Graphical} = .432$. A Fisher r-to-z transformation test (again, not technically appropriate) yields a one-tailed $p = .048$, suggesting that the graphical depiction may be better than the textual at linking confidence to correctness of landing decision. Once again, however, we cannot formally state such a conclusion, due to the statistical considerations mentioned.

Pilot Opinions Concerning the Technology

All pilots were informally debriefed after finishing the testing session. The most salient aspect of their interviews was that *every pilot— unanimously—expressed the opinion that he or she preferred the graphical depiction to the textual, and this was the same outcome noted in our previous experiment.* We therefore now have two separate studies showing unanimous support for the graphical depiction of landing-wind information.

Pilot Heuristics

Heuristics are simplifying rules used in decision making. Interestingly, a few pilots revealed to us their "trick" for assessing landing difficulty in the Easy and Hard textual scenarios. This was simply that they ignored the runway angle, and landed, if the wind speed values were very low, and diverted if the values were very high.

This made sense with very low values, because trigonometry dictates that no separate wind component can ever be larger than the overall reported wind speed.[7] So, if the overall speed is lower than the pilot's personal minimums, the speed of its components has to be lower still, thus guaranteeing acceptable landing winds.

That approach was more of a gamble when applied to high reported wind speeds, though. If the overall wind speed was somewhat higher than a pilot's personal maximums, they might also ignore the runway angle, and chose *not* to land. However, because the wind components' speed would be lower than the reported overall wind speed, a safe landing opportunity might sometimes be missed. In signal detection theory, this would be equivalent to a false alarm (declaring detection of a signal of high risk when none was indeed present, box superscripted "C" in Table 6). And, indeed, we see five instances of that mistake in Table 6 for the textual depiction, versus zero for the graphical display.

The use of such methods is important for two reasons. Most importantly, we now know that at least some (if not most) pilots use heuristics, rather than trigonometry, in judging risk due to winds. Second, this needs to be a consideration in designing this type of research (see section **Suggestions for Future Research**). It essentially means that testing very low or very high wind speeds is a waste of effort.

B2 Results and Their Significance

Modeling Pilots' Risk-Evaluation Processes

Previously, we alluded to a special purpose for the Moderate-difficulty (B2) scenarios, namely that we wanted to gain some insight into how pilots cognitively process the graphical depictions in our scenarios. Appendix B provides detail. To quickly summarize, we constructed B2 scenarios to have Easy headwinds and Difficult crosswinds. And, we surmised that one of three cognitive risk-models would influence pilots' *perception* of landing difficulty:

1. A "conservative" one-factor model — Pilot would pick the *worse* of either crosswind or headwind
2. A "liberal" one-factor model — Pilot would pick the *better* of either crosswind or headwind
3. A two-factor model — Pilot would factor-weight-and-sum the crosswind and headwind

We hoped we might be able to distinguish which model was operating, on the basis of pilots' B2 *PLD* scores. For instance, the "conservative" one-factor model should result in relatively high *PLD*, since pilots would have two risk values to choose from, and would always choose the larger. Therefore, even though the objective B2 score would be 50, the *perceived* B2 scores would be closer to B3 scores.

Comparing the Data to the Models

Indeed, this is precisely what we see in Figure 15. Note that, in every one of Figure 15's eight panels, the middle arrow, which represents the B2 mean, is much closer to the B3 (Hard) mean than it is to the B1 (Easy) mean. The probability of this being true by chance for all eight means in Figure 15 is $p = 1/2^8 = .0039$.

Figure 17 shows this more dramatically, with the data simply grouped by difficulty level (B1-3). Note how the frequency distribution of Moderate scenarios (OLD = 50) far more resembles the Hard (OLD = 90) distribution than it does the Easy (OLD = 20) distribution.

[7] The basic formula for triangles is sqrt($x^2 + y^2$)= hypotenuse h (= stated wind speed, for our purposes). Implicitly, no component x or y can be larger than h.

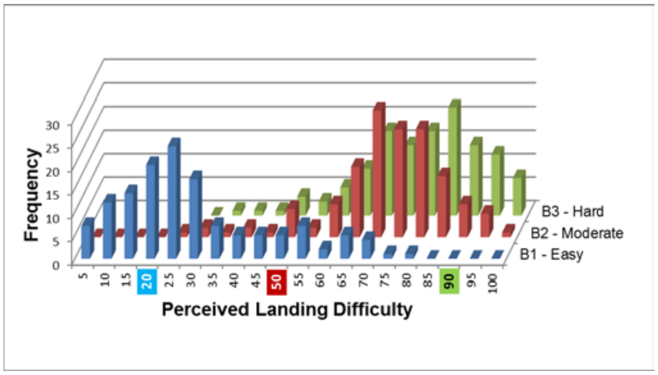


Figure 17. PLD frequency distributions, grouped by scenario difficulty. This disconfirms the “liberal” one-factor model.

Examination of the 17 individual pilots in Table 8 shows that—for every pilot but #17—the mean of their B2 *PLD* scores was closer to B3 than it was to B1 (p = .00027, 2-tailed sign test).

Table 8. Pilots’ individual mean *PLD* scores. The larger of B2 and B3 is highlighted gray.

S	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	mean
B1	19.38	22.38	25.50	19.88	14.88	25.25	31.38	23.88	39.63	14.38	32.00	28.25	13.13	45.63	26.88	44.13	55.88	28.38
B2	70.50	68.50	79.50	77.25	54.75	76.13	71.75	58.63	76.75	69.50	84.63	65.50	70.38	73.25	74.25	63.25	66.25	70.63
B3	79.13	77.13	73.50	67.25	84.88	84.00	68.13	72.63	78.00	66.75	94.13	70.38	77.00	88.50	66.25	81.38	77.88	76.88
B2 closer to	B3	B3	B3	B3	B3	B3	B3	B3	B3	B3	B3	B3	B3	B3	B3	B3	B2	**B3**

These experimental data clearly rule out the “liberal” one-factor cognitive risk model for the graphical depictions. Instead, they seem to support either the two-factor model weighted far more heavily toward the more-dangerous wind component, or the one-factor model, with the more-dangerous component considered exclusively.

The one caveat we need to present is that further research should be considered on this point. Because our B2 scenarios always presented Easy headwinds and Hard crosswinds, an experimental design using all combinations of headwinds, tailwinds, and crosswinds should be considered (Easy+Easy, Easy+Hard, Hard+Easy, and Hard+Hard). Such a design would address the issue of the relative importance of crosswind-versus-headwind/tailwind components.

One conclusion is certain, however: On average, *pilots tend to overestimate intermediate levels of risk* a bit, in apparent compensation for their lower level of confidence about the true value of the separate wind components.

DISCUSSION

This is the fourth in a series of studies aimed at developing an enhanced method to support in-flight pilot decisions by showing wind information on a mobile electronic device such as a tablet computer. In doing so, we hoped to reduce an identified gap in the general pilot skill base, namely that calculating wind components requires excessively high cognitive workload, leading to a safety risk.

The key to safe landing under windy conditions lies in *accurate* and *timely* assessment of runway wind components (headwind/tailwind and crosswind). “Accurate” means *correctly assessing landing difficulty by estimating runway wind component speeds at the area of touchdown, within the context of the pilot’s individual level of skill and risk tolerance.* “Timely” means that these *estimates will be made as fast as good accuracy allows, and as near to the area of touchdown as safely possible.*

In our previous (Phase 3) study, we discussed pilot perception of low-altitude runway winds as a *speed-accuracy process.* Pilots are trained to read information sources such as METARs, and extract wind components. As long as the necessary data are present, they try to accurately estimate those wind components to the best of their ability, no matter how long it takes. That study showed that they were reasonably adept at the task.

However, wind information depictions can vary in *efficiency* as well. Some depictions require more cognitive processing than others. Ultimately, the Phase 3 study showed that, *given equivalent levels of accuracy, a two-component graphical depiction was faster, and therefore more efficient, than the same information in textual form.*

As Figure 18a shows, the Phase 3 graphical wind depiction was quite sophisticated. It had two large arrows showing the runway-relative wind components and their speeds. This was designed to eliminate having to mentally estimate separate headwind/tailwind and crosswind components, or having to figure them from a chart.

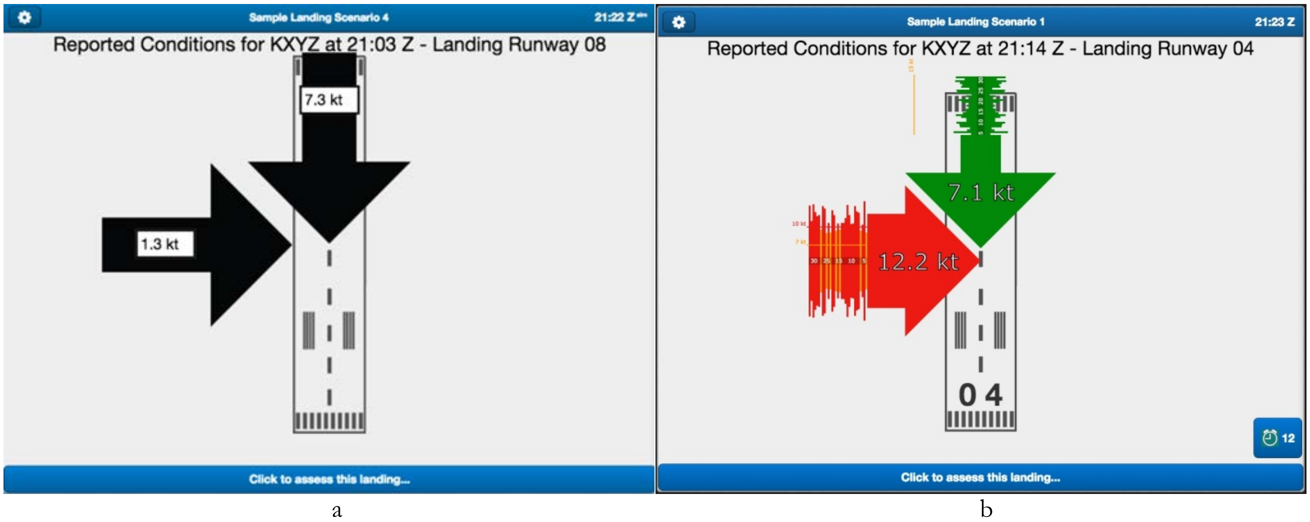
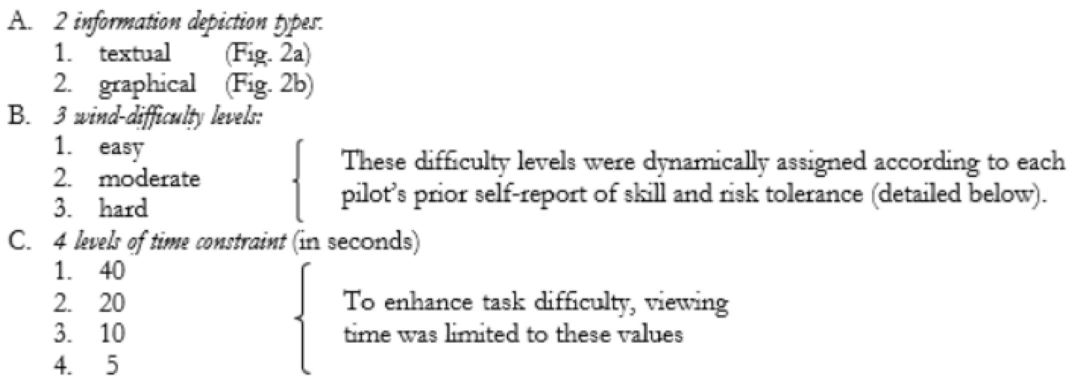

Sample Landing Scenario 4
21:22 Z
Reported Conditions for KXYZ at 21:03 Z - Landing Runway 08
7.3 kt
1.3 kt
Click to assess this landing...
a
Sample Landing Scenario 1
21:23 Z
Reported Conditions for KXYZ at 21:14 Z - Landing Runway 04
15 kt
5 10 15 20 25 30
7.1 kt
10 kt
7 kt
30 25 15 10 5
12.2 kt
0 4
12
Click to assess this landing...
b

Figure 18. a) the Phase 3 two-component graphical depiction, b) the enhanced Phase 4 graphical depiction, where green meant "This is within my personal comfort zone," orange meant "This is where I start to hesitate," and red meant "This is above and beyond my comfort zone."

Figure 18b shows that the current Phase 4 depiction was even more sophisticated. Its two arrows were now color-coded to represent *risk scaled according to each individual pilot's personal risk-tolerance and skill.*

Additionally, Phase 4 was set up to investigate how limited viewing time might affect judgmental accuracy. How might performance change if pilots were sometimes given only a bare minimum of time to view the weather depictions, as often happens in real practice?

This Phase 4 study therefore manipulated three independent variables (IVs):



A. *2 information depiction types:*
   1. textual (Fig. 2a)
   2. graphical (Fig. 2b)

B. *3 wind-difficulty levels:*
   1. easy
   2. moderate
   3. hard

   These difficulty levels were dynamically assigned according to each pilot's prior self-report of skill and risk tolerance (detailed below).

C. *4 levels of time constraint* (in seconds)
   1. 40
   2. 20
   3. 10
   4. 5

   To enhance task difficulty, viewing time was limited to these values

This 2×3×4 design therefore had 24 combinations, all shown to each pilot in repeated-measures format. The effects of those IVs were measured on three dependent variables (DVs):

1. Speed
2. Accuracy
3. Confidence-in-estimate

Like the Phase 3 study, Phase 4 relied on creating landing scenarios with *known* objective landing difficulty (*OLD*). This *OLD* was different for each pilot, customized according to their level of skill and risk-tolerance. Then, during each test scenario, based on the wind information shown to them, pilots reported their *perceived* landing difficulty (*PLD*). This allowed us to compute an error score by subtraction (*PLD-OLD*) representing the *accuracy* of risk judgment for every pilot on every scenario.

We also recorded how much time each pilot spent viewing each scenario's weather information. That elapsed time (*ET*) became a proxy for cognitive processing *speed. Decisions to land* or not land were also recorded. And, lastly, we asked pilots to tell us their level of *confidence* in each of their decisions.

The Phase 4 data were far more complex and difficult to analyze and explain than Phase 3's data. Phase 4 had more variability in responses, and more unexplained outliers in both the raw *PLD* and *ET* data.

Three main effects were examined regarding *PLD*. First, overall, textual depiction seemed to produce about as good average judgment of landing difficulty as graphical depiction (see Fig. 7 and Table 2, A1 vs. A2). However, averages did not tell the complete tale because the results of two relatively easy conditions to judge (low winds and high winds) were averaged in with one difficult condition to judge. This introduces the second point.

Second, *intermediate-speed runway winds are the hardest to decide upon*. Pilots could usually distinguish between easy landings and difficult ones (Table 2, B1 vs. B3) by using a simple heuristic requiring no mathematical calculation or chart lookup: *"Land if the wind speed is very low. Do not land if it is very high."* But, intermediate-speed winds required thought and work.

Third, on the whole, available viewing time, from 40 seconds down to 5 seconds, exerted no significant effect on *PLD*. However—once again—the entire story went deeper than it first appeared.

The full story was, of course, much richer than merely main effects of *PLD*. First—as in Phase 3—*graphical depiction was significantly more efficient than textual depiction*. Graphical depictions took only 70% as much cognitive processing time compared to textual—with no penalty in accuracy (Table 5). Moreover, when available viewing time was severely constrained to just 5 seconds, pilots timed out significantly fewer times with graphical depictions (Fig. 14, 53 graphical timeouts vs. 75 textual). Both these results indicate higher graphical efficiency.

Second, *graphical depiction produced significantly fewer misclassifications of landing difficulty* than textual depiction (Table 4, 19 vs. 56 misclassifications).

Third, *graphical depiction ultimately produced fewer mistakes in deciding whether or not to land* (Table 6).

Fourth, *graphical depiction may produce higher pilot confidence in their landing decisions, particularly when available viewing time is severely constrained* (Table 7). And, because those decisions may, indeed, be better, this higher confidence may be warranted.

Fifth, pilots appear to employ heuristics (simplifying rules) when estimating risk due to runway winds. *In textual depiction, especially when time is limited, pilots may ignore trigonometry and instead base their risk estimate as if the stated wind speed were the crosswind component.* Similarly, *in graphical depiction, pilots seem to focus on the more severe wind component as the limiting risk factor for that landing.* Interestingly, both heuristics lead to overestimation of risk, and more conservative landing behavior, which is not necessarily a bad thing when safety is the goal.

Finally, *pilots unanimously preferred the graphical depiction*, both in this study, and the previous one. Unanimity of preference is rare in product development, and manufacturers will take note of this.

SUGGESTIONS FOR FUTURE RESEARCH

A number of suggestions present themselves for future research. First, since trend and variability information are critical to real-world assessment of landing difficulty, both could be explored by the method suggested in Figure 5. Namely, that information could be represented in the tails of the wind-component arrows.

Second, two approaches might help alleviate the problem of "regression to the mean" in the data frequency distributions, which militated against the use of repeated-measures ANOVA in the current study. Focusing on graphic depictions only, and then spanning the entire range of *OLD* with easy, moderate, and difficult scenarios (instead of just easy vs. difficult) would reduce the non-normality of *PLD*.

Finally, exploring further which kind of heuristic pilots use to determine intermediate-risk situations with graphical depictions would be interesting and useful, since this appears to represent a natural bias toward increased safety with this kind of display.

ACKNOWLEDGMENTS

This research was funded by the FAA's Weather Technology in the Cockpit Program (WTIC Program), ANG-C61. Arnaud Dumont and Paddy McCarthy of NCAR's Research Applications Laboratory (RAL) designed and coded the mobile application itself. They also contributed significantly to the study design and methodology, as did RAL Program Manager Bob Barron. NCAR is a Federally Funded Research and Development Center sponsored by the National Science Foundation. We are grateful to Gary Pokodner and Andrew Mead for reviewing this manuscript and providing helpful comments. Special thanks go to David Lodes, Chief Flight Instructor, and Kenneth Carson, Program Director of the University of Oklahoma Department of Aviation, and their staff, for recruitment of pilot participants. Finally, we thank Daniel Jack for assistance with data collection.

This research is in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA.
FAA Contract: DTFAWA-15-D-00036
Task Order Number and Task Name: #21101 WTIC Mobile Meteorology (MET) Application
Deliverable: 2c – Analysis Report of Demonstration Results of Prototype Tool Enhancement
Delivery Status: Final

REFERENCES

Ahlstrom, U, Caddigan, E., Schulz, K., Ohneiser, O., Bastholm, R., & Dworsky, M. (2015). *Initial assessment of portable weather presentations for general aviation (GA) pilots.* (Technical Report DOT/FAA/TC-15/42).

Batt, R., & O'Hare, D. (2005). *General aviation pilot behaviors in the face of adverse weather.* (Report no. B2005/0127). ACT, Australia: Australian Transport Safety Bureau.

Knecht, W.R., & Dumont, A. (2017). *Tailoring surface winds information for mobile meteorological applications: Phase I, beta-testing.* (OAM technical report, in review).

Heitz, R.P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*. Retrieved February 13, 2017 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4052662/

National Transportation Safety Board (2005). *Risk factors associated with weather-related general aviation accidents* (Safety Study NTSB/SS-05/01). Washington, DC: Author.

APPENDIX A

Computation of Objective Scenario Difficulty

In this experiment, scenario winds were constructed to be "Easy," "Moderate," or "Hard," according to their *Objective Landing Difficulty* (*OLD*). The statistical challenge was to try to factor in each pilot's individual levels of skill and risk-acceptance, so that any given scenario would subjectively *feel* equally as difficult to all pilots. That "normalization" of the difficulty level would then allow the repeated-measures ANOVA to calculate a separate mean score for each pilot, and thereby detect *relative deviations from that individual mean*—changes in our dependent variables supposedly caused by the different ways we presented the wind information (i.e., our independent variables). This is how RM-ANOVA is able to statistically control for "individuality," by "using each subject as their own 'control'," resulting in very sensitive, powerful analysis.

The process of normalizing scenario difficulty began before testing, by collecting each pilot's "Low Threshold" and "High Threshold" (see Fig. 4). The Low Threshold was defined as "a wind speed below which you wouldn't worry about landing," while the High Threshold was defined as a "value above which you would be hesitant to land."

These two thresholds were then transformed to fit a normal 0-100 scale, according to the mapping equation

$$x_{0-100} = \frac{50\left(x_{4-10} - T_{Low}\right)}{T_{High} - T_{Low}} + 25 \tag{1}$$

which guarantees that the Low Threshold (T_{Low}) will map to 25 on the new scale, and the High Threshold (T_{High}) will map to 75. For example, if T_{Low} = 4 and T_{High} = 10, if x = 4 and 10, respectively, then

$$x_{0-100} = \frac{50(4-4)}{10-4} + 25 = 25 \text{ and } x_{0-100} = \frac{50(10-4)}{10-4} + 25 = 75$$

If x_{4-10} is halfway in between 4 and 10, then x_{4-10} = 7, and

$$x_{0-100} = \frac{50(10-7)}{10-4} + 25 = 50$$

Now we also needed an inverse[8] of Equation 1 to create working wind speeds for the experiment, and this was

$$x_{(T_{low} - T_{high})} = \frac{\left(x_{0-100} - 25\right)\left(T_{High} - T_{Low}\right)}{50} + T_{Low} \tag{2}$$

so plugging, for instance, x_{0-100} = 25, 50, and 75 into Equation 2 where, again, T_{Low} = 4 and T_{High} = 10,

$$\frac{(25-25)(10-4)}{50} + 4 = 4, \frac{(50-25)(10-4)}{50} + 4 = 7, \text{ and } \frac{(75-25)(10-4)}{50} + 4 = 10$$

being the original values we started with.

Given the definition of "objective landing difficulty," it now became possible to calculate δ, the estimate of *pilot error at recognizing the objective difficulty*, given a certain weather-information depiction.

[8] An inverse is a transformation that "undoes" some initial transformation. For instance, suppose some variable x_{old} were transformed by taking its square root, so that x_{new} = sqrt(x_{old}). The inverse transform would then be the one that restores x_{new} back to x_{old},. So, that inverse would be the square of x_{new}, or $(x_{new})^2 = x_{old}$.

Method of Generating Landing-Wind Scenarios

Generation Algorithm

The present study used the same computational method (Eqs. 1, 2) as the previous study (Knecht & Dumont, 2017). However, in the present study we mainly used two of the three wind-level difficulties (Easy and Hard) to test pilots' reactions to the various winds. The third (Moderate) difficulty level was used to test cognitive models pilots might be using to estimate risk (discussed in the section *B2 Results and Their Significance* section, sub-heading *Modeling Pilots' Risk-Evaluation Processes*).

The three scenario difficulty levels were thus constructed as follows:

1.	EASY	Easy headwinds, Easy crosswinds
2.	MODERATE	Easy headwinds, Hard crosswinds
3.	HARD	Hard headwinds, Hard crosswinds

Runway orientation was changed for each scenario, in order to force participants to analyze each scenario as unique. Runway orientations were randomly generated for 18 different compass directions, rounded to the nearest 10 degrees, with the north (360°), east (90°), south (180°), and west (270°) directions excluded as being too easy. The intent was to force participants to perform a different, non-trivial geometric transformation for each scenario. The same random runway orientations were used for each participant, but since their scenarios were in a different order, each combination of scenario and runway direction was unique.

As previously mentioned, each test began with the participant entering their low and high threshold wind speeds for headwind, tailwind, and crosswind. The application then calculated a wind speed and direction for every scenario. This was done by first calculating the wind vectors for the scenario difficulty, along the runway and perpendicular to the runway, then rotating those vectors for the random runway orientation of the scenario.

Tailwind landing conditions were excluded, based on the logic that the combination of tailwind and crosswind would introduce too much uncertainty into the pilot's decision-making, even despite the sufficiently long runway (8000') described to participants. Therefore, headwinds were used for every scenario. The Easy headwind speed was set to a constant value, which was 20% below the participant's low headwind threshold. This resulted in a uniform "easy" wind vector along the runway axis.

Working values were set according to the difficulty of the scenario. For "easy" scenarios, the working crosswind speed was set at 20% below the participant's low normalized crosswind threshold.

$$x_{working\_low} = \frac{(0.8x_{0-100} - 25)(T_{High} - T_{Low})}{50} + T_{Low} \qquad (3)$$

For "difficult" scenarios, working crosswind speed was set at 20% above the high normalized crosswind threshold.

$$x_{working\_high} = \frac{(1.2x_{0-100} - 25)(T_{High} - T_{Low})}{50} + T_{Low} \qquad (4)$$

For "moderate" scenarios, working crosswind speed was set as the mean of the low and high thresholds.

$$x_{working\_med} = \frac{x_{working\_high} - x_{working\_low}}{2} \qquad (5)$$

The direction of the crosswind was reversed for each *Enhanced* scenario in order to present a mirror-image runway-relative wind direction as the corresponding *Traditional* scenario. Additionally, to create realistic variations in the Minutely scenario reports, each of the Minutely crosswind speeds s_t was perturbed along a normal distribution ranging above and below the calculated scenario value v., producing a range of $.8v \leq s_t \leq 1.2v$.

The calculated wind vectors were then combined into a runway-relative wind direction and magnitude before being rotated around the compass relative to the random runway orientation. This resulted in a unique north-relative wind speed and direction used for each scenario. The test was, then, to determine how quickly and accurately the participant could determine whether the given wind regime was easy, moderate, or hard.

Controlling for Order Effects

Each odd participant (1,3,5…N-1) was assigned a pseudo-randomly ordered set of the 18 scenarios. Each even participant (2,4,6…N) was assigned a set of scenarios ordered the exact reverse of the previous participant. Therefore, S2's presentation order was a mirror-image of S1, and so forth. This mirroring served to cancel any learning or fatigue effects that might cause later scenarios to be interpreted differently from earlier ones. All scenarios orders were pre-generated and stored for later analysis, if needed.

APPENDIX B

Modeling Pilots' Risk-Evaluation Processes: B2 Results and Their Significance

The results stemming from responses to Moderate-difficulty (B2) wind speeds were actually designed to test a hypothesis about pilots' mental risk models during processing of a wind display showing two components.[9]

We had first figured that each pilot would first more or less normalize the headwind and crosswind components—meaning that whatever actual number they saw for a component on-screen, they would first mentally translate that into an internal "fear factor" based either on events having happened to them first-hand, or second-hand events (real or fictional) having happened to others, or perhaps on notions of risk gleaned from exposure to statistical information concerning accident frequencies and rates.

Second, we surmised that pilots would mentally rely upon one of three basic risk-model types.

1. A "conservative" one-factor model	Pilot would pick the *worse* of either crosswind or headwind
2. A "liberal" one-factor model	Pilot would pick the *better* of either crosswind or headwind
3. A two-factor model	Pilot would factor-weight-and-sum the crosswind and headwind

Here, the "conservative" one-factor model would result in the higher-than-nominal scores, because a pilot would have two risk values to choose from, and would always choose the larger.

To try to tease out which model was operating, we structured scenario difficulties according to this template:

1. "Easy" (B1) scenarios always showed both headwinds and crosswinds as easy (with mean objective landing difficulty = 20).
2. "Moderate" (B2) scenarios always showed an Easy headwind and a Hard crosswind (with mean OLD = 50).
3. "Hard" (B3) scenarios always showed both a Hard headwind and a Hard crosswind (with mean OLD = 90).

In this manner, B2 could function as a discriminator. If pilots tended to greatly overrate their B2 PLD scores, that would support Model 1. Severe underrating would support Model 2. And, scores close to nominal would support Model 3.

[9] Keep in mind that we presented no tailwinds, only headwinds. Nonetheless, there can only be *either* a tailwind or headwind—not both—so, the risk model would still always be two-component, even if we had presented tailwinds.

APPENDIX C

Instructions Given to Pilots

THANKS

The National Center for Atmospheric Research (NCAR) and the FAA's Civil Aerospace Medical Institute (CAMI) want to take this opportunity to thank you for agreeing to participate in this study. Without the help of pilots like you, we couldn't do research like this.

BACKGROUND

This study is generally about how pilots gather weather information just prior to landing. Specifically, we're studying **low-level winds** today. Landings are always challenging and, of course, any significant wind going over the approach and runway makes them even harder.

As you know, various aspects of winds such as speed, direction, variability, and trend normally present a particular challenge just at landing.

This is the kind of information you want to get just prior to landing, and your goal, naturally, is to create a **mental picture** in your head of these winds and how to deal with them.

Today, we're going to test several different presentations on a mobile device to see how well they help you **create** these mental pictures. This is **not** a test in the regular sense. There's no "pass" or "fail," you won't be graded, and nothing goes into your Airman Record. So relax and enjoy yourself. This is a science experiment, and the goal is only to figure out easier, faster, and better ways to present wind info in the cockpit just before landing.

Today we'll ask you to **imagine you're flying the small GA aircraft you fly most**.

Then, we'll show you 18 brief landing scenarios, one at a time. For some of these scenarios, you'll be shown one type of weather-information presentation and, for others, a different kind of presentation. Both will be on an iPad (and we'll spend plenty of time showing you how to make that work).

Your job will be to gather weather info and make 4 short decisions about landing.

1. Given the weather report you'll see,
    a) how hard would you **expect** the landing to be for you?
    b) what's the **easiest** it might be?
    c) what's the **hardest** it might be?
2. Would you land at that airport, go around, or divert to an alternate?

In Question 1a, you'll give us your best guess about how difficult this landing would be for you personally, in your usual GA aircraft.

In 1b and c, you'll tell us how easy—and then how difficult—it might be, given your experience of how weather sometimes changes during the last 10 minutes of approach.

In Question 2, you'll say whether you'd normally land, go around, or divert, given the report you see.

Those are the basics. This thing is easy. We'll go over it all in a little more detail in just a minute. Then, we'll have a nice, thorough practice session before starting.

INSTRUCTIONS

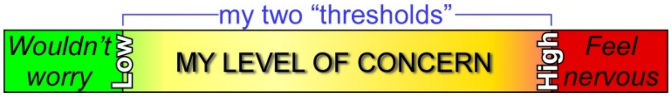
1. First fill out your information on the **Demographic Worksheet** (and this is where you get your Participant ID number, so remember that number for Step 2).
2. On the tablet, fill out your **Test Information**, including your Participant ID number and your personal “threshold” minimums and maximums for runway-relative wind components.

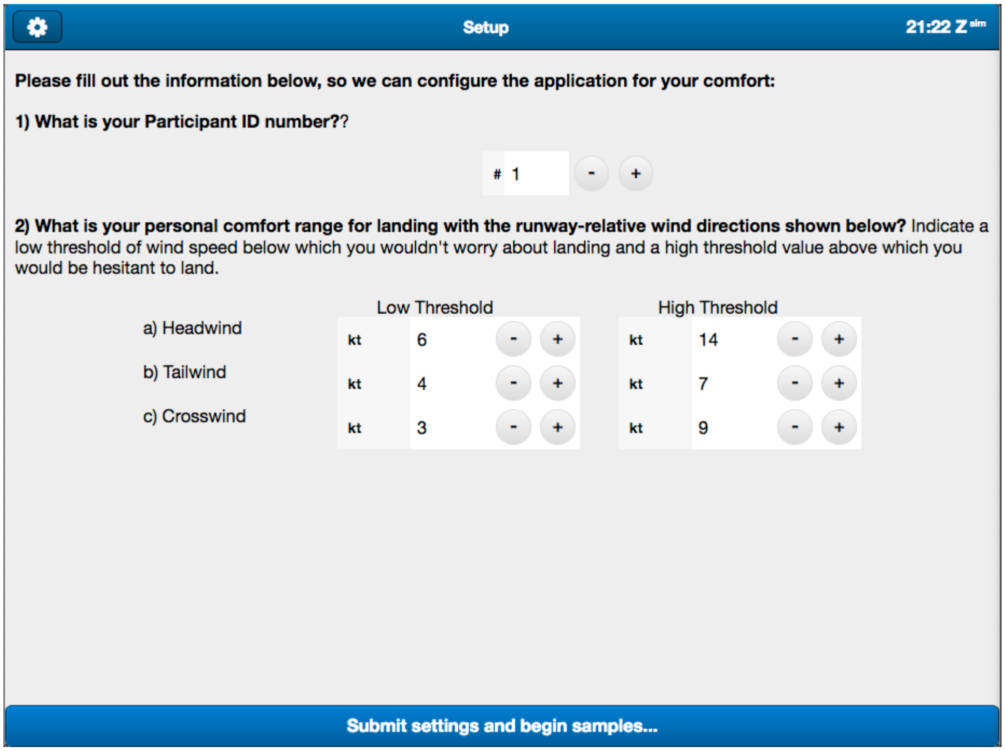
   •Remember, *these numbers apply to the GA aircraft you fly the most*.

   •“Low Threshold” means “below that speed = I wouldn’t worry about landing with that wind component.”

   •“High Threshold” means “above that speed = I’d feel nervous about landing with that wind component.”

   •What we’re really making is 3 range scales that look like this, → one for each wind component





Please fill out the information below, so we can configure the application for your comfort:

1) What is your Participant ID number??

1

2) What is your personal comfort range for landing with the runway-relative wind directions shown below? Indicate a low threshold of wind speed below which you wouldn't worry about landing and a high threshold value above which you would be hesitant to land.

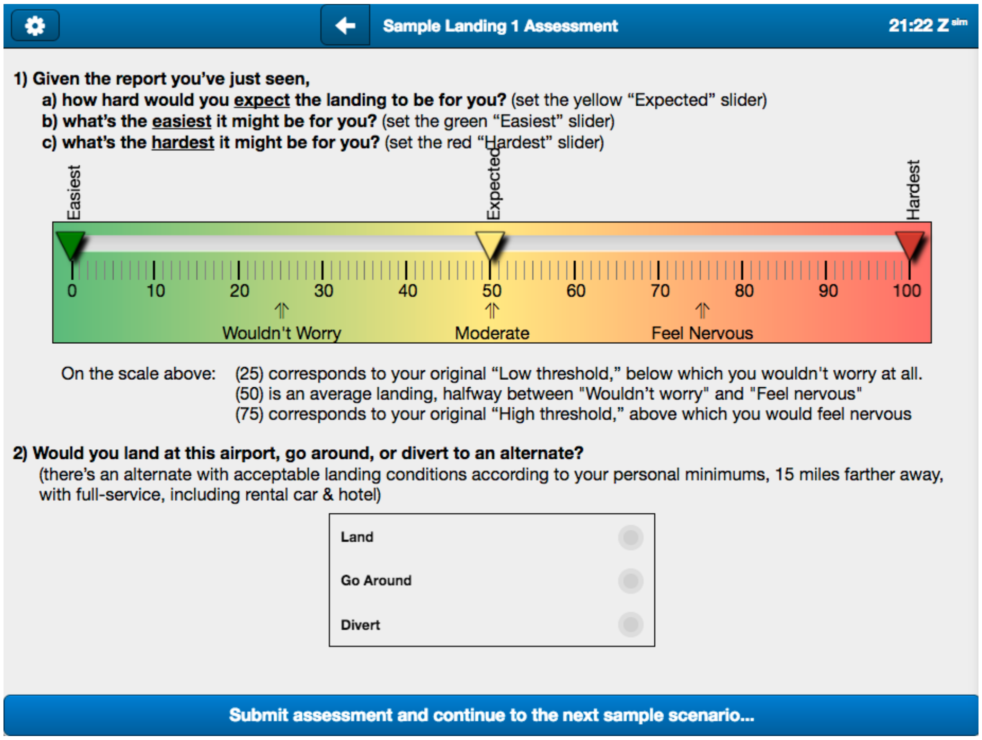
	Low Threshold	High Threshold
a) Headwind	kt 6	kt 14
b) Tailwind	kt 4	kt 7
c) Crosswind	kt 3	kt 9

   •The app uses this to create custom test scenarios made specially for you.

3. Look at the **Sample Sheet**, which has examples of each type of scenario you will encounter. You may keep the Sample Sheet to refer to at any time during the study.
4. There’ll be **18 scenarios**.
5. For 9 you’ll use one kind of weather app on your iPad, for the other 9, you’ll use a second kind.
6. We expect each scenario to take 2-5 minutes, but no hurry. Take as much time as you need.
7. We’ll have a good practice session beforehand, so you can get comfortable with the setup.
8. Here’s the **setup**:
   1. You’re flying **the small aircraft you fly most often**.
   2. **No time pressure whatsoever**.

3. You’re approaching an **untowered airport**, **15 minutes out**.
4. **Dry, concrete** runway**, 100 wide, 8000’ long** (i.e., not a problem).
5. **ASOS** but **no LLWAS**
6. To simplify things, today, don’t worry about wind variability or trend. **Focus on wind speed and direction.**

9. Once you feel that you have a good understanding of the landing conditions, answer these 4 questions on that scenario’s “**Assess View” page.** The first 3 you do by moving the 3 sliders on the “**Expected Difficulty Scale**” (*IMPORTANT*: This scale is **NOT** “wind speed 1-100”. It’s “expected landing difficulty” 1-100, as explained on the Assess page. Very important).



1) Given the report you've just seen,
a) how hard would you expect the landing to be for you? (set the yellow “Expected” slider)
b) what's the easiest it might be for you? (set the green “Easiest” slider)
c) what's the hardest it might be for you? (set the red “Hardest” slider)

On the scale above: (25) corresponds to your original “Low threshold,” below which you wouldn't worry at all.
(50) is an average landing, halfway between "Wouldn’t worry" and "Feel nervous"
(75) corresponds to your original “High threshold,” above which you would feel nervous

2) Would you land at this airport, go around, or divert to an alternate?
(there’s an alternate with acceptable landing conditions according to your personal minimums, 15 miles farther away, with full-service, including rental car & hotel)

Land	○
Go Around	○
Divert	○

Submit assessment and continue to the next sample scenario...

Don’t worry, we’ll practice all this ↑

PRACTICE

The easiest way to understand what this study is about is to start by seeing a couple of practice scenarios. You can practice until you feel comfortable—and ask questions, too—so there’s no time pressure like with a pass/fail test. All we ask is that you go in “well-motivated,” meaning that you treat these situations with the same seriousness you’d treat an actual landing.

Once you feel comfortable with the practice sessions and give us the go-ahead to start the experiment, over the next 60-90 minutes, moving at your own easy pace, we’ll show you 18 short experimental landing scenarios, each one followed by a few brief questions. Again, these won’t be pass/fail situations, so relax. They’ll be more at seeing which weather presentations seem better for you, perhaps faster, or more accurate, or easier to understand.