Federal Aviation Administration

DOT/FAA/AM-18/11
Office of Aerospace Medicine
Washington, DC 20591

Tailoring Surface Winds Information for Mobile Meteorological Applications, Part 1: Beta-Testing

William R. Knecht[1]
Arnaud Dumont[2]

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

[2]Research Applications Laboratory
National Center for Atmospheric Research
Boulder, CO

July 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.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-18/11		

4. Title and Subtitle	5. Report Date
Tailoring Surface Winds Information for Mobile Meteorological Applications, Part 1: Beta-Testing	July 2018
	6. Performing Organization Code
7. Author(s) Knecht WR,[1] Dumont A[2]	8. Performing Organization Report No.
9. Performing Organization Name and Address [1]FAA Civil Aerospace Medical Institute, Oklahoma City, OK 73125 [2]NCAR Research Applications Laboratory, Boulder, CO 80307	10. Work Unit No. (TRAIS) 11. Contract or Grant No.
12. Sponsoring Agency name and Address 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

The purpose of this study was to learn how to help pilots land aircraft more safely and efficiently under windy conditions. While other factors such as visibility and runway conditions are obviously part of real landings, this research, expressly controlled for and tried to remove other factors from concern to let pilots focus on winds alone.

The key to safe landing under windy conditions lies in accurate and timely assessment of runway wind components (headwind/tailwind and crosswind). We can visualize pilot perception of low-level runway winds as a speed-accuracy process. Pilots are trained to assess reports such as METARs and extract wind components. As long as the necessary data are present, they seem to try to estimate those wind components to the best of their ability, even when the geometry of a given situation is complex. This study shows that they are reasonably adept at the task.

However, wind depictions vary in efficiency, some requiring more cognitive processing time than others. So, in judging the usefulness of these various wind information depictions, we cannot expect to see great differences in accuracy. Instead, given equivalent levels of accuracy, the faster depiction can be considered the more efficient.

To those ends, we tested 25 general aviation pilots on 18 runway wind scenarios, varying in crosswind and headwind/tailwind components, and runway orientation. In each scenario, pilots manipulated a tablet computer (iPad) to see a wind information page with one of four different formats for depicting the airport wind information. Two depictions were text-based, the other two, graphical.

The results of this study are clear and simple. The runway-relative, two-wind component depiction was significantly the fastest, most efficient of the four depictions tested. Moreover, this was the depiction pilots unanimously said they preferred, saying it removed the difficult task of having to mentally calculate the wind components.

17. Key Words	18. Distribution Statement
crosswind landing, runway wind, aviation weather, graphical display	Document is available to the public through the Internet: www.faa.gov/go/oamtechreports/

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

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

Tailoring Surface Winds Information for Mobile Meteorological Applications, Part 1: Beta-Testing

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

Arnaud Dumont
Research Applications Laboratory (RAL)
National Center for Atmospheric Research (NCAR)

EXECUTIVE SUMMARY

The purpose of this study was to learn how to help pilots land aircraft more safely and efficiently under windy conditions. While other factors such as visibility and runway conditions are obviously part of real landings, this research, expressly controlled for and tried to remove other factors from concern to let pilots focus on winds alone.

The key to safe landing under windy conditions lies in *accurate* and *timely* assessment of runway wind components (headwind/tailwind and crosswind). We can visualize pilot perception of low-level runway winds as a *speed-accuracy process.* Pilots are trained to assess reports such as METARs and extract wind components. As long as the necessary data are present, they seem to try to estimate those wind components to the best of their ability, even when the geometry of a given situation is complex. This study shows that they are reasonably adept at the task.

However, wind depictions vary in *efficiency*, some requiring more cognitive processing time than others. So, in judging the usefulness of these various wind information depictions, we cannot expect to see great differences in accuracy. Instead, *given equivalent levels of accuracy, the faster depiction can be considered the more efficient*.

To those ends, we tested 25 general aviation pilots on 18 runway wind scenarios, varying in crosswind and headwind/tailwind components, and runway orientation. In each scenario, pilots manipulated a tablet computer (iPad) to see a wind information page with one of four different formats for depicting the airport wind information. Two depictions were text-based, the other two, graphical.

The results of this study are clear and simple. The runway-relative, two-wind component depiction (Fig. 1d) was significantly the fastest, most efficient of the four depictions tested. Moreover, this was the depiction pilots unanimously said they preferred, saying it removed the difficult task of having to mentally calculate the wind components.

INTRODUCTION

Adverse winds remain a persistent challenge for all pilots and, therefore, a perennially high priority for the FAA (FAA, 2017). Moreover, wind is particularly problematic to the general aviation (GA) community and continues to be, per NTSB statistics, the weather element that is the highest cause of GA accidents (Fultz & Ashley, 2016, Table 2)."

Weather technology for use in the cockpit is widely seen as a leading contender for promoting maximally efficient flight while minimizing weather-related risk. That technology has now matured to the point where weather data can be efficiently delivered to the flight deck in real time. This raises a number of scientific issues that could be easily rephrased as hypotheses:

1. Does in-flight weather information actually increase GA flight safety and efficiency?
2. If so, is there a way to provide GA pilots with a low-cost method of receiving such information?
3. In any given flight situation, do textual or graphical depictions of information work better?

The current research pursues such hypotheses. Its main objective is to continue empirical testing of a low-cost, portable GA device designed to deliver timely weather information to the flight deck. This device is a mobile meteorological application that runs on a tablet computer. It is currently under development by the Research Applications Laboratory (RAL) of the National Center for Atmospheric Research (NCAR), which has also developed the Aviation Digital Data Service that served as the testbed for the National Oceanic and Atmospheric Administration's Aviation Weather Center Web site.

Testing of "mobile meteorology" (MMET) applications has been performed in multiple phases. Phase 1 evaluation began at the FAA's 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 is the focus of this report and centers on the landing phase of flight.

METHOD

Hypotheses

At a strategic level, we hypothesize benefits to displaying, on a mobile electronic device, weather information that is specifically tailored to the approach and landing phase of GA flight—as opposed to presenting information with a traditional general-purpose (encyclopedic) interaction design.

Showing benefit, of course, requires empirical support. At a tactical level, we therefore hypothesize specific benefits to presenting

1. one-minute wind-measurement information (e.g., wind speed and direction, as compared to less-frequent reports, or to only the most-recent measurement)
2. runway-relative headwind and crosswind information, as compared to the traditional north-relative airport wind information
3. low-level wind information in a graphical format, as compared to the traditional text format

To better convey a sense of what was involved, we begin by describing the experimental method and design, including the "look and feel" of the application. This will hopefully lay a foundation for the abstract concepts to follow.

Experimental Method

Hardware and Software

The application ran on a tablet computer (an iPad), and presented wind information such as shown in Figure 1.

Measuring Quality of Information Depiction

This research aimed to determine which ways of displaying (depicting) runway wind information (Fig. 1) work better than others. In order to support a claim that one depiction is "better" than another, there has to be some method of measuring display quality.

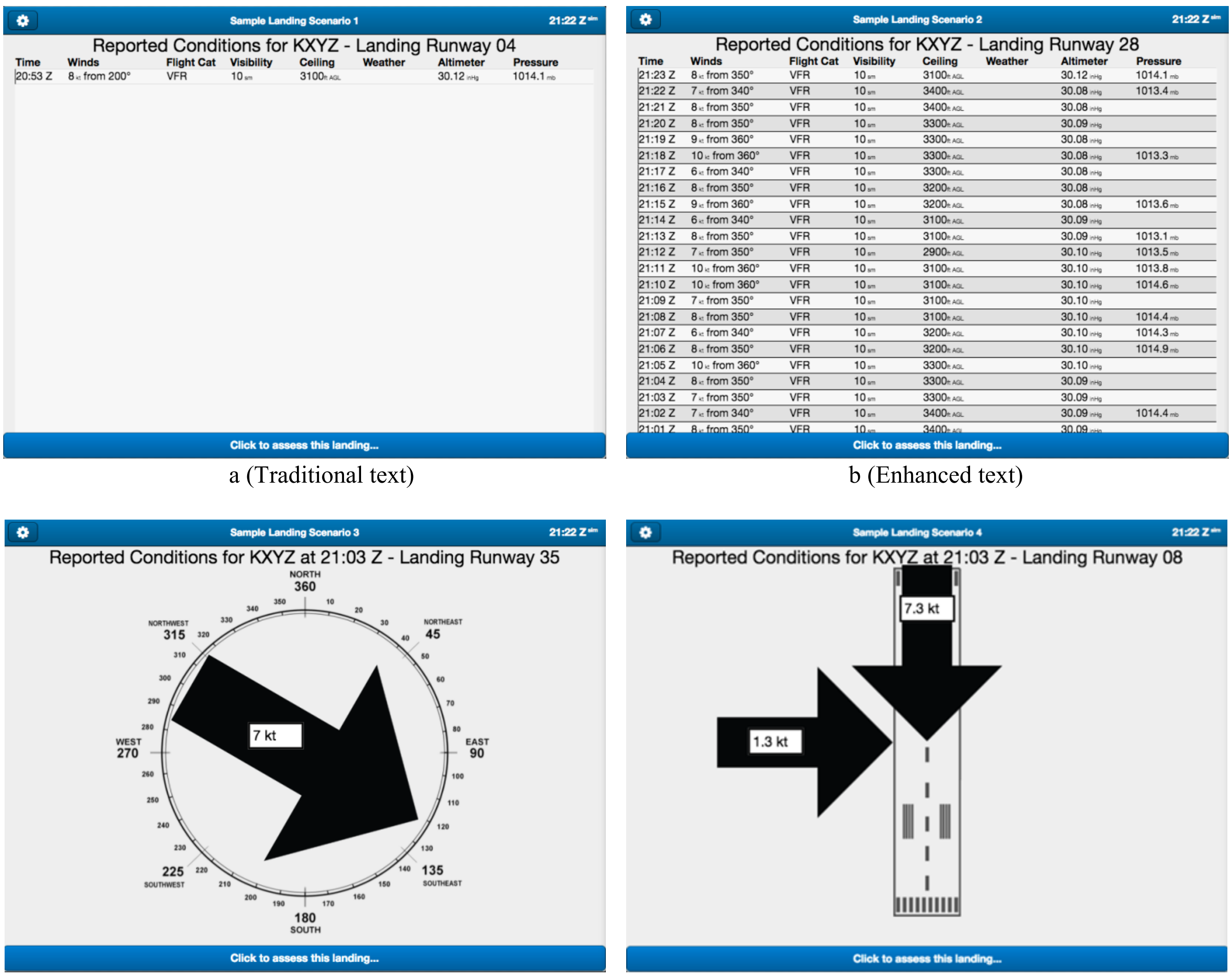
"Quality" is usually best measured by user performance on logical, quantifiable metric(s). For our purposes, the obvious metrics to measure were *accuracy* and *speed* of the pilots' mental wind-evaluation process. We also sought to develop a measure of *user confidence*—how confident a pilot was in his or her estimate of scenario difficulty—on the assumption that, the better the display, the more confident the user should be that the information depicted is accurate.

To summarize, the quality of each information depiction was measured by pilots'

1. Speed
2. Accuracy

3. Confidence-in-estimate (of scenario difficulty)

"Speed" is simply how much time it takes to make a decision whether or not to land, given low-level wind information such as shown on one of the pages in Figure 1. (Note that the participants were instructed to take as much time as needed to evaluate the wind information.) "Accuracy" of a display, however, is an altogether-different and harder quality to assess.



a (Traditional text) b (Enhanced text)

c (Traditional graphical) d (Enhanced graphical)

Figure 1. a) "Traditional," text-based runway wind information, similar to an aviation routine weather report (METAR). This sample is supposedly 19 minutes old; b) "Enhanced" runway wind information, showing similar information, but updated each minute, with the newest information no more than 1 minute old; c) "Traditional" graphical wind depiction, a north-up view with an arrow showing wind direction and textual depiction of speed; and d) "Enhanced" graphical wind depiction, a runway-relative view with separate arrows for crosswind and runway-aligned wind components.

To assess accuracy, we developed a way to compare "objective landing difficulty" to "perceived landing difficulty." This was based on the assumption that *the closer the **perceived difficulty** of a wind scenario was to its **objective difficulty**, the better the wind display*. This assumption was central to the statistical analysis. Figure 2 illustrates.

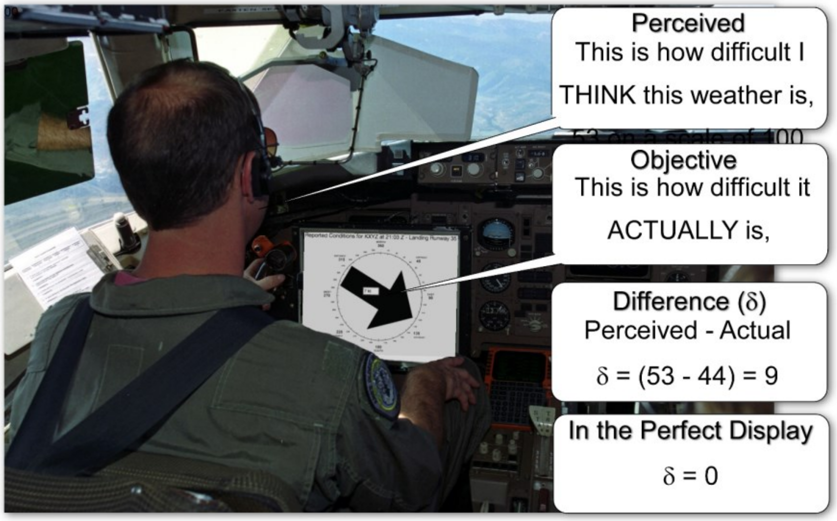


Figure 2. “Display quality” was measured as the difference δ (delta), defined as participant’s perceived scenario difficulty minus their objective scenario difficulty, both on a scale of 0—100. In a “perfect” display δ would equal zero; the display enabled them to correctly assess the scenario difficulty.

Now, defining perceived and objective difficulty depends considerably on understanding some key details of how this experiment was set up. Before presenting detailed definitions of these concepts, it may therefore be useful for us to describe the experimental design. This will establish a more concrete conceptual basis upon which we can return later and finish detailing what is meant by our central precepts of “perceived landing difficulty” versus “objective landing difficulty.”

Experimental Design

The experiment presented wind information in a within-participants (repeated measures) statistical design. Each participant/pilot saw and responded to 18 runway wind landing scenarios, each scenario displaying a single page of information similar to those in Figure 1, each with a slightly different set of wind parameters that were manipulated as independent variables.

Independent Variables (IV)

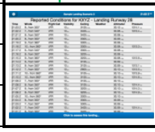
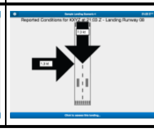
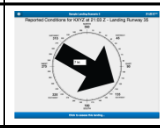
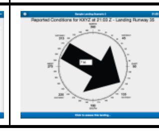
Independent variables are the factors manipulated in an experiment to see how they affect desired outcomes (the dependent variables). The current design involved three IVs in a 2 × 3 × 3 = 18 total-trial design:

1. *2 Information Display types*:
   a. Traditional (Figs. 1a, c)
   b. Enhanced (Figs. 1b, d)
2. *3 Wind-Information Scenario types* that varied:
   a. report frequency: Low (hourly reports, Fig. 1a) vs. High (1-minute reports, Fig. 1b),
   b. wind orientation: North-Up wind report (Fig. 1c) vs. Runway-Relative wind report (Fig. 1d), and
   c. depiction of winds: low-level wind Graphical vs. Textual formats.
3. *3 Scenario Difficulty levels:*
   a. Easy
   b. Moderate
   c. Hard

   These difficulty levels were individually created according to each pilot’s prior self-report of skill and risk tolerance (detailed below).

Table 1 shows specifically how the three IVs were organized and these IVs relate to Figure 1’s information depictions. See Appendix A for greater detail.

Table 1. Proposed research design, **2 x 3 x 3** (Display Type (A) x Information Type (B) x Scenario Difficulty (C))

	A1-Enhanced Display			**A2-Traditional Display**		
Scenario Difficulty[A]	B1-High-Frequency Reports	B2-Runway-Relative Orientation	B3-Graphical Wind Depiction	B1-Low-Frequency Reports	B2-North-Up Orientation	B3-Textual Wind Depiction
C1-Easy						
C2-Moderate						
C3-Difficult						

[A]All scenarios were normalized according to individual S's answers for "Low Threshold" and "High Threshold" values on their Setup screen (see text for details).

The "Traditional vs. Enhanced" distinction (independent variable "A") centered on the difference between the ways wind information has been traditionally displayed (Figs. 1a, c) versus NCAR's vision of how to display it more effectively (Figs. 1b, d).

The three "Wind-Information Scenario" (IV "B") types were selected to represent three important aspects of wind information, namely, how wind reports

1. can vary in frequency (high-frequency reports should contain more information collectively, be more up-to-date than low-frequency reports, and potentially convey variability or trends);
2. can either be oriented in a) a traditional north-up direction (forcing pilots to estimate trigonometric sine and cosine wind components—already a difficult task—but now one made relative to a non-standard runway orientation—a doubly difficult task), or, b) can present runway-relative winds directly (easing the pilot's job by eliminating the mental-estimation step); and
3. can describe winds through text or graphics (graphical depictions being easier to interpret for certain kinds of information).

Having more concretely described the experimental setup, we can now return to the abstract notion of "Scenario Difficulty" (IV "C" in Table 1).

Dynamic Individualization of Scenario Difficulty

Operationalizing our experimental method required wind scenarios with various objective levels of difficulty. The problem with this is that it required controlling for each pilots' *skill* and *risk-tolerance*. For instance, if one pilot thought a 3-kt crosswind was "easy" and another pilot thought a 5-kt crosswind was easy, 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 pilot. The mathematical term for this kind of individual tailoring is called *normalization*, and its goal is to create a single "normal" scale (e.g., 0—100 "difficulty units") that can be applied to all pilots, no matter what their skill or risk tolerance. This allows us to compare them with one another statistically.

To create such a "normal scale," during the Setup page (Fig. 3) at the very beginning of each pilot's test session, we gathered information to be used in normalization, by having pilots give us their individual "thresholds" for wind-component speeds.

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 value halfway between the two extremes.

Appendix B details the exact algorithm used to construct the three levels of objective scenario difficulty for each pilot. Essentially, what it does is 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 fits the new, "normal" 0—100 scale.

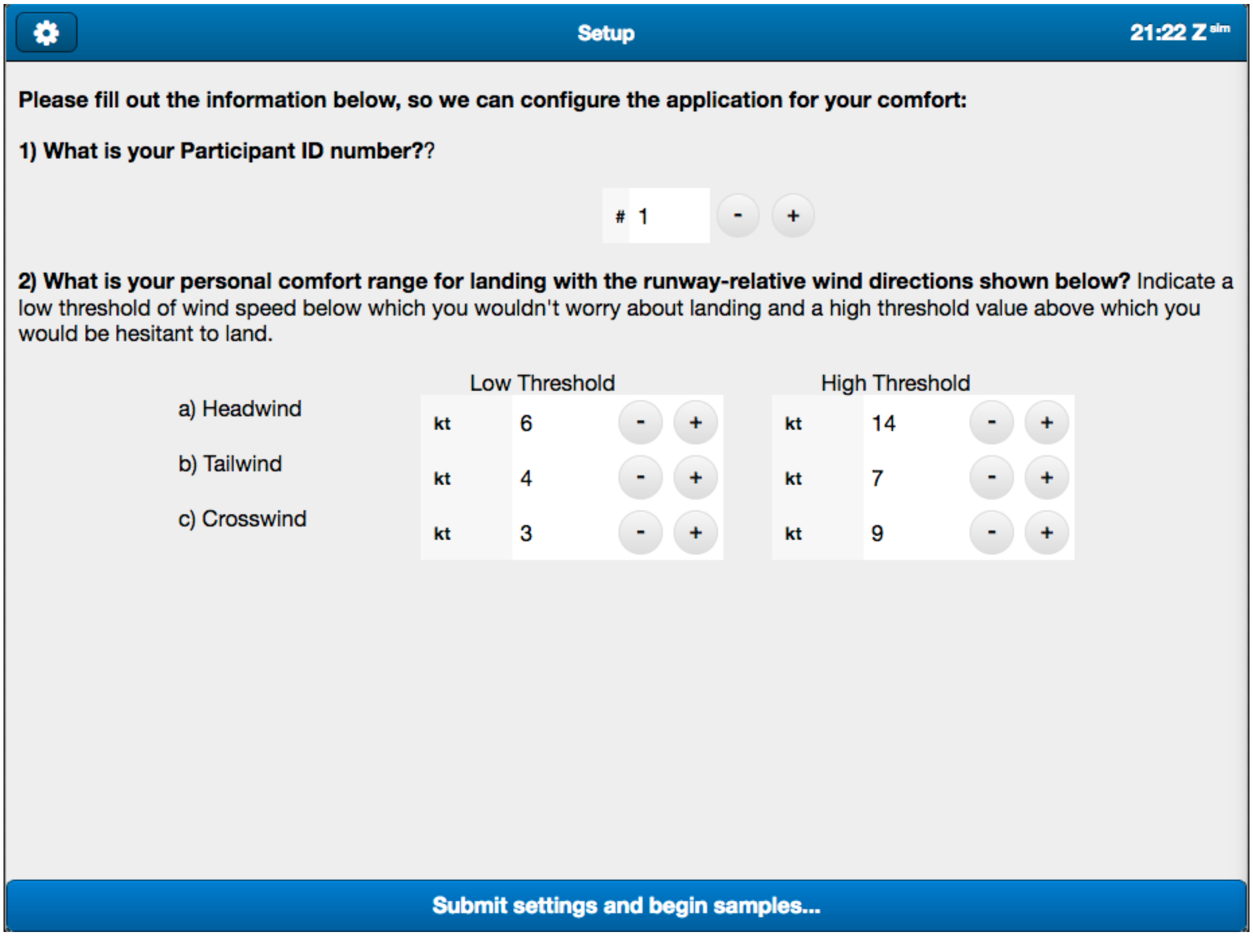


Figure 3. 1- Screenshot of the Setup page, showing the example of a "Low Headwind Threshold" of 6 kt and a "High Crosswind Threshold of 9 kt" for a hypothetical pilot.

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 δ and thereby compare one pilot to another in terms of how closely their perceptions of a given scenario's difficulty matched its objective difficulty.

Given test scenarios with known objective difficulty, the next step was to measure how well pilots could perceive, or cue into, that difficulty. This perceived scenario difficulty could now be tested for two major factors, *speed* and *accuracy*, our dependent variables supposedly influenced by our IV of objective difficulty.

Dependent Variables (DV)

Measuring *Speed* was straightforward. As previously stated, *Speed* was merely the time it took each pilot to assess the wind situation. This was defined as the elapsed time, in milliseconds, from when the wind information page (Fig. 1a, b, c, or d) was first shown to the pilot until the instant when they moved on to the subsequent assessment page.

For each scenario assessment, the pilot was asked to indicate difficulty by moving sliders along a scale. Figure 4 illustrates. These sliders showed the "normal scale" of 0—100, representing how difficult he or she *expected* the landing to be. The middle slider setting was defined as the pilot's best estimate of *Perceived Landing Difficulty,* given their own personal level of skill and risk tolerance within the context of their aircraft's capabilities.

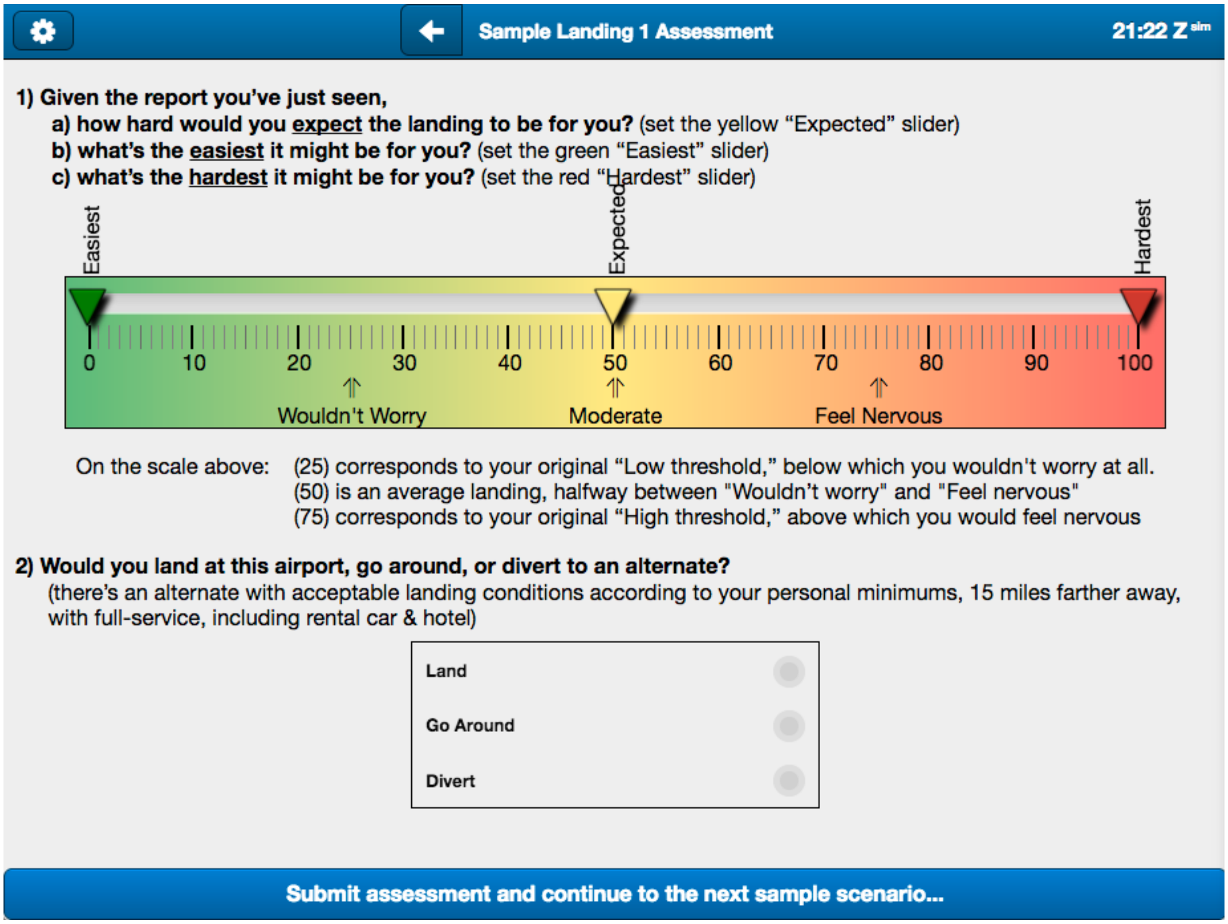


Figure 4. Screenshot of the Evaluation page.

Meanwhile, recall that each scenario's *Objective Landing Difficulty* 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 δ. And if, as hypothesized, one wind depiction was higher quality than another, we would expect most of the δ scores to be smaller.

Confidence-of-Estimate as a Third Potential Method of Measuring Display Quality

As Figure 4 shows, three slider values were collected on the assessment page. The leftmost slider represented the easiest possible landing in the pilot's estimation, while the rightmost slider represented the hardest possible landing.

This allowed for the creation of a possible third metric of display quality. If we subtracted the leftmost ("easiest") expected landing-difficulty slider value from the rightmost ("hardest") expected slider value, this would produce a Hardest-Easiest score for each scenario.

Although we were uncertain of the eventual outcome, conceptually, the magnitude of each individual *Hardest-Easiest* score might be thought of as a *confidence estimate*—a measure of how certain each pilot was about the middle *Perceived Landing Difficulty* score. High confidence in a pilot's estimate of their own judgment might result in very little difference between the lowest, middle, and highest *Difficulty* estimates. Conversely, low confidence might result in a wide spread between the lowest and highest estimate. Moreover, over the course of the entire experiment, 18 such *confidence-of-estimate* scores could result for each pilot, producing a "frequency distribution of variability scores" resembling Figure 5, whose mean and variance could itself be analyzed the same way we analyzed the middle ("expected") scores. While it is unusual to analyze the variance of a set of variability scores, there is no theoretical reason for avoiding it, as long as the score frequency distributions meet the assumptions of ANOVA.

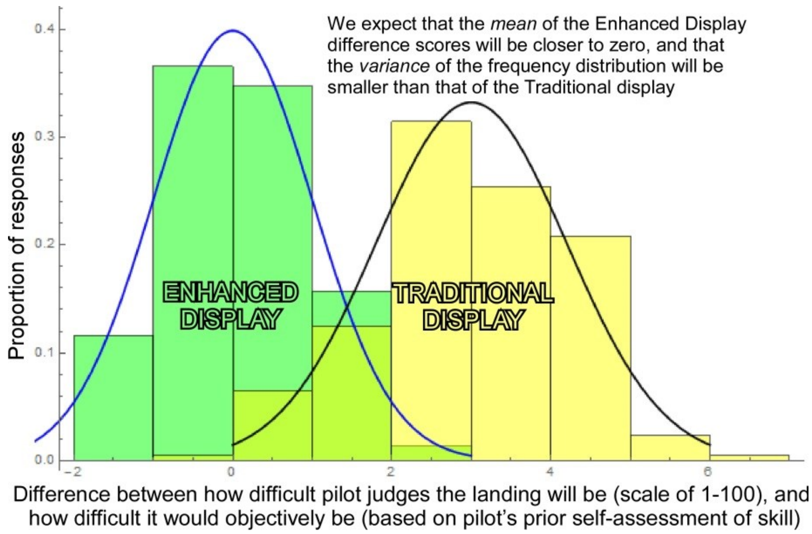


Figure 5. What a frequency distribution of Confidence-of-Estimate scores might resemble if one of the Enhanced display methods were truly of higher quality than its corresponding Traditional display.

Control for Unwanted Experimental Effects

One more 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, with far too many DV combinations to counterbalance (i.e., to present every possible scenario presentation order to an equal number of participants).

Therefore, to help 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-set D I P E O N R M K F L Q B C G A J H was presented to the next pilot.

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 four wind depiction types in Figure 1, with text descriptions of all their important features.

Pilots were then walked through four practice pages, one for each of the four depiction types, before beginning formal data collection on the main study.

RESULTS

Pilot Participants

Twenty-five general aviation pilots were recruited from a local flight school and paid $50 USD for their participation. Table 2 summarizes demographics. One pilot's data are missing due to experimenter oversight. Note that a small number of older, very high-hour pilots skewed the means, making medians more representative of the general group.

Table 2. Pilot demographics (N = 24).							
Student	0	**CFII**	5	**Age-mean**	28.8	**TFH[1]-mean**	959
Private pilot	24	**Commercial**	8	**Age-median**	21.5	**TFH-median**	185
Instrument-rated	9	**ATP**	2	**Age-SD**	14.6	**TFH-SD**	2150
CFI	7	**Multi-engine**	7	[1] Total Flight Hours			

Preliminary Examination of Data

Refer back to Table 1 for the organization of the three experimental IVs. Recall that the research design was set up as 2 × 3 × 3 = 18 treatments of (Display Type (A) × Information Type (B) × Scenario Difficulty (C).

Data Normality

A critical assumption of analysis of variance (ANOVA) is that parent data frequency distributions be normal (bell-shaped). This is all too often overlooked. Analysis of raw-scores here showed non-normalities requiring correction. Various standard methods were used. First, *Expected Difficulty* (δ) scores were treated with a one-step winsorization (where the highest and lowest score among the 25 * 18 = 450 raw scores were replaced by copies of the next-highest and next-lowest values, respectively), in order to scale back one, single outlier among the 450. Second, *Elapsed Time* scores were log-transformed ($T_{corrected} = log_{10}(T_{raw})$) to shrink the long right-hand tails in the frequency distributions. Third, one pilot's 18 scores were removed from the *Hardest-Easiest Difficulty* scores because that pilot failed to move the "Hardest" slider during four scenarios, creating outliers.

No Correction for Familywise Error

Since this is an exploratory study, no correction was made for familywise error. Familywise error is the inflation of *Type I error* (finding "significant differences" where there truly are none) due to conducting multiple analyses. 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 once, one would expect the probability of getting a black marble to be p = .05. However, if one replaced the withdrawn marble, and then repeated the entire process 100 times, the chance of getting a black marble one or more times purely by chance would increase greatly, even though the underlying ratio of black to white never changed. In other words, the more times we repeat a chance process, the more times "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 each separate analysis—its ability to detect an effect if one truly exists—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.

Analysis of Perceived Landing Difficulty

Unless otherwise specified, all analyses were tested for, and passed Mauchley's Test of Sphericity, which tests the equivalence of difference-score variances when there are three or more levels of a DV. Nonetheless, to be conservative, p-values cited will all be based on a Greenhouse-Geisser correction (which, had there been any correctable problem, would have compensated for it, statistically).

The Overall 2 × 3 × 3 Repeated-Measures ANOVA

The Data

The overall results generally support pilots having a fair level of expertise at discriminating between objectively "Easy," "Moderate," and "Hard" landings.

Table 3. Perceived Scenario Difficulty.			
Scenario Difficulty	**Perceived Mean Difficulty**	**Objective Difficulty**	**δ (Perceived-Objective)**
“Easy”	28.5	20	8.5
“Moderate”	43.2	50	-6.8
“Hard”	66.8	90	-23.2

Recall that judgment of *Perceived Landing Difficulty*, via derivation of wind components, was our primary way of getting at the relative quality of each of our four information depictions (i.e., Fig. 1). We presumed that “depiction quality” would affect pilots’ *accuracy* at determining landing difficulty. Accuracy is a key factor in psychology, often being associated with the co-factor of speed of task completion, as in the famous *speed-accuracy tradeoff* (Heitz, 2014).

The overall 2 × 3 × 3 analysis of *Perceived Landing Difficulty* scores, however, showed significance only for the three *Objective Landing Difficulty* levels. As Table 3 implies, perceived mean differences between difficulty levels were significantly distinct ($p = 1.55 * 10^{-13}$, even with Greenhouse-Geisser correction), partial $eta^2 = .763$ (meaning that the *Objective Landing Difficulty* of each scenario accounted for a relatively large proportion of the variance—or “spread”—in the data).

Pilots were generally able to at least correctly rank-order the difficulty levels no matter how the wind information was depicted. In support of that conclusion, analysis of the three individual pairwise comparisons of “Easy vs. Moderate,” “Moderate vs. Hard,” and “Easy vs. Hard” for all the individual pilot scores shows few *reversals* (e.g., an “Easy” rated harder than a “Moderate”), only 9.3% of all the 18 * 25 = 450 scores, with the 2-arrow display having the lowest percentage (2.7%). Additionally, pairwise post-hoc ANOVA comparisons indicated that each of the three objective difficulty levels was perceived by pilots significantly different from the other two at $p = .00001$ or better.

Accurate rating of scenarios’ objective difficulty was fair, as stated, although certainly not perfect. As Figure 6 shows, there was a degree of variation in the ratings.

Figure 6 also shows some skewing of the frequency distributions toward the center. This is plausibly an example of *floor* and *ceiling* effects, respectively. Here, the “floor” of Figure 6’s *Perceived Difficulty* distribution is zero. Scores are inherently free to vary *above* zero, but not below it (hence the term “floor effect”). This effectively cuts off the extreme end of any left-hand tail in a distribution. Similarly, scores are free to vary *below* 100, but not above it (hence “ceiling effect.). This cuts off the end of right-hand tails. Given such a floor and ceiling, trying to center broad, bell-shaped distributions around our stated difficulty levels of 20 and 90 (detailed in Appendix B) would *logically* truncate the left-hand side of the “Easy” distribution and the right-hand side of the “Hard” distribution, producing exactly the kind of results we see in Figure 6.

This is admittedly an arcane point. Nonetheless, it meant we needed to correct the normality of these frequency distributions before analyzing them with ANOVA.

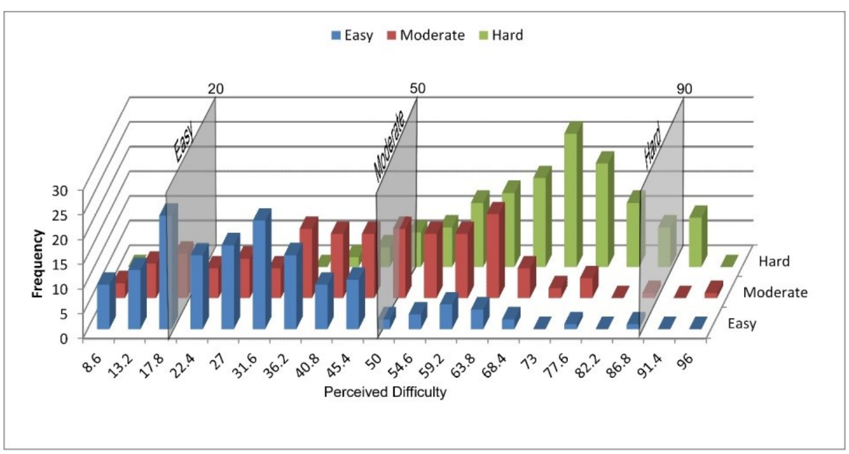
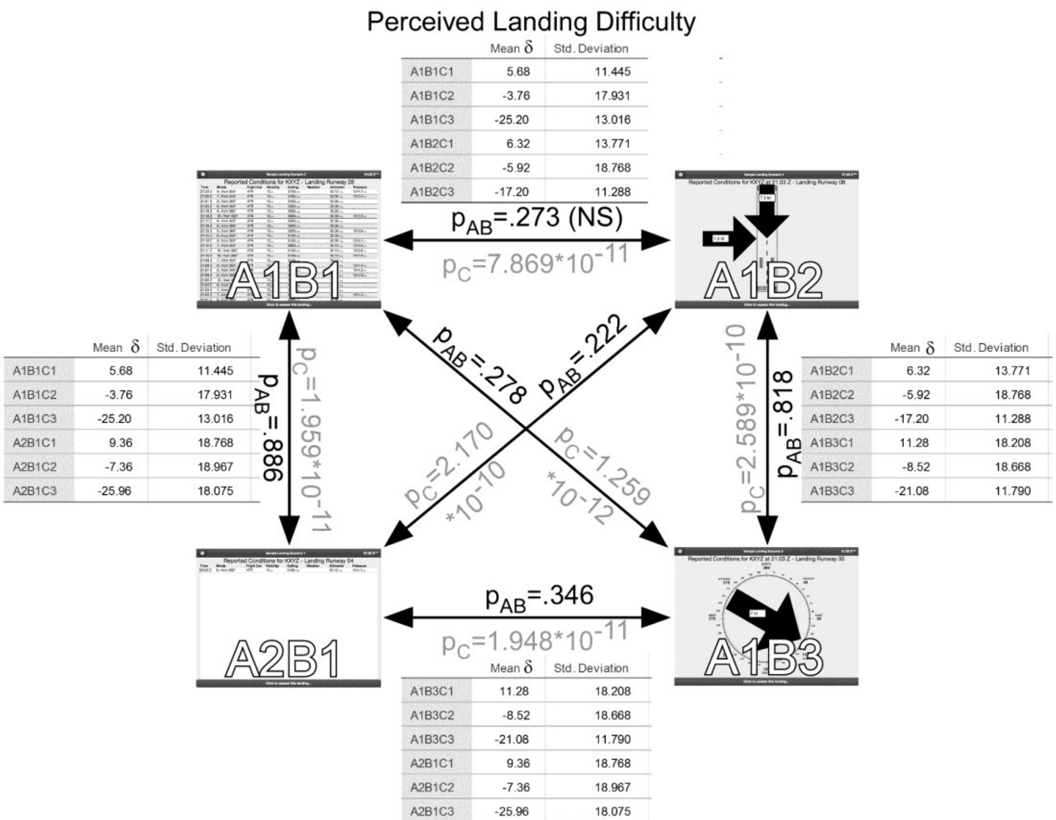


Figure 6. How pilots rated the three objective scenario difficulties.

Pairwise 2 × 3 Analysis

Referring back to our 2 × 3 × 3 setup in Table 1, one could argue that we had grouped several disparate displays into our “Enhanced” (A1) and “Traditional” (A2) groups more or less arbitrarily. This suggests reorganizing the analysis into a more direct examination of separate pairwise 2 × 3 (“AB”, depiction × difficulty) comparisons between the four major information depiction types themselves (A1B1, A1B2, A1B3, A2B1). Each comparison would include all three of the objective difficulty levels to maximize power. Figure 7 summarizes the result.



	Mean δ	Std. Deviation
A1B1C1	5.68	11.445
A1B1C2	-3.76	17.931
A1B1C3	-25.20	13.016
A1B2C1	6.32	13.771
A1B2C2	-5.92	18.768
A1B2C3	-17.20	11.288

	Mean δ	Std. Deviation
A1B1C1	5.68	11.445
A1B1C2	-3.76	17.931
A1B1C3	-25.20	13.016
A2B1C1	9.36	18.768
A2B1C2	-7.36	18.967
A2B1C3	-25.96	18.075

	Mean δ	Std. Deviation
A1B2C1	6.32	13.771
A1B2C2	-5.92	18.768
A1B2C3	-17.20	11.288
A1B3C1	11.28	18.208
A1B3C2	-8.52	18.668
A1B3C3	-21.08	11.790

	Mean δ	Std. Deviation
A1B3C1	11.28	18.208
A1B3C2	-8.52	18.668
A1B3C3	-21.08	11.790
A2B1C1	9.36	18.768
A2B1C2	-7.36	18.967
A2B1C3	-25.96	18.075

Figure 7. Statistical relations between wind-information depictions for *Perceived Landing Difficulty*. P_{AB}-values for the four basic information depiction types are shown in black. P_C-values for the three *Objective Landing Difficulty* (“C”) levels are shown in gray.

The p-values in Figure 7 imply that *pilots seem to be able to extract the wind-component information, no matter whether it is displayed textually or graphically*. This is evidenced by Figure 7’s black-colored p_{AB}-values all being greater than .05, which tells us that no type of information depiction produced significantly better judgment of landing difficulty than did any other type.

Pilots also seem reasonably good at figuring out how relatively difficult a landing scenario was, as described previously. Figure 7’s gray-colored p_C values are far, far below .05, indicating highly significant discrimination between the three objective scenario difficulty levels.

Analysis of Elapsed-Time

The Overall 2 × 3 × 3 Repeated-Measures ANOVA

So, pilots did *not* differ significantly at estimating landing difficulty when using the four different depictions of wind information. But, perhaps they differed on how much *time they spent processing the wind information* for each depiction. Recall that the DV *Elapsed Time* consisted of the time between when the wind information depiction was first presented, until the time when the pilot moved on to the assessment screen. They were instructed to take as much time as needed to review the wind information. While not a perfect proxy for cognitive processing time (e.g., daydreaming, or unintended interruption could affect it), one can argue a case for using it as such.

Pairwise 2 × 3 Analysis

Recall that *Elapsed Time* analysis was done on $\log_{10}$-transformed milliseconds. Unlike *Perceived Scenario Difficulty*, however, *Elapsed Time* showed statistical significance and considerable effect sizes for all three main IVs ($p_A = .002$, partial $eta^2 = .348$, $p_B = 1.377 * 10^{-10}$, partial $eta^2 = .655$, $p_C = .001$, partial $eta^2 = .296$). We can thus claim that differences existed between "Enhanced" and "Traditional" ("A") depictions, between the four types of depiction-pairs ("B") themselves, and between the three objective scenario difficulty levels ("C").

Similar to Figure 7, we can zoom in on pairwise relations between the four "AB" wind information depiction types. Once again, the three landing difficulties differed significantly on *Elapsed Time*, with "Hard" landings taking more time to process. Figure 8 and Table 4 summarize.

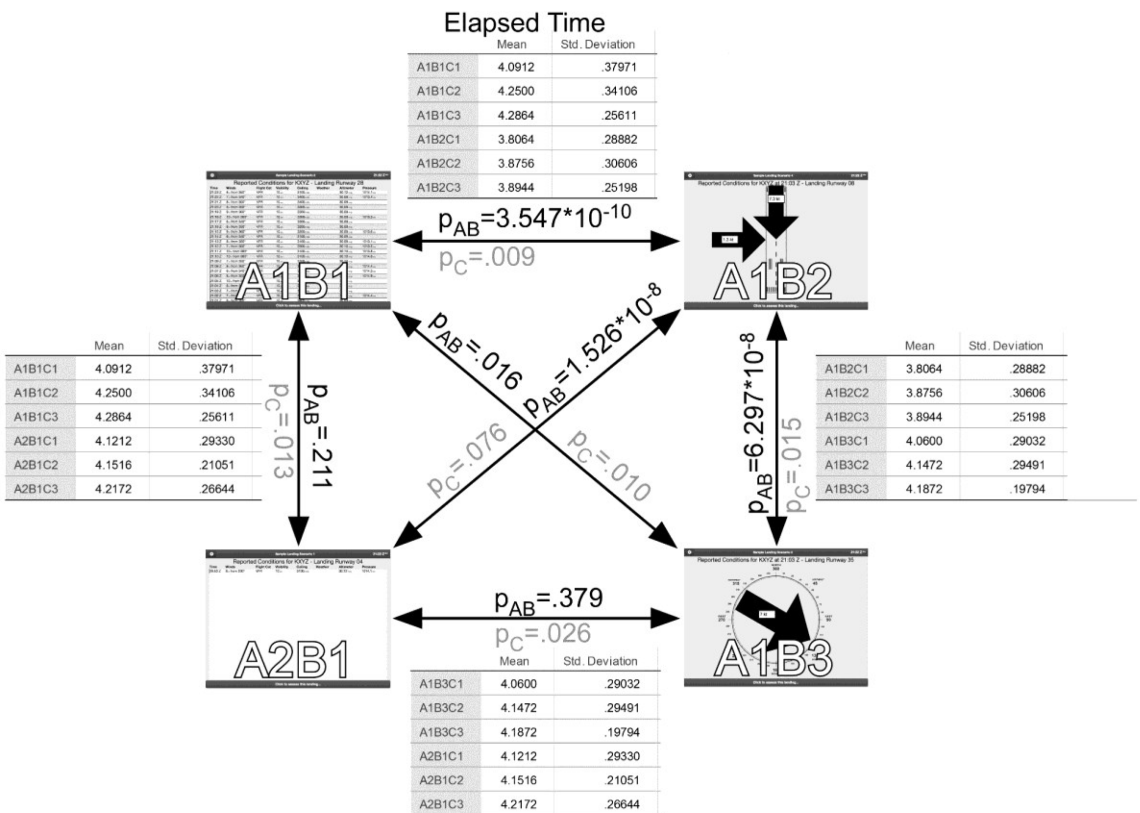


Figure 8. Statistical relations between wind-information depictions for *Elapsed Time* (log-transformed milliseconds). P_{AB}-values for the four basic ("AB") information depiction types are shown in black. P_C-values for the three *Objective Landing Difficulty* ("C") levels are shown in gray.

Table 4. Elapsed Time (seconds).		
Scenario Difficulty	**Mean ET**	**SD**
"Easy"	14.6	11.8
"Moderate"	16.4	12.3
"Hard"	18.0	12.4

Pilots processed graphical data significantly faster. Figures 8 and 9 show that the two textual depictions were the slowest, the one-arrow depiction was middling, and the two-arrow depiction was significantly fastest.

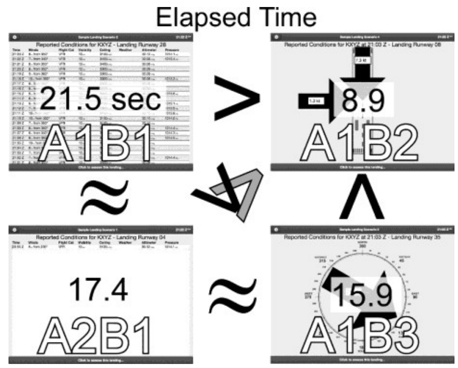


Figure 9. Greater-than (>)/approximately equal (≈)/less-than (<) relations between wind information depictions for *Elapsed Time*. The two-arrow (A1B2) depiction was the fastest of all.

Analysis of "Hardest-Easiest" Perceived Landing Difficulty

The Overall 2 × 3 × 3 Repeated-Measures ANOVA

Recall that "Hardest-Easiest" analysis was meant to serve as a measure of pilots' confidence in the various depictions' ability to convey relevant information.

Unfortunately, the overall 2 × 3 × 3 analysis told us little. Ostensibly, there was a significant effect for "A," the "Enhanced" vs. "Traditional" information depictions, but it was weak ($p_A = .020$, partial $eta^2 = .215$), with the overall mean "spread" for "Highest–Lowest" for "Enhanced" depictions being 32.7 units (on our scale of 0–100), versus 34.2 units for "Traditional" depictions, a difference of only 1.5 units out of 100.

Pairwise 2 × 3 Analysis

Moreover, pairwise analysis, similar to what was conducted for *Perceived Landing Difficulty* and *Elapsed Time*, revealed no significant differences between any of the depictions taken as individual pairs.

What this means is that only when considering all three of the "Enhanced" depictions grouped together was there any statistical discrimination apparent. But, given that these three displays looked so different from each other, we should not assign too much meaning to this individual result. It was only by combining the three "Enhanced" depictions into a single group that any "significance" arose in the first place.

DISCUSSION

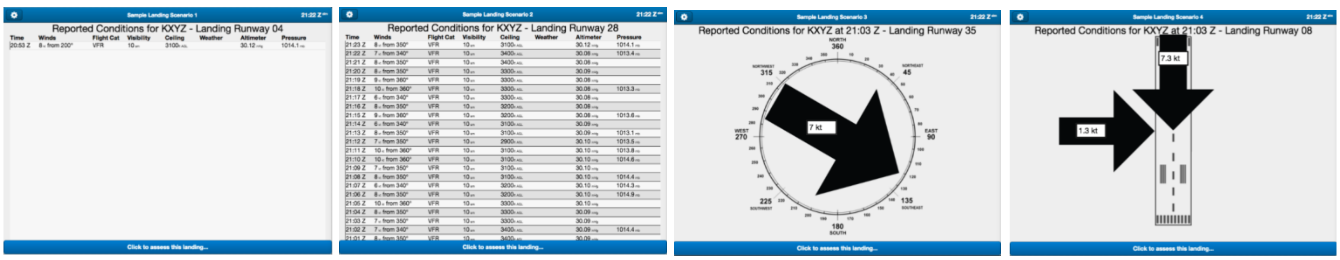
The primary purpose of this study was to identify an enhanced method for showing wind information on a mobile electronic device designed to support pilot decisions in flight. In doing so, we hoped to reduce an identified gap in the general pilot skill base, namely that calculating wind components can require 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 your 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 time of touchdown as safely possible*.

Specifically, we tested three specific hypotheses about in-flight, low-level, METAR-like wind information depictions:

1. that one-minute wind-measurement information reports would produce better pilot estimates of landing difficulty than reports containing only the most-recent measurement,
2. that providing runway-relative headwind and crosswind information would produce better pilot estimates of landing difficulty than depictions showing traditional north-relative airport winds, and
3. that graphical depictions would produce better pilot estimates of landing difficulty than traditional text depictions.

To those ends, we tested 25 general aviation pilots on 18 runway wind scenarios, varying in crosswind and headwind/tailwind components, and runway orientation. In each scenario, pilots saw a wind information page with one of four different formats for depicting the airport wind information. Two depictions were text-based, the other two graphical.



a b c d

Figure 10. 1/4-scale screenshots of the four basic wind information depictions. See Fig. 1 for a 2/3-scale version.

Our intent was to see which depictions worked better—meaning *which ones enabled pilots to make more accurate and/or faster subjective estimates about landing difficulty*. We could do this because we had engineered the objective landing difficulty of each scenario according to wind component tolerances that each individual pilot had already told us before starting the experiment. Therefore, we could calculate an error score for each scenario by subtracting *Objective Landing Difficulty* from *Perceived Landing Difficulty*. Scores closer to zero implied more accurate pilot judgment, which, in turn, implied a better wind information depiction.

Statistical analysis of these error scores showed no statistically significant differences in pilot *accuracy* over the four wind information depictions. This implied that pilots were reasonably good at discriminating between "easy," "moderate," and "hard" landings, and were just about as good at judging landing difficulty using one wind depiction type as another.

Where the four depictions differed significantly was mainly in pilots' *speed* of cognitive processing. Figure 11a summarizes. The two-arrow graphical depiction—depiction A1B2 (or Fig. 10d), which showed separate runway relative crosswind and headwind/tailwind components—was 78.7% faster than the one-arrow graphical depiction (10c), and about twice as fast as the two textual depictions (10a, b).

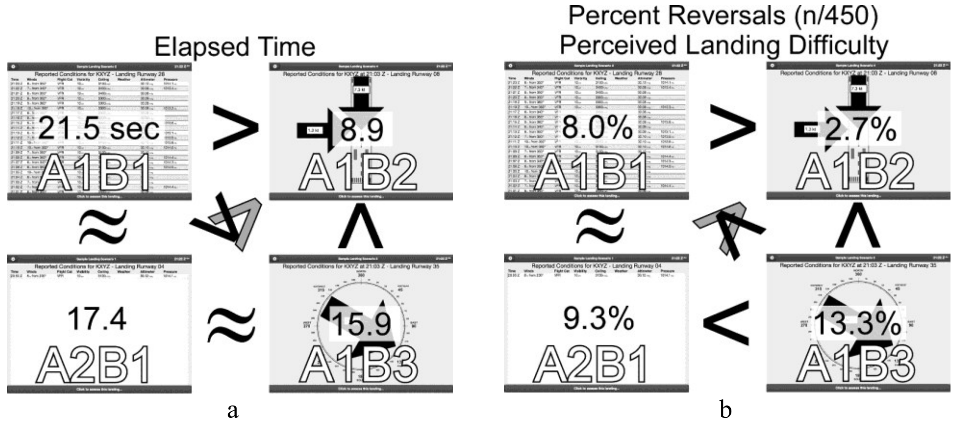


Figure 11. a) Mean Elapsed Time values for the four basic wind information depictions The two-arrow (A1B2) depiction was much faster, b) It also had a lower incidence of reversals (see text for definition).

Additionally, the two-arrow depiction showed directionality toward fewer *reversals* (Fig. 11b, cases where, for instance, an "Easy" landing was judged harder than a "Moderate" one, or vice versa). None of these directionalities was significant (p_{X^2} = .17, NS) since the overall number of reversals was small. But, as Figure 11b shows, the relations between the four depictions did unfold in a way consistent with superiority of the two-arrow depiction.

CONCLUSIONS

The Runway-Relative, Two-Wind Component Graphical Depiction is Most Efficient

The story supported by this study is clear and simple. We can visualize pilot perception of low-level 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. This study shows that they are reasonably adept at the task.

However, wind information depictions obviously vary in *efficiency*, some requiring more cognitive processing time than others. When evaluating wind information formats to be used in-flight during the approach phase, cognitive processing time is key. So, in judging the usefulness of these various wind information depictions, we cannot expect to see great differences in accuracy. Instead, *given equivalent levels of accuracy, the faster depiction can be considered the more efficient*.

Regarding the issue of whether graphical displays are generally superior to textual displays, we cannot categorically conclude that *both* the graphical depictions tested here were more efficient than *both* textual depictions. The issue is slightly more complicated than that. As Figure 11a showed, the one-arrow graphical (A1B3) depictions were slightly faster than the one-line textual (A2B1) depictions, but not significantly so.

What is unequivocal, however, is that, of the four types of low-level runway wind information depictions tested, the graphical runway-relative, two-wind component depiction (Fig. 10d) was significantly faster/more efficient than the other three.

Pilot Preference

Finally, we asked pilots which depiction they liked most. As one might expect, they preferred the two-arrow graphical depiction, saying it removed the difficult task of having to mentally calculate the wind components. Notably, this was a unanimous response, rarely seen in preference polling.

Half a dozen pilots mentioned the "minutely" (A1B1) depiction, not because it was easy to use, but because it presented the opportunity for trend and variability information. We therefore recommend a follow-up study wherein some attempt is made to graphically represent wind trend and variability information.

ACKNOWLEDGMENTS

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

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).

Federal Aviation Administration. (2017). FAQ: Weather delay. Retrieved May 22, 2017 from https://www.faa.gov/c/nextgen/delivering/nextgen_weather/faq_weather_delay.html.

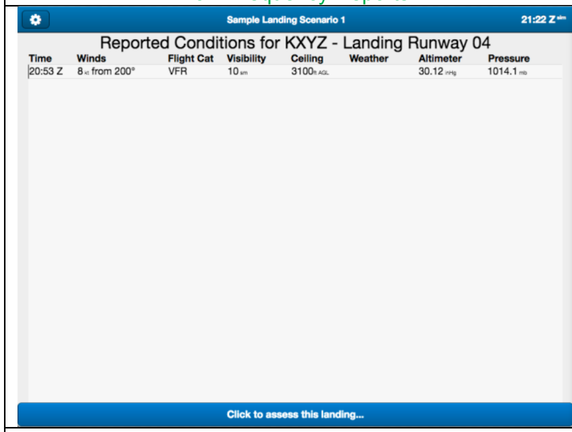
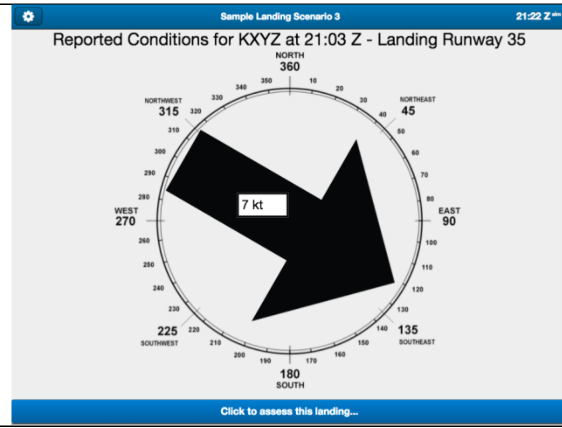
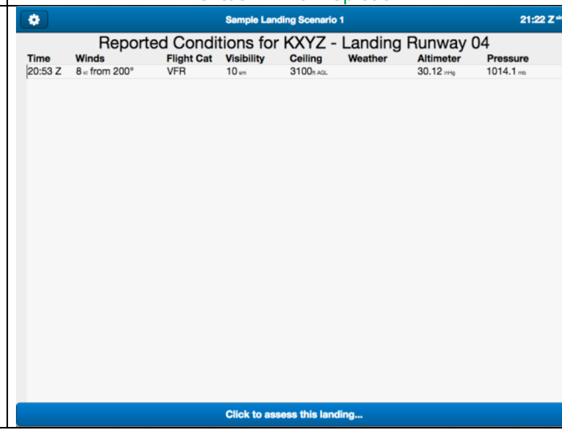
Fultz, A.J., & Ashley, W.S. (2016). Fatal weather-related general aviation accidents in the United States. *Physical Geography,37*(5), 291—312.

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/.

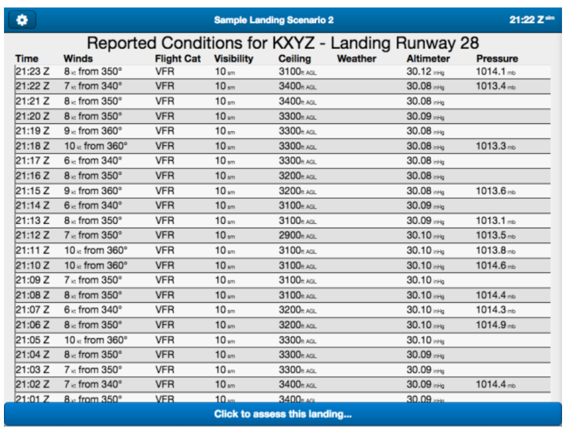
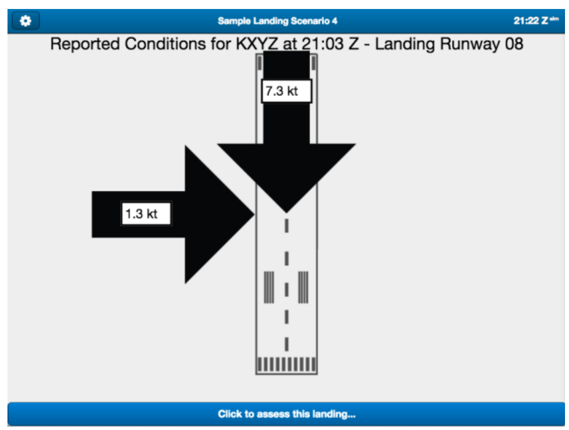
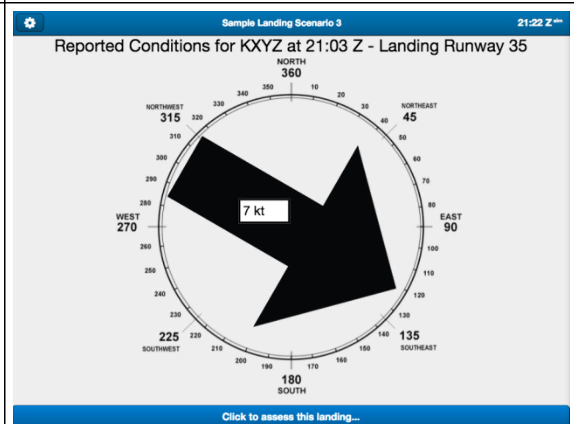
APPENDIX A

Screenshots of research design. Each of these 2-IV combinations would have 3 difficulty levels, to create a total of 18 scenarios.

Traditional Display

Low-Frequency Reports	North Wind Orientation	Textual Wind Depiction
		

Enhanced Display

High-Frequency Reports	Runway-Relative Orientation	Graphical Wind Depiction
		

APPENDIX B
Wind Scenario Calculation

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.

The runway orientation was also 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 and the north (360°), east (90°), south (180°), and west (270°) directions were 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.

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 never used because the combination of tailwind and crosswind was deemed to introduce too much uncertainty into the pilot's decision-making despite the sufficiently long runway (8000') described to participants. Instead, a constant slight headwind was used for every scenario. The 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.

Crosswind was set according to the difficulty of the scenario. For "easy" scenarios, the crosswind speed was set at 20% below the participant's low crosswind threshold. For "difficult" scenarios, crosswind speed was set at 20% above the high crosswind threshold. For "moderate" scenarios, crosswind speed was set as the mean of the low and high thresholds. 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 $.8v \le s_t \le 1.2v$.

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.

Calculation of Objective Scenario Difficulty

Given the way scenario winds were constructed as "Easy," "Moderate," or "Difficult," objective scenario difficulty (a.k.a., *Objective Difficulty*) was determined on our 0—100 difficulty scale as follows:

1. "Easy" scenarios were rated difficulty 20.
2. "Moderate" scenarios were rated difficulty 50.
3. "Hard" scenarios were rated difficulty 90.

This resulted from the process of normalizing each pilot's personal wind component Low and High Thresholds to our 0—100 scale (see Fig. 4), where a value of 25 corresponded to the Low Threshold and 75 to the High Threshold. Since "Easy" scenarios were constructed to be 80% of the Low Threshold, then, on the 0—100 scale, the equivalent was .80 * 25 = 20. Similarly, "Hard" scenarios were constructed to be 120% of the High Threshold, therefore 1.20 * 75 = 90 on the 0—100 scale. "Moderate" scenarios were simply halfway between Low and High, therefore halfway between 25 and 75, being 50.

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

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 <u>how difficult this landing would be for you personally, in your usual GA aircraft</u>.

In 1b and c, you'll tell us how easy—and then how difficult—it <u>might</u> 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

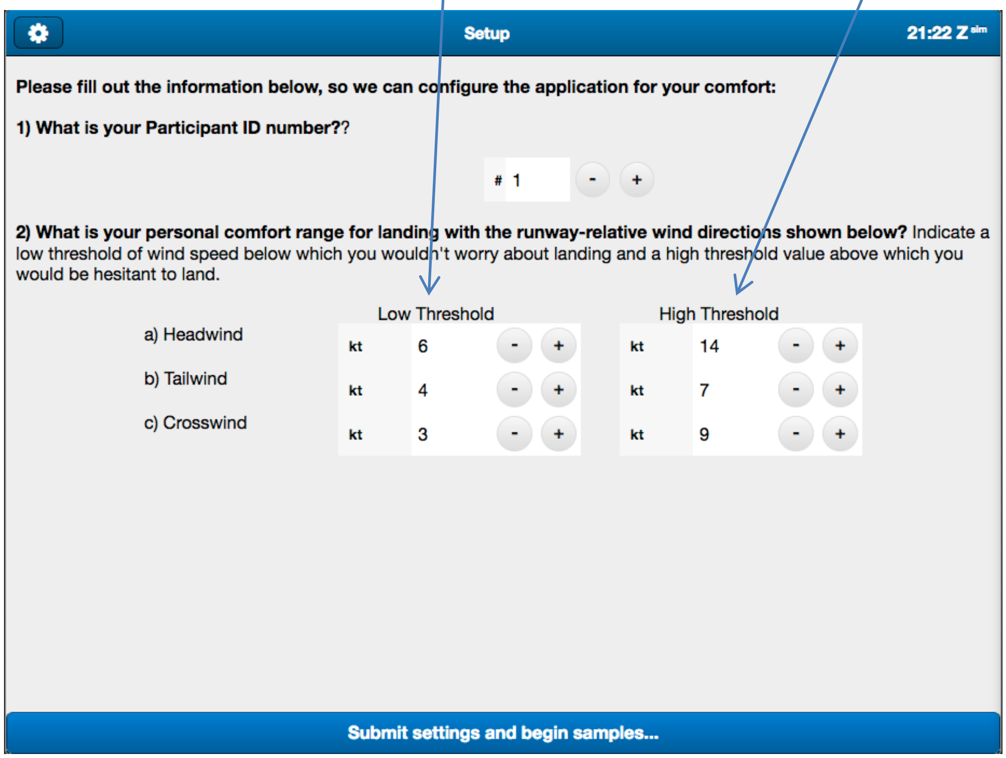
1. First fill out your information on the **Demographic Worksheet** (and this is where you get your Participant ID number, so remember that number for Step 2).
2. On the tablet, fill out your **Test Information**, including your Participant ID number and your personal “threshold” minimums and maximums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

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

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

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

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

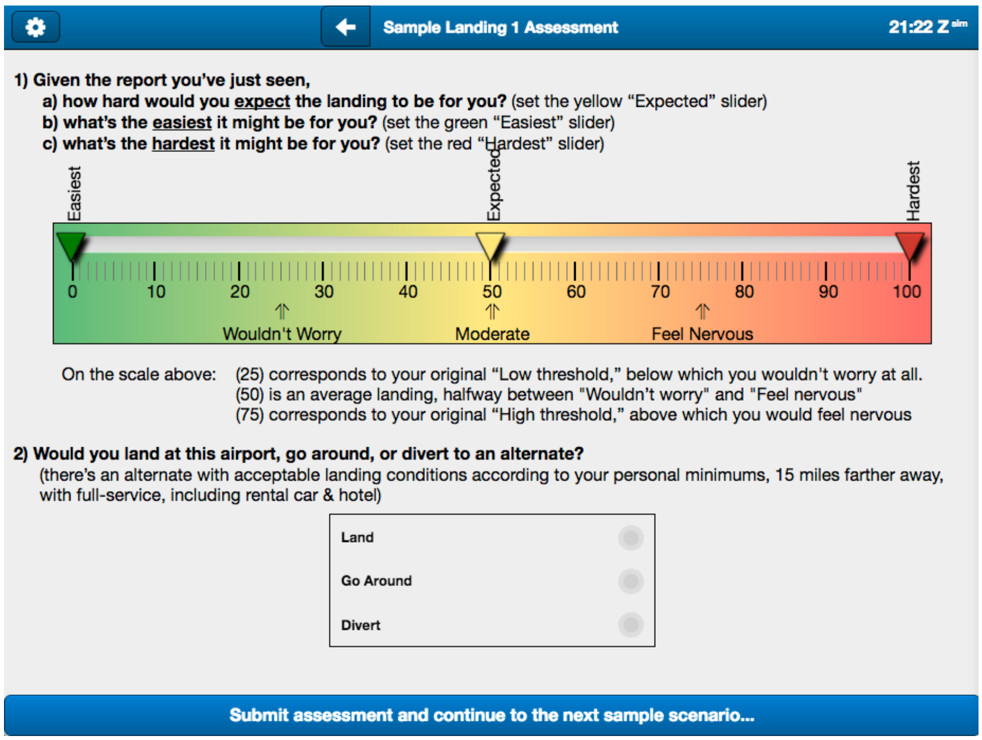


   •The app uses this to create custom test scenarios made specially for you.

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).



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.