Federal Aviation Administration

DOT/FAA/AM-15/9
Office of Aerospace Medicine
Washington, DC 20591

Pilots' Risk Perception and Risk Tolerance Using Graphical Risk-Proxy Gradients

William R. Knecht
Civil Aerospace Medical Institute
Federal Aviation Administration
Oklahoma City, OK 73125

Eldridge Frazier
Office of Advanced Concepts &
Technology Development
Federal Aviation Administration
Washington, DC 20024

May 2015

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 Federal Aviation Administration website.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-15/9		

4. Title and Subtitle	5. Report Date
Pilots' Risk Perception and Risk Tolerance Using Graphical Risk-Proxy Gradients	May 2015
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Knecht WR,[1] Frazier E[2]	

9. Performing Organization Name and Address		10. Work Unit No. (TRAIS)
[1]FAA Civil Aerospace Medical Institute P.O. Box 25082 Oklahoma City, OK 73125	[2]FAA Office of Advanced Concepts & Technology Development 800 Independence Ave., SW Washington, D.C. 20024	11. Contract or Grant No.

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

15. Supplemental Notes

Work was accomplished under approved task AHRR521

16. Abstract

Graphical weather displays such as Next-Generation Radar (NEXRAD) radar reflectivity maps are now extensively being used by general aviation (GA) pilots. Human factors issues associated with such risk-proxy displays are of great interest to researchers, aviation policymakers, manufacturers, and aircraft insurers alike.

To that end, this study is a simple, three-page test of risk tolerance. With risk defined as the chance of "significant damage to your aircraft," and motivation as "fuel cost combined with time pressure," three graphical NEXRAD-like risk gradients were created, each with a different starting value, and logarithmically color-coded with eight different levels of risk posed by potential weather. Each risk gradient was given two different motivation levels. The study utilized 30 GA pilots to draw six flight paths from a departure point to a destination point and estimated each pilot's risk tolerance for each flight, based on flight path length (an efficiency measure) and the highest-risk area traversed (a safety measure).

Three major quantitative findings emerged. First, higher motivation generally led to shorter flight paths, but at the cost of higher risk. Second, in more than half the flights tested here, pilots appeared to exhibit risk tolerances in excess of formal national policy goals. Third, however, the numerical risk values themselves appeared confusing to many pilots.

All three of these findings could be effectively and easily addressed by training.

This study explores plausible theoretical explanations for these findings, including pilots' use of risk heuristics—simplifying mental rules, which substitute for complex mental calculations. Some of these heuristics could benefit from training. The remainder need only be "tuned" to meet policy goals. Finally, the study recommends that the color schemes in flightdeck displays be kept simple and consistent with color schemes pilots already know.

17. Key Words	18. Distribution Statement
General Aviation, Cockpit Displays, Weather Displays, NEXRAD, Risk Perception, Risk Tolerance, Training, Heuristics	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	33	

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

ACKNOWLEDGMENTS

This study was sponsored by FAA ANG-C61, Weather Technology in the Cockpit Program, and supported through the FAA NextGen Human Factors Research and Engineering Division, ANG-C1. Special thanks go to Dr. Jason Rogers, AAM-510, for brainstorming and for help with data collection, to Gena Drechsler, AAM-510, for help with scoring, and to David McClurkin, Chief Flight Instructor at the Department of Aviation, University of Oklahoma, without whose assistance and counsel this study would have been far more difficult. Finally, special thanks to Mike Wayda, AAM-400, for his thoughtful editing and technical support.

Contents

Pilots' Risk Perception and Risk Tolerance
Using Graphical Risk-Proxy Gradients

EXECUTIVE SUMMARY

Graphical weather displays such as Next-Generation Radar (NEXRAD) radar reflectivity maps are now extensively being used by general aviation (GA) pilots. Human factors issues associated with such risk-proxy displays are of great interest to researchers, aviation policymakers, manufacturers, and aircraft insurers alike.

To that end, this study is a simple, three-page test of risk tolerance. With *risk* defined as the chance of "significant damage to your aircraft," and *motivation* as "fuel cost combined with time pressure," three graphical NEXRAD-like risk gradients were created, each with a different starting value, and logarithmically color-coded with eight different levels of risk posed by potential weather. Each risk gradient was given two different motivation levels. The study utilized 30 GA pilots to draw six flight paths from a departure point to a destination point and estimated each pilot's risk tolerance for each flight, based on flight path length (an efficiency measure) and the highest-risk area traversed (a safety measure).

Three major quantitative findings emerged. First, higher motivation generally led to shorter flight paths, but at the cost of higher risk. Second, in more than half the flights tested here, pilots appeared to exhibit risk tolerances in excess of formal national policy goals. Third, however, the numerical risk values themselves appeared confusing to many pilots.

All three of these findings could be effectively and easily addressed by training.

This study explores plausible theoretical explanations for these findings, including pilots' use of risk *heuristics*—simplifying mental rules, which substitute for complex mental calculations. Some of these heuristics could benefit from training. The remainder need only be "tuned" to meet policy goals. Finally, the study recommends that the color schemes in flightdeck displays be kept simple and consistent with color schemes pilots already know.

Pilots' Risk Perception and Risk Tolerance Using Graphical Risk-Proxy Gradients

INTRODUCTION

Motivation for the Current Study

Inclement weather is hazardous for general aviation (GA) flight. Estimates vary, but weather is judged to be at least a secondary factor in about 20% of GA accidents (FAA/ASIAS, 2010). Historically, a disproportionately high number (≈50-78%) of weather-related GA accidents prove fatal (Batt & O'Hare, 2005; NTSB, 2005). Lately, that figure seems slightly lower; the 2011 *Nall Report* cites adverse weather as being the primary cause of 43 of 1,160 (3.7%) non-commercial, fixed-wing GA accidents in 2010 (AOPA, 2011). The GA fatal accident rate has flattened over the past six years, with 259 fatal accidents in 2013, at a cost of 449 lives. Nonetheless, improving general aviation safety remains a top priority for the FAA and industry, and they are working together to raise awareness to prevent weather related accidents.[1]

Cockpit Weather Information (CWI)

CWI is widely seen as a leading contender for mitigating weather-related risk. This makes CWI a priority research topic with stakeholders, including FAA, National Weather Service, NASA, industry, and the Department of Defense. As such, the FAA Next Generation Air Transportation System (NextGen) Implementation Plan specifically discusses "up-to-date weather and airspace status information delivered directly to the cockpit" as part of its vision for modernizing the National Airspace System (FAA, 2012, p. 8).

Research for this effort is being performed by groups such as the FAA's Weather Technology in the Cockpit (WTIC Program), the National Center for Atmospheric Research (NCAR, Steiner et al., 2010), NASA, and MIT's Lincoln Laboratory. The WTIC Program, for instance, recently completed a human factors study of probabilistic CWI and its effect on navigation through convective weather (ATSC, 2013).

The current study was inspired by that WTIC Program study. However, this study is not a study of "probabilistic weather" in the sense defined in aforementioned study ("the probability of a given *type of weather* being in a given *location* at a given *time*"). Instead, this study explores how GA pilots perceive and use *risk gradients*—the direct graphical display of *risk* at a given location at a given time, with weather information being represented as risk gradients on a graphical display.

The underlying motivation is, of course, that risk perception is likely something derived from both static and dynamic information in the graphical display itself. This could involve *direct perception,* in the Gibsonian sense of not requiring higher cognitive processing (Gibson, 1979). Or, it could involve *constructivism*, in the sense of assembling more-complex schemata from elementary percepts (von Glasersfeld, 1995). Regardless, the phenomenon of risk perception has to be based on information present in the display, and it is important to know what that specific information is.

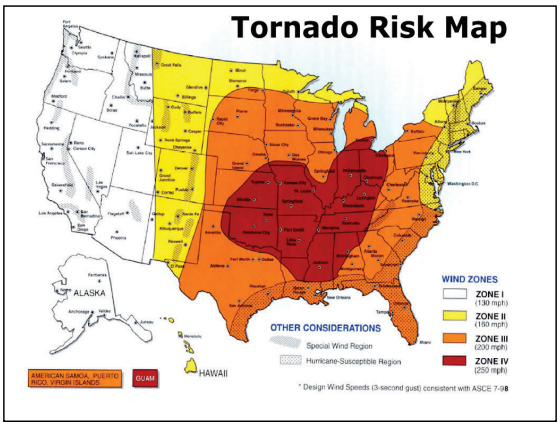


Figure 1. Map of tornado risk (NWS, 2013).

Risk and Risk Gradients

Risk is defined as the chance of a given loss or injury (Merriam-Webster, 2013). *Risk gradients* are graphical representations of risk. For instance, Figure 1 shows a color-coded National Weather Service (NWS) map of tornado risk, the chance of experiencing a tornado in various parts of the United States.

Risk gradients can be *static* (motionless, e.g., Figure 1) or *dynamic* (moving over time). Since risk itself is rather difficult to represent *directly*, it is typically represented indirectly *by proxy,* where some stimulus attribute such as color stands in for the more abstract quality of risk.

Figure 2 shows three frames of a composite NEXRAD radar reflectivity movie showing movement of a weather system across the Midwest. Each frame shows a time snapshot of radar reflectivity levels. Pilots interpret these radar reflectivity levels as a *proxy* for risk. Red is taken to be potentially more dangerous than yellow, which is taken as potentially more dangerous than green, and so forth.

Risk proxies are typically not perfect representations of risk. Instead, they are *correlates*. They *correlate with* risk, meaning they bear a statistical relation to risk, can serve as representations of risk, and can function as predictors of risk to some degree. Risk proxies are extremely important, since the accurate numerical calculation of risk can be extremely complex, and governmental entities such as the National Weather Service (NWS) are simply not in a position to calculate actual numerical risk estimates for every type of aircraft in every type of weather situation—let alone to take legal responsibility for such predictions. So, instead, they supply weather information—risk proxies—and let individual pilots estimate their own individual risk from that information.

[1]http://www.faa.gov/news/press_releases/news_story.cfm?newsId=15634

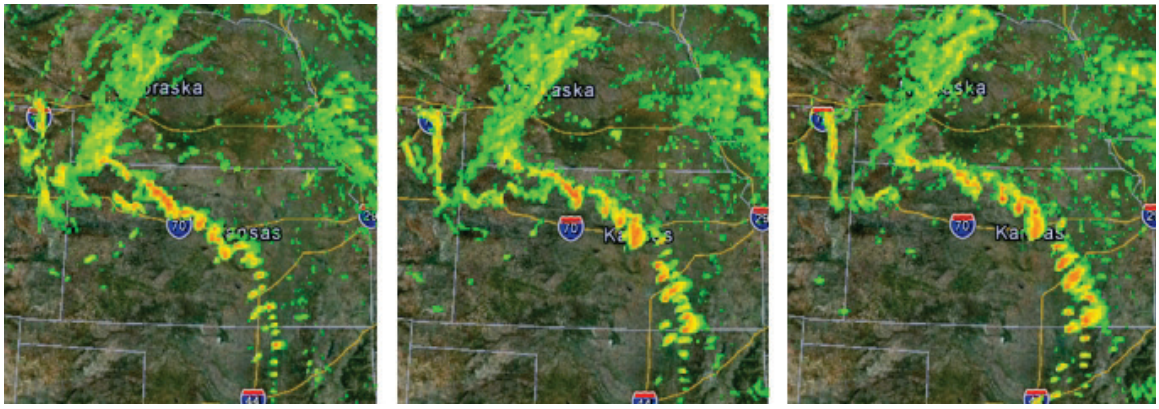


Figure 2. Three Looping NEXRAD Movie Frames

Risk Perception and Tolerance

Despite the common notion that "brains are computers," people rarely calculate risk the way a computer program might. Instead, one usually relies on risk-proxies and *heuristics*. Heuristics are simple rules, often used unconsciously. Heuristics come in different kinds, two of which are particularly relevant to this paper.

Perceptual heuristics

Many object qualities are not tangible things that can be perceived directly in the same sense as color and taste are perceived directly.[2] Instead, object qualities are often things one infers from physical stimuli that usefully correlate with a given quality. For instance, one often unconsciously uses *visual clarity* to judge how close an object is, because closer objects tend to look clearer, while distant objects look blurrier (Kahneman, Slovic, & Tversky, 1982, p. 3).

Certain kinds of risk can be accurately estimated with perceptual heuristics. A crawling infant suddenly stops at the edge of a staircase, transfixed by the famous "visual cliff," the visual stimulus of a sudden drop-off (Gibson & Walk, 1960). The drop-off is a cue to, or correlate of, an impending high-impact, potentially dangerous event. This cue is so important to survival that nature has hardwired in a heuristic to handle it (ibid., 1960). Similarly, birds such as gannets, which hunt by diving at fish close to the surface of water, seem to know exactly when to pull out of their dive by perceiving a visual cue of *time to contact* mathematically inherent to looming objects (Lee & Reddish, 1981).

Clearly, perceptual heuristics embody survival functions vital to life, and often work remarkably well, despite their relative simplicity.

Cognitive heuristics

Cognition occurs at a higher neurological level than perception and involves higher-level constructs made up of, or derived from, lower-level percepts (Papert & Harel, 1991). Again, heuristics appear intimately involved in constructionism, as we mentally construct models of the world and its characteristics. A good example of a cognitive heuristic is *satisficing* (Simon, 1955, 1990). Rather than seeking perfectly optimal solutions to complex problems, which can be tremendously time-consuming, most of us stop searching once we find a choice that "sufficiently satisfies" our selection criteria.

Many risk situations are complex but can be estimated, well-optimized, and satisficed with cognitive heuristics (Marsh, Todd, & Gigerenzer, 2004). Viewing a predictive weather display, making sense of it, and using it to minimize risk while circumnavigating weather probably falls into this category.

Therefore, the search for the perceptual and cognitive heuristics underlying those processes forms a theoretical justification for the current research.

Risk tolerance

Heuristics such as satisficing involve setting decision criteria or *thresholds*. A threshold embodies the degree of something below which or above which represents a cutoff for decision making.

If one can perceive risks directly, or correlates thereof, or if one can mentally construct risk estimates, then the *thresholds* of those qualities, below which no behavioral alteration is necessary, can be called one's *risk tolerance*.

As previously stated, many of our risk estimates are surprisingly accurate. Yet, the human mind appears particularly poor at estimating *high-impact, low-probability events* (Camerer & Kunreuther, 1989). Aviation accidents, and the events leading up to them, fall into this category, making aviation safety not only an area of great practical concern for all but one of considerable interest for decision theory, as well.

[2]It is not the intention of this paper to enter the technical debate about Gibsonian direct perception versus a perceptual heuristics approach (see Hecht, 1996). That may merely be an issue of how many layers deep we allow a network of neurons to be defined as "direct" perception, versus the number at which we define "heuristic" perception to emerge. Instead, for the sake of simplicity, we will use "heuristic perception" to temporarily subsume its direct sibling.

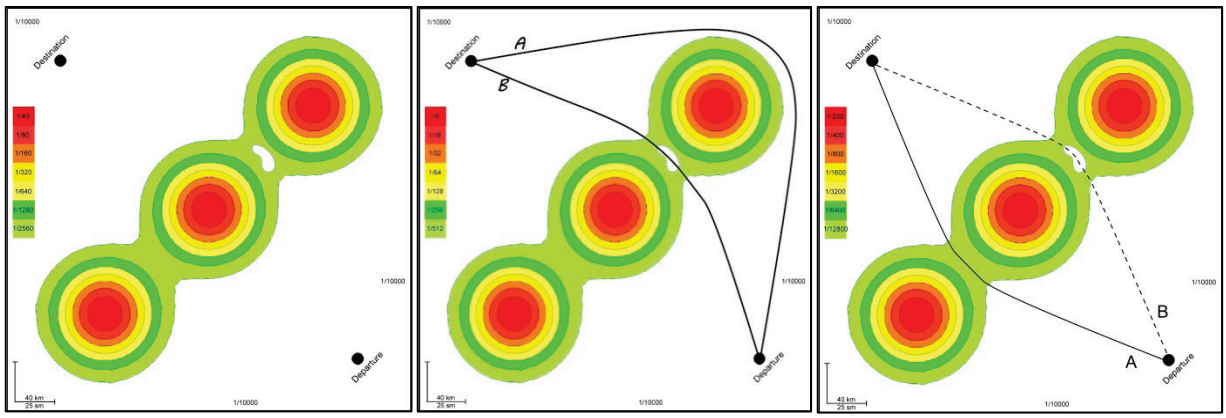


Figure 3. The a) High, b) Medium, and c) Low-risk test pages, shown here at ¼ their actual size. Each page contains three smaller risk gradients, separated by light-green "narrows" between them.

MATERIALS AND METHODS

General Approach

The current study focuses on aviation weather risk behavior. The general approach was to capture pilots' "first impressions" about risk in convective weather situations. Because convective weather is often chaotic and unpredictable, impulsive decisions based on first impressions can lead to situations difficult for pilots to later escape. By studying these first impressions, this study's goal is to discover *weather features* or *cues* (Wiggins, Azar, Hawkins, Loveday, & Newman, 2014) that serve to induce or trigger incorrect judgments, which training and/or technology could then mitigate or improve.

Test setup

Figure 3 illustrates the test pages used in the study. These are shown at ¼ their actual size (Appendix A shows a two-thirds-size version of a sample page, for greater detail).

An abbreviated scale of eight colors was used. These eight represented a subset of the NEXRAD color scheme, which is itself consistent with the latest available FAA Advisory Circular on thunderstorms (AC00-24, FAA, 2013). In our color scheme, risk effectively increased as one passed from light green toward dark red, and pilots essentially needed to avoid *orange-coded areas* by at least 20 nautical miles, to comply with FAA regulations.

As Figure 3 illustrates, the basic experimental task was for each pilot to simply draw a line between a point labeled "Departure" and another labeled "Destination." This line represented the flight path they would choose, given the particular set of conditions presented. Since this path then intersected the color-coded "weather system" (risk gradient) at various points, that gave us information about each pilot's risk tolerance in that particular flight-planning situation.

Figure 3 details how the test was constructed. The risk gradients were shown as three colored structures arranged symmetrically along an imaginary line extending from lower-left to upper-right on the test page. Note how the lightest-green color forms two narrower areas ("narrows" or "gaps") between the three separate gradients. Also note that the shortest possible path (not shown) between Departure and Destination would fall directly through a red area denoting maximum risk. We expected pilots to avoid these red areas by varying degrees, thereby giving valuable statistical information about their risk tolerance.

Statistical design

The main analysis employed a 3x2 repeated-measures statistical design, with three levels of a first independent variable (IV) and two levels of a second IV, both described below. In repeated measures, all participants see all test conditions (as opposed to between-groups designs, where different experimental groups each see only a subset of test conditions). Repeated measures has the powerful advantage of letting each participant serve as his or her "control," meaning that each person's innate, baseline preferences and tendencies (e.g., for risk-taking) will presumably be fairly stable across all scenarios, and can be statistically subtracted out by comparing each scenario's results to that individual's average scores (a.k.a., analysis of *change scores*)

Independent variables

The first main IV was *risk level*, with three levels (High, Medium, and Low), each shown on a separate page. The colors and topology (the exact shape of the figure) were identical on each page, but the odds scale for each risk gradient was different. This will be detailed momentarily, using Figures 3 and 4.

The second main IV was *motivation level*. Each page's instructions presented two different motivation levels, described in the test instructions as:

1. Suppose fuel is its normal price and you're in no hurry. Draw a line showing the shortest flight path acceptable to you from Departure to Destination. Label it "A."
2. Suppose fuel is twice as expensive and you are late to an important engagement. Draw a second line showing the shortest flight path acceptable to you. Label it "B."

A number of other secondary demographic and "hypothesis-specific" IVs were also collected for modeling purposes (explained in Appendix B). These included certificates and ratings, pilot age, total flight hours, and two self-rating scales.

Dependent variables

Several behavioral performance measures were used as dependent variables (DVs) during statistical analysis:

1. *Path length*.[3] As we can see in Figure 3, shorter flight paths generally meant greater risk tolerance because short flight routes cut through the weather, not around it.
2. M*aximum risk encountered* along each path. Most of the planned routes intersected more than one colored area. The color representing the highest value along a given route was used as the measure of risk tolerance.

Note that these DVs were consistent with the "economics of flight." This involves two factors: *safety* and *efficiency.* Pilots desire to fly safely to a destination in the best time- and most fuel-efficient route as possible. Therefore, path length provides a metric of efficiency. The shorter the route, the more efficient the flight is in terms of time and fuel. Concurrently, maximum risk encountered is an estimate of weather hazard. Therefore, it is a metric of safety.

The economics of adverse-weather flight typically dictate a cost-benefit tradeoff. Safety and efficiency are frequently negatively correlated, in that, as one increases, the other decreases. The shortest flight path through weather is usually not the safest.

Critical Methodological Issues

Minimizing experimenter bias

This type of study raises certain issues having to do with experimental procedures (i.e., methodology). One is the risk of *experimenter bias*. Expectations of the experimenter(s) may be either overtly or inadvertently transmitted to participants, through speech or even nonverbal behaviors. This may, unfortunately, bias participants to behave in ways they otherwise might not behave.

To minimize experimenter bias, standard practices were strictly followed. These included a) giving verbal instructions to the effect that "Realistic behavior is what the study is looking for," b) giving assurances of strict confidentiality of data, c) making a conscious effort to minimize nonverbal cues, and, d) generally keeping instructions as succinct as possible. This included giving pilots no special risk training for this study. They were told only what the task was, what the colors and Figures represented, and how to complete the task.

Stability of the "weather system"

One methodological issue had to do with the test itself. Pilots were instructed that, given the limitations of this paper-and-pencil test, they could assume that the "weather" (the risk gradients) would remain static for the duration of their flight. Therefore, they could trust their graphical display, and that gaps between different gradients would not close unexpectedly.

This assurance was given to ensure that the pilots would not assume (because real weather rarely remains motionless) that they ought to "build in a larger safety margin" into their flight route. Statistically, this would have resulted in increased unpredictability (variance) in the data, which would decrease statistical power (the ability to detect true IV effects where they did, indeed, exist).

Specifics

Path length measurement

Path lengths were measured with a Scalex *PlanWheel SA* electronic measuring wheel. This is a mechanical wheel that rolls along a surface, converting the number of revolutions into a measurement of length, with an accuracy of ±0.25%. One principal investigator and an assistant separately scored all paths twice (i.e., each path was scored four times). Each separate scorer's two values were first compared to eliminate errors, which were corrected by re-measurement, and then averaged into a single value. The two scorers' separate averages were then checked for interrater discrepancies. Finally, all four individual values were averaged to create the official path length for each data point. Interrater reliability of scoring will be discussed in the **Results** section.

Risk gradients

Figure 4 shows how the three risk gradients were colored to make them logically similar to an abbreviated NEXRAD composite radar reflectivity color scheme.

Each of the three color-coded scales represented one risk gradient used on one test page (e.g., Figure 4a's scale is $1/8$, $1/16$… $1/512$, used on the test page of Figure 3a). Within each risk gradient, light green therefore represented the lowest-risk "weather" on that particular scale (e.g., $1/512$), while dark red represented the highest risk (e.g., $1/8$).

The odds themselves were defined in the pilots' instructions (Appendix A) as the "probability of significant damage to your aircraft by the time you arrive at that place."[4] Note that each individual odds value is "twice as dangerous" as the one below it (e.g., $1/8 = 2*(1/16)$). Thus, each scale follows a base-2 logarithmic (log) scale. Log scales were chosen because they allow a wide range of values to be represented, using only a handful of numbers, while still allowing fine discrimination between small values.

[3]The term "path" is used here as the variable name because it is a term common to both aviation and mathematics. Otherwise, we can use "path" and "route" interchangeably.

[4]The authors are aware that risk is technically defined as some probability related to some specified length of time (e.g., say, if you flew through a "1/512" risk area for one hour, there might be a 1/512 chance of significant damage to your aircraft. However, it was decided to omit from the instructions that reference to time, assuming that this level of nuance would be simply be confusing to all but a very few pilots with formal training in probability theory (which, as it turned out, would have been all but one pilot).

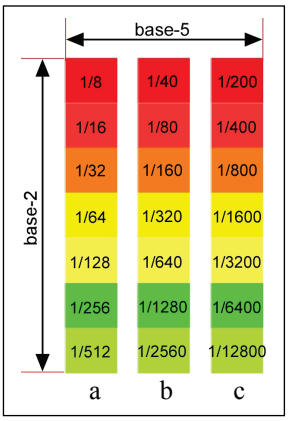


Figure 4. The three color-coded risk scales, a) High, b) Medium, c) Low-risk.

Moving from left to right in Figure 4, the three scales followed a base-5 log (e.g., 1/8 = 5*(1/40)). Thus, 4a is "five times as dangerous" as 4b, and so forth.

Going back to Figure 3, note that about 2/3 of the area on each test page is white. White represented the ambient, or *base-rate* risk probability—essentially, the odds of coming to harm just flying along, well clear of weather. This base rate was arbitrarily set at 1/10,000 to logically have a number that was smaller for flying outside the weather than flying inside it.[5] Essentially, the informal logical rule was "White is safest."

Probing for depth of understanding

The one exception to the "White is safest" rule was the very lowest risk value of all (Figure 4c, lightest green cell, risk = 1/12,800). This was actually *lower* than the base rate of 1/10,000 just discussed. In fact, this was a *probe* designed to see which pilots were paying close attention to the risk odds and deeply understood how to use them. This probe requires some explanation.

Recall that each pilot saw three scenarios. In the one scenario where lightest green represented "1/12,800" (Figure 3c) any pilot deeply understanding probabilities would be *less* likely to choose the "long way around the weather," since risk and path length could both be minimized simultaneously by "shooting the gap," that is, going through one of the two lightest-green "narrows" between the three risk gradients (i.e., Figure 3c, path "A" or "B").

So, this first probe consisted of that one probability scale (1/12,800) in that one scenario (Figure 3c). And, the "output" of that probe for that one scenario was whether or not the pilot shot the gap. In this one, special scenario, shooting the gap would make one suspect that perhaps the pilot understood the basics of minimizing risk and maximizing efficiency using this type of display.

To get an even better sense of this, a second probe was engineered—a small, white, kidney bean-shaped "island," visible in Figure 3 within the rightmost narrows of all three scenarios.

If a pilot chose to shoot a gap, then it should be done one way for Figures 3a and b, and a different way for 3c. If they shot a gap in 3a and b, the flight path should go *through* the white "island" (e.g., Figure 3a, path "B"), because the island's risk was defined as *lower* (1/10,000) than the rest of the gap (1/512 for Figure 3a, and 1/2,560 for Figure 3b, respectively).

In contrast, shooting the gap in Figure 3c should *not* go through the white island, because the island's risk was *higher* (1/10,000) than the surrounding light green area's defined risk (1/12,800). In Figure 3c, path "A" is safer than path "B."

In full honesty, these probes could not prove that pilots understood the display. But, they could show that they misunderstood it, or were not paying attention, both of which are important human factors concerns.

Order effects

Because of three test risk gradients, presentation order might exert an unwanted effect. For some reason, it could be that presenting the gradients in, say, Low/Medium/High (LMH) sequence might produce different results than, perhaps, High/Medium/Low (HML). This is a standard concern in research design and would not be a desirable result.

To guard against such order effects, the 30 three-page tests were counterbalanced by risk severity. That is, six different test versions were created, each having one of the six mathematically possible presentation orders (LMH, LHM, MLH, MHL, HLM, HML). Versions were then handed to pilots at random as they walked in, presumably distributing any order effect impartially across the entire experiment.

RESULTS

The way the test was constructed, the expected main effects were *risk level* and *motivation level*. Between the three scenarios, each on its own page, *path length* was logically expected to increase as overall scenario risk level increased, meaning that pilots should fly farther to avoid increased risk. Ideally, however, *maximum risk encountered* would remain relatively stable across pages, *if* pilots truly understood the concepts of risk and odds.

In contrast, within each separate scenario, as motivation level increased, path length was expected to decrease while maximum-risk-encountered increased. This would logically reflect the universal animal-behavioral tendency to take slightly more risk if either pushed by circumstance of punishment avoidance or pulled by the possibility of reward (Mackintosh, 1974).

As things turned out, this proved to be a rich dataset with interesting individual-difference themes at work as well. Therefore, the data will be described both quantitatively and qualitatively.

The pilot demographics are described first, followed by a standard data check, main effects, individual differences, correlations between variables, and simple modeling of risk-taking behavior.

[5]Again, as noted in footnote 3, the experimenters tried to minimize confusion by simply presenting the odds, with minimal explanation. Readers highly skilled in probability theory will recall that total risk is technically $Risk_{total} = 1 - \prod_{t=0}^{T}(1 - p_t)dt$, that is, a product of infinitely small risks associated with infinitely small time slices, which nonetheless represent a non-zero risk over some time T. As you may suspect, this paper does *not* hypothesize that this is how the average person actually calculates risk.

Pilot Demographics

Thirty GA pilots were recruited from a local flight school. Table 1 shows pilot demographics, including certificates and ratings. The point of collecting demographics is to assess how likely the results of an experiment are to generalize to the population at large. Because of the relatively high variability reflected in the standard deviations (SD), medians are also shown where appropriate.

Participants were mostly flight instructors, as reflected in the high percentage of CFIs. Consequently, despite the relatively low median age of 24, these pilots had a considerable amount of flight experience, with a median total flight hours (TFH) of 475.

Note that 2/3 of these pilots were instrument-rated (IR). This important detail will be discussed later.

Pilots were asked to self-rate their own weather flying skill on a 1-7-point Likert scale (shown in Appendix B). As Table 1 indicates, these "self-rating" responses ranged from 1 ("Lower 5%") to 7 ("Upper 5%), with both the mean and median equal to 4.0, and a standard deviation of 1.9. This implies that, as a group, these pilots considered themselves something more than novices, yet something less than seasoned experts.

Taken as a group, therefore, these pilots arguably fall somewhere between a "perfectly representative sample" (which is virtually impossible to obtain) and an unacceptably skewed or biased sample. In a very apt sense, they "satisfice" the experimental conditions for generalizability of results.

Preliminary Data Check

Appendix D details the standard data-check for distributional normality of DVs, outliers, treatment order effects, and interrater reliability in the scoring of path length. To summarize, interrater reliability is excellent ($r_{Spearman}$=.958, p<.0001)), and there are no significant treatment order effects.

Certain DVs were expected to be non-normal (e.g., bimodality of path length, pilot age, TFH, maximum risk taken), and that proved to be the case. Therefore, the majority of data will be analyzed with nonparametric statistics, particularly ones based on rank-ordering (Hollander & Wolfe, 1999).

Relations Between Path Length and Maximum Risk Taken

Table 2 shows the Spearman (rank-order) intercorrelation matrix for our two main DVs of path length and maximum risk taken along each flight path.

As expected, there is a within-subjects effect, evidenced by the positive sign and high significance of correlations within the same category. This implies that pilots tend to behave with significant consistency across the six scenarios, both on safety (maximum risk taken) and efficiency (path length). This supports the use of repeated measures to analyze other data.

At the same time, these correlations are not perfect. Shared-variable variance (R^2) across the two DVs (i.e., path length×max risk) varies between .154 (-.392^2) and .867 (-.931^2). This indicates that, to some degree, these are different measures, and justifies examining them both.

Finally, note that the correlations between safety and efficiency measures are negative. As expected, this implies that shorter path lengths tend to be riskier.

Main Effects

Ruling out trivial explanations

Any frank discussion of risk tolerance must deal openly with potentially trivial reasons why people might exaggerate their own willingness to take risks. Society, of course, values bravery, and this is a big part of the social status of being a pilot. Yet, scientific research ideally needs to distinguish bravery from bravado, particularly in situations where there is absolutely no chance of anyone's actually getting hurt. This is a thorny issue, since the Institutional Review Boards that approve research protocols typically feel constrained to rule out direct punishment for bad behavior (e.g., mild electroshock administered after a simulated "crash" is forbidden). This forces the experimenter to devise more subtle ways of determining exactly how the independent variables influence the dependent variables.

Table 1. Pilot demographics

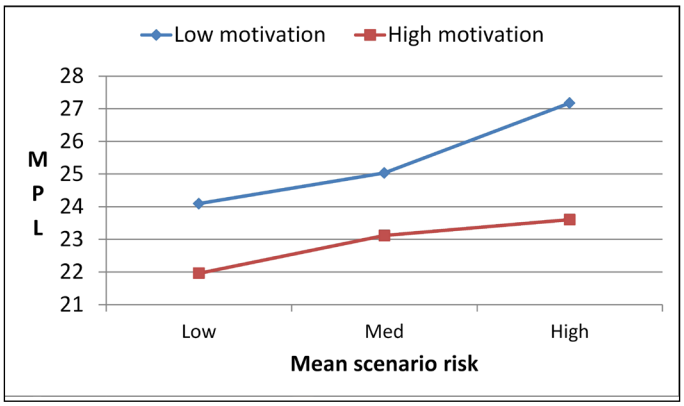
Certificates & ratings	Count	%		Min	Max	Mean	SD	Median
Instrument-rated	20	67	Age	20	72	29.2	12.5	24
CFI[a]	18	60	TFH	15	12000	1368.7	2506.5	475
CFII[b]	15	50	Self-rating	1	7	4.0	1.9	4
Commercial	18	60						
ATP[c]	2	7						
Multi-engine	15	50						

[a]Certified Flight Instructor. [b]Certified Flight Instructor--Instrument. [c]Air Transport Pilot.

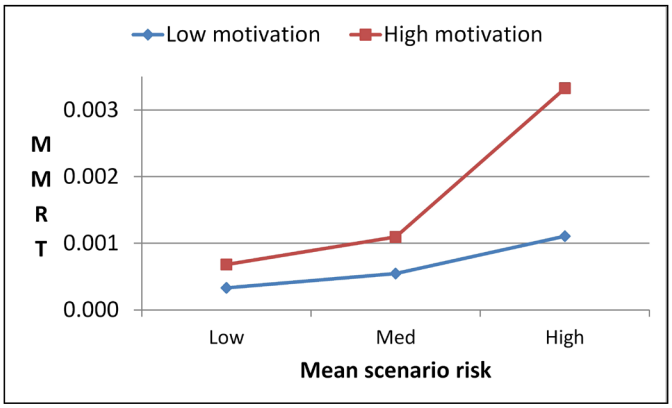
Table 2. Spearman correlations between path length and maximum risk taken.

			Path length				Max risk taken			
	Risk gradient		High		Med		High		Med	
		Motivation	Higher	Low	Higher	Low	Higher	Low	Higher	Low
Path Length	High	High	1							
		Low	.663	1						
	Med	High	.759	.745	1					
		Low	.535	.803	.793	1				
Max risk	High	High	-.833	-.686	-.587	-.392	1			
		Low	-.556	-.879	-.686	-.741	.644	1		
	Med	High	-.740	-.755	-.851	-.698	.720	.690	1	
		Low	-.442	-.804	-.699	-.931	.403	.835	.707	1

p<.05 highlighted dark gray. .01<p<.05 light gray. All others p<.001



a



b

Figure 5. a) *Mean path length* (in cm) and b) *mean maximum risk* taken as a function of mean scenario risk.

Table 3. Mean path lengths (MPL) and mean maximum risk taken (MMRT).

Risk gradient	Motivation	Mean path length[A]	MPL, disregarding motivation	Mean max risk taken	MMRT, disregarding motivation
High	Low	27.2	25.4	.00111	.00221
	High	23.6		.00332	
Med	Low	25.0	24.0	.00054	.00082
	High	23.1		.00109	
Low	Low	24.1	23.0	.00033	.00051
	High	22.0		.00068	

[A] Path-length units are centimeters-on-paper.

Recall that the statistical design involves repeated measures (a.k.a., within-subjects design). One of the attractive features of repeated measures design is that each participant serves as his or her own control. Statistically, change scores are examined, and the effect of examining change scores is that each individual's baseline "bravado level" is theoretically subtracted out during the analysis, leaving a purer measure of the effect of the IV(s).

IVs, DVs, and hypotheses

Recall that the main IV involved three different risk gradients and two motivation levels. Common sense supposes that risk-taking is a function of both risk level and motivation. As risk level rises, so should risk-avoidance. As overall motivation level rises, so should risk-taking.

Given the way the test was constructed, *mean path length* (MPL) and *mean maximum risk taken* (MMRT) are two reasonable measures of risk tolerance (dependent variables, or DVs). The longer the path, the less risk generally taken.[6] Similarly, the greater the maximum risk encountered along the path—the greater risk generally taken.

Figure 5 graphically shows the MPL and MMRT data, organized by risk and motivation level. Table 3 gives a more complete description, also showing MPL and MMRT collapsed across motivation level.

Groupwise main effect of risk level

Groupwise main effect of scenario risk level. The mean scenario risk level significantly affected pilot group behavior, but in in a way that initially seemed contradictory. As Figure 5a shows, risk avoidance (MPL) increased monotonically[7] as mean scenario risk increased, for both motivation levels ($p_{low\ motivation}$ = .0011 and $p_{high\ motivation}$ = .0005, nonparametric Friedman test of rank-orderings). However, so did the level of risk accepted (MMRT, Fig. 5b, $p_{low\ motivation}$ = .0004 and $p_{high\ motivation}$ < .0001). So, why did pilots, on average, accept increasing risk as the scenarios got riskier? This requires explanation.

We suspect that, as a group, these pilots did perceive and respond to labeled differences in risk, but more within scenarios than between them. In other words, colors representing greater risk did stimulate more risk-avoidance behavior. However, some pilots seemed not to pay strict attention to the numerical values as they changed from one scenario to the next. Instead, they may have adopted an simplified, non-numeric rule such as "Green is pretty safe, red is dangerous, and everything else lines up in between."

Therefore, as a group, these GA pilots did perceive and respond to labeled differences in risk. On average, colors representing greater risk appeared to stimulate risk-avoidance behavior. However, this conclusion is not cut-and-dried. There was considerable variation between pilots, as shall be seen in the **Individual Differences** section.

Main effect of motivation level

Motivation level also significantly affected behavior. Figure 5a shows that higher motivation led to shorter MPL. This held across all three risk levels (MPL=24.1 vs. 22.0 with $p_{low\ risk}$ = .0002, 25.0 vs. 23.1 with $p_{med\ risk}$ = .0003, and 27.2 vs. 23.6 with $p_{high\ risk}$ = .0004 for the nonparametric Wilcoxon test).

Individual Differences

It is important to discover average group characteristics about risk displays and how those affect aircraft navigation in the presence of differential motivation levels. At the same time,

[6] Pilots appeared to take the test seriously. There were no curlicue or "meant-as-a-joke" paths.

[7] Monotonic functions steadily increase or decrease (e.g., are not "U-shaped" or "upside-down U-shaped").

Table 4. Percentage of pilots with maximum risk taken ≥ .00125 (1/800) chance of serious damage to the aircraft.			
	Overall scenario risk level		
Motivation	Low	Med	High
Baseline (low)	6.7	6.7	36.7
Higher	17.2	24.1	75.9

individual differences are also important for a deeper understanding of the *range* of how individual pilots think about risk (or fail to). Did all pilots uniformly and completely understand all underlying principles of probability? Or, did understanding vary, as do most other kinds of knowledge and skill? If so, what mechanisms were most likely at work causing this variation in risk perception and/or tolerance?

Appendix C details the analytic method, the results of which are summarized here.

1. Many pilots superficially appeared to tolerate an extraordinarily high degree of risk (Table 4),
   a. particularly in the higher-motivation conditions and
   b. during the single highest-risk scenario
2. Some pilots may simply not have been paying close attention to the numerical risk scales.
3. Many pilots may misunderstand the concept of *absolute risk* (as opposed to relative risk).
   a. *Relative risk* varied according to the color scheme within each scenario. Green was always relatively safer than yellow, which was safer than red, and so forth.
   b. *Absolute risk* was stated as the numerical odds of harm associated with each color. This changed across scenarios.

Heuristics and Themes

If people rarely coldly calculate numerical risks, what are some simplifying heuristics they use as substitutes? Do all people use all heuristics, or do different people adopt different heuristic "styles?" If so, what are some of those styles, and what evidence can be found for their use? In instances where no heuristics are evident, can pilots at least be grouped with common characteristics into *themes*?

Appendix F presents detailed evidence seen for the use of heuristics and themes. This is summarized below. Heuristics are presented in quotation marks; themes are not.

An "index of confusion" (I_c) is presented where possible. Developed for this study and detailed in Appendix C, the I_c measures the variability (across the six scenarios) of pilots' maximum flight path risk divided by their minimum risk. Recall that—ideally—pilots with perfect understanding about probability and risk odds should have a strict idea about how much risk is acceptable and should, therefore, ideally show the same maximum risk tolerance across all three risk gradients. I_c therefore measures that degree of uniformity (variability). Low I_c is evidence for "low confusion" and greater understanding of risk probabilities.

1. *"Green means 'safe,' red is 'unsafe,' and yellow means 'exercise caution.'"* Figure A-3 in Appendix C amply shows that yellow and red tend to be avoided. But, beyond that, there is almost certainly a strong bias to resist assigning radical new meanings to these colors, perhaps because of lifelong experience with highway traffic lights. This conclusion is supported by heuristics 2 and 3 below. Many pilots have also flown successfully through light rain (green on NEXRAD), perhaps leading to a deep-seated feeling that "green isn't particularly dangerous," no matter what risk an experimenter might artificially try to assign it.
2. *"Always pick green."* (n=5, Table A-11, I_c=19.1). One intriguing heuristic involved five pilots who always flew through green areas, regardless of what numerical risk was actually associated with those areas. They apparently meant not to take high risk, but did so accidentally, in not fully understanding the risk gradients. This also relates to heuristic 3.
3. *The high-risk gradient was confusing to many.* By examining various threshold values for maximum-risk-tolerated, we could see if any of the three risk gradients may have seemed confusing to pilots. It appears that the high-risk (H) scenario was often confusing. We hypothesize that the WSR-88D Intensity Legend (Appendix F, Figure A-9) already associated with NEXRAD proactively interfered with learning the odds associated with our color scheme. Pilots already familiar with "safe colors" versus "unsafe colors" probably had difficulty replacing their existing notions for the arbitrary new ones we asked them to temporarily learn.
4. *"Avoid all colored areas."* (n=2, Table A-7, I_c=2.0). Two pilots always deviated around all colored areas, regardless of those areas' numerical risk or the scenario's motivation level. While arguably staying fairly safe by avoiding all colored areas, these pilots failed to pay close attention during the lowest-risk scenario, where light green meant risk below ambient level.
5. *"Top-rated pilots take risks."* (n=10, Table A-8, I_c=7.2). Non-instrument-rated (non-IR) pilots are supposed to avoid convective weather. Yet all 10 of the non-IR pilots were willing to fly at least into green areas. Nine of these were students at a top local flight school, actively pursuing advanced training. This tempts us to speculate that those pursuing a higher status may think and act a lot like those already holding it (Kolman, 1938).
6. *"Avoid risk unless there's a compelling reason."* (n=8, Table A-9, I_c=8.1). Eight pilots uniformly accepted greater risk under all higher-motivation conditions (high fuel price plus being late to an engagement) than they did under low-motivation conditions. Six of these varied widely in risk tolerance across gradients, implying confusion over what the numerical risk scales meant.
7. "*Resist small motivations.*" (n=14, Table A-10, I_c=7.9). The exact opposite of the "compelling reason" heuristic is that some pilots appear "motivation-resistant." For the most part, this heuristic encourages safety. Nine pilots were uniformly motivation-resistant across all three scenarios. An additional five were resistant across two scenarios. The relatively high average I_c was mainly due to high scores for just four pilots.

Table 5. Spearman *r* between selected variables (significance level is in parentheses if p < .10).

	Path length	Max risk	Age	TFH	IR	SR	PKA	Eq. 2
Age	.264	-.209	1					
TFH	.346 (.091)	-.188	.386 (.057)	1				
IR	.130	.005	-.183	.769 (<.001)	1			
SR	-.099	.151	.073	.608 (.001)	.639 (<.001)	1		
PKA	.370 (.048)	-.377 (.044)	-.010	-.050	-.160	-.266	1	
Eq. 2	.418 (.053)	-.218	.367	.420 (.058)	.168	.158	-.051	1
Eq. 3	.413 (.056)	-.210	.367	.420 (.058)	.178	.166	-.051	1.00 (<.001)

SR = Self-rating. PKA = Personal knowledge of aviation accident.

8. *"No risk too great."* High risk tolerance is a tricky category to assess. For one, it depends on how one defines "high." Also, it can reflect confusion, not true risk tolerance. With that in mind, there were just two pilots showing uniformly high risk tolerance, defined as 1/256 (.00391) chance of inflicting serious damage to the aircraft.
9. *Overt risk calculators.* Only one pilot overtly wrote his acceptable risk threshold on the test sheets. This was a private pilot working on his instrument rating. Unfortunately, his risk tolerance was also quite high (1/100, .01).

On the one hand, one should not read too much into this one case. One cannot assume that understanding probabilities leads to being more risk-tolerant.

On the other hand, it probably is safe to assume that most pilots—just like the vast majority of people—are not highly trained in probability theory. That is not surprising. Nor is it terribly hard to remedy, if research such as this finds it necessary.

Modeling Flight Risk Factors

Modeling should be reserved until after one has some feel for the data. Ideally, one desires to know what it was about these particular pilots (IVs) that modulated their risk tolerance (DVs). Models can often shed light on the relations between IVs and DVs.

It is wise to first examine simple relations, such as correlations, before exploring more complex models. Table 5 shows Spearman rank-order correlations between our DVs of path length and maximum risk tolerated (for the medium risk gradient/high-motivation condition), and the IVs of age, TFH, instrument rating (IR), the self-rating scale, whether or not a pilot had personal knowledge of anyone involved in an aviation accident, and two "risk-availability" metrics that will be explained below.

The medium-risk gradient is chosen here because its risk values seemed to best represent a range we might see in the real world. It also seemed the least confusing to pilots.

Unsurprisingly, Table 5 shows near-significant positive correlations between age and TFH, and significant correlation between TFH and IR (both highlighted gray). As expected, older pilots and IR pilots also tend to have more flight hours, although there is actually a negative correlation between age and IR here, since many pilots in this sample were young flight instructors.

Oddly, self-rating (SR) appears unrelated to age but, rather, highly related to TFH and IR (green-highlighted cells). Yet, self-rating is unrelated to risk tolerance.

The decision-making theory of *availability* (Kahneman et al., 1982) states that cognitive/affective constructs that are easily accessible ("available") to awareness are the ones most likely to bias decision making. For our purposes, the basic idea can be summed up as "If it happened to someone I know, it could happen to me."

This comports well with common sense. Logically, having personal knowledge of someone who had an aviation accident should reduce risk tolerance, as measured by our test. The more serious the accident(s), the closer the people involved to the pilot, and the closer in time the accident(s) was/were, the greater the correlation should be between an availability metric and path length and/or maximum risk tolerated.

Table 5 embodies the results for three preliminary, very simple risk-tolerance models. First, personal knowledge of someone involved in an aviation accident (PKA) was a simple yes/no measure, allowing for up to two accidents, so the scale ran from 0-2.

Equation 1 represents a second, additive model; Equation 2 represents a third, multiplicative model. In these two models, the "availability effect" is assumed to decay over time.

$$R_i = f\left(\frac{S_{1i} + C_{1i}}{t_{1i}} + \frac{S_{2i} + C_{2i}}{t_{2i}}\right) \quad (1)$$

$$R_i = f\left(\frac{S_{1i}C_{1i}}{t_{1i}} + \frac{S_{2i}C_{2i}}{t_{2i}}\right) \quad (2)$$

Here, the *i*th pilot's risk tolerance is a function *f* of the seriousness *S* of the accident, the closeness *C* of people involved, and how long ago in time *t* it happened. Appendix B shows the 7-point Likert scale forming *S* and the 3-point scale forming *C*. As previously stated, up to two accidents could be coded for each pilot.

Table 5 shows that PKA—the simplest model of all—was actually a slightly better performer than the Equations 1 and 2 models. This may have been because the Equations 1 and 2 models were plagued by missing data from seven pilots who left the *t*-values blank for 11 accidents they otherwise supplied data on.

In the end, since the models' correlations with our DVs are only barely significant, and since none of the other IVs' correlations are significant, we should halt modeling here. What Table 5 gives us is some preliminary indication that an availability-based

model is probably worth exploring in future studies. The form of that modeling equation should probably resemble

$$R_i = \frac{k_1 S_{1i} + k_2 C_{1i} + k_3 S_{1i} C_{1i}}{t_{1i}^{k_4}} + \frac{k_1 S_{2i} + k_2 C_{2i} + k_3 S_{2i} C_{2i}}{t_{2i}^{k_4}} \qquad (3)$$

to capture the effects of both main and interaction terms. Around 60 participants would be needed (15 per parameter), and they should be monitored to ensure that they provide complete data.

DISCUSSION

Graphical weather displays such as NEXRAD are now extensively being used by general aviation pilots. NEXRAD provides radar reflectivity mosaics useful to assessing flight risk during flight planning.

NEXRAD is not a "risk display," per se. Rather, it displays colored regions, which represent radar reflectivity values. These serve as a *proxy* for relative risk as pilots plan flight routes.

Human factors issues associated with such risk-proxy displays are an area of interest to the FAA. Most particularly, our study's goal was to understand the broader safety and efficiency benefits that graphical risk displays in general might bring, versus any tendencies they might have to induce undue tactical risk-taking (e.g., by pilots attempting to pick their way through overly narrow gaps between chaotic convective weather cells).

The current study attempted to address such questions. Using a simple paper-and-pencil test (depicted in Appendix A), 30 GA pilots drew six flight paths from a departure point to a destination point. Eight different numerical levels of risk (supposedly posed by weather) were combined on each page to create colored *risk gradients* similar to the colored "topographic" radar-reflectivity maps seen in displays like NEXRAD. Three different risk gradients were presented, each in the context of two different levels of motivation. Pilots' risk tolerance, based on measures such as *flight path length* and *maximum risk tolerated* were estimated. The data were then analyzed quantitatively and qualitatively.

Two significant quantitative findings emerged. First, with "motivation" operationalized as "fuel cost combined with time pressure," higher motivation led to shorter flight routes (greater efficiency), but at the cost of higher stated risk tolerance (greater risk). This is not surprising and is consistent with what is known about risk-benefit tradeoffs and the psychology of decision making.

Second, however, many pilots' exhibited surprisingly high apparent risk tolerance. Shown in Figure A-3 and Table 4, even given a threshold level of "1/800 chance of serious risk to the aircraft"—nearly 20 times higher than actual U.S. accident rates per flight hour—almost 37% of pilots accepted at least that much risk in the low-motivation/high-risk gradient condition, and nearly 76% in the high-motivation/high-risk gradient condition. Moreover, in more than half the flights tested here, pilots appeared to exhibit risk tolerances in excess of formal national policy goals (Appendix E shows derivation of that estimate).

There are several plausible explanations for this finding of high apparent risk tolerance, none of which imply that these were foolhardy pilots. First, it could simply be that people act braver on artificial tests than they do in actual circumstances where they could face serious harm. This can never be completely ruled out in any study of this type (although our quantitative methodology was designed to minimize its effects).

Third, aviation accidents are rare events, and comprehension of low probabilities is simply not some-thing many are good at. Beneath what may seem like risky behavior may be a very human tendency to think "It doesn't matter as much how dangerous something may be, if I don't do it very often."

All these potential explanations are reasonable and worth considering. In all likelihood, multiple effects may be at work to affect a pilot's risk tolerance.

Moreover, pilots may assess complex qualities, such as risk, through the use of simplifying rules (heuristics). For that reason, qualitative individual-differences analysis was conducted. This led to the identification of several likely themes and heuristics. These are detailed in the Results section.

To summarize these heuristics, only one of 30 pilots gave evidence of being an "overt risk-calculator." The rest appeared to "guesstimate" in one way or another. As mentioned above, there ap-peared to be a strong tendency for pilots to associate the meanings of colored areas with prior experience, specifically "red means danger," "yellow means caution," and "green is relatively safe." This prior association seemingly caused many pilots to disregard the actual, stated risk values and to substitute their own, preexisting meanings. This has profound implications for the design of such displays—namely that it is unwise to assign meanings to colors that seriously challenge what users already think those colors mean.

CONCLUSIONS

Pilots—like most people—vary in their stated levels of risk tolerance. But, there can be many reasons why.

For one, society reveres pilots for their courage, "skewing the mean upward," and perhaps creating inordinately high risk tolerance for those it casts its reverence upon.

Second, most pilots—like most people—use risk heuristics. These are simplifying rules, rather than complex mental calculations. A small number of these heuristics are clearly mistaken, unsafe, and need to be recognized and trained-out. The remaining majority is usually based on an intention of reasonable risk-taking and safe-yet-efficient flying, and need only be "tuned" to make sure they accomplish those goals.

That "tuning" should involve making these heuristics plain, not leaving them unconscious. It should involve training pilots' minds to clearly understand the risks associated with things like various kinds of adverse weather. They should, for instance, understand the NWS WSR-88D Weather Radar Precipitation Intensity scale and Weather Radar Echo Intensity Legend (see Figure A-9), and know that "Orange and worse are the colors you need to clear by at least 20 nautical miles."

Emotions also need to be trained. Pilots have to become "motivation-resistant," and maintain a clear "do-not-cross line" when it comes to risk under normal circumstances. They should be comfortable with not having to prove their courage to themselves or anyone else.

Finally, the color schemes in predictive displays should be kept relatively simple and consistent with those that pilots already know, for example from NEXRAD. Any great changes will simply result in confusion and lowered overall safety in the National Airspace System.

FUTURE RESEARCH

A modification of this paper-and-pencil study, using a third motivation condition,[8] has already been conducted. Forty-two pilots have been tested, and those data await analysis. That analysis should greatly contribute to the development of methodology for the general research paradigm.

The most important addition to this overall paradigm would be to computerize it, enabling visual presentation of dynamic, NEXRAD-like weather and/or risk-proxy gradients. That capability could be inexpensively set up as a part-task psychophysics test, giving greater control over the gradients in question, their severity, and movement. Such a paradigm would bring us closer to identifying the *information in the stimulus* that forms the beginning of this type of pilot decision making.

[8]Appendix A shows the two current motivation conditions. The added condition, meant to evoke "extremely high motivation for safety," is "Suppose you have a very close family member on board (wife, child, sibling, or parent)."

REFERENCES

Airline Owners and Pilots Association, Air Safety Institute. (2011). Nall Report. Frederick Maryland: AOPA.

Atmospheric Technology Services Corporation. (2013). Final report: Demonstration comparing the effects of probabilistic and deterministic forecast guidance on pilot decision-making and performance. Washington, DC: Federal Aviation Administration.

Batt, R., & O'Hare, D. (2005). General aviation pilot behaviors in the face of adverse weather. (Report B2005/0127). ACT, Australia: Australian Transport Safety Bureau.

Camerer, C.F., & Kunreuther, H. (1989). Decision processes for low probability events: Policy implications. Journal of policy analysis and management, 8(4), 565-92.

Federal Aviation Administration, Aviation Safety Information Analysis and Sharing (ASIAS). (2010). Weather-related aviation accident study. Downloaded 14 Feb., 2013 from http://www.asias.faa.gov/portal/page/portal/asias_pages/asias_studies/pdfs/2003-2007weatherrelatedaviationaccidentstudy.pdf

Federal Aviation Administration. (2012). NextGen implementation plan, March 2012. Downloaded 15 Feb., 2013 from http://www.faa.gov/nextgen/implementation/media/NextGen_Implementation_Plan_2012.pdf

Federal Aviation Administration (Feb. 19, 2013). Advisory Circular 00-24C. Downloaded 14 Aug., 2013 from http://www.faa.gov/documentlibrary/media/advisory_circular/ac%2000-24c.pdf

Federal Aviation Administration. (2014). FY2014 Scorecard. Downloaded 3 Nov, 2014. from http://www.faa.gov/about/plans_reports/performance/quarter_scorecard/media/2014/q2/General_Aviation_Fatal_Accident_Rate.pdf

Gibson, E.J., & Walk, R.D. (1960). Visual cliff. Scientific American, 202 (4), 64.

Gibson, J.J. (1979). The ecological approach to visual perception. Boston: Houghton Mifflin.

Hecht, H. (1996). Heuristics and invariants in dynamic event perception: Immunized concepts or nonstatements? Psychonomic Bulletin & Review, 3 (1), 61-70.

Hollander, M., & Wolfe, D.A. (1999). Nonparametric statistical methods. New York: Wiley.

Kahneman, D., Slovic, P., & Tversky, A. (1982). Judgment under uncertainty: Heuristics and biases. New York: Cambridge.

Kolman, H.C. (1938). Compliance, identification, and internalization: Three processes of attitude change. Journal of Conflict Resolution, 2(1), 51-60.

Lee, D.N., & Reddish, D.E. (1981). Plummeting gannets: A paradigm of ecological optics. Nature, 293, 293-294.

Mackintosh, N.J. (1974). The psychology of animal learning. New York: Academic Press.

Marsh, B., Todd, P.M., & Gigerenzer, G. (2004). Cognitive heuristics: Reasoning the fast and frugal way. In J.P. Leighton & R.J. Sternberg (Eds.). The nature of reasoning (pp. 273-87). New York: Cambridge University Press.

Merriam-Webster. (2013). Merriam-Webster's online dictionary. Downloaded 15 Feb., 2013 from http://www.merriam-webster.com/dictionary/risk

National Transportation Safety Board. (2005). Risk factors associated with weather-related general aviation accidents. (Report NTSB/SS-05/01). Downloaded 14 Feb., 2013 from http://www.ntsb.gov/doclib/safetystudies/SS0501.pdf

National Weather Service. (2013). Tornado risk map. Downloaded 11 Feb, 2013 from http://www.nws.noaa.gov/climate/local_data.php?wfo=LMK

Papert, S., & Harel, I. (1991). Constructionism. New York: Ablex.

Simon, H.A. (1955). A behavioral model of rational choice. Quarterly Journal of Economics, 69, 99-118.

Simon, H.A. (1990). Invariants of human behavior. Annual Review of Psychology, 41, 1-19.

Steiner, M., Bateman, R., Megenhardt, D., Liu, Y., Xu, M., Pocernich, M., & Krozel, J. (2010). Translation of ensemble weather forecasts into probabilistic air traffic capacity impact. Air Traffic Control Quarterly, 18(3), 229 – 254.

von Glasersfeld, E. (1995). Radical constructivism: A way of knowing and learning. London: Routledge-Falmer.

Wiggins, M.W., Azar, D., Hawkins, J., Loveday, T., & Newman, D. (2014) Cue-utilisation typologies and pilots' pre-flight and in-flight weather decision-making. Safety Science, 65, 118-24.

APPENDIX A
Sample "High Risk" Test

This shows a sample "high-risk test," one of three risk gradients presented to pilots. The task was for each pilot to draw a line between the "Departure" and the "Destination," representing the flight path they would choose, given the set of conditions presented. This was intended to elicit individual risk tolerance. Two different motivation conditions were described in the instructions (shown below), resulting in two paths that could be compared for path length, maximum level of risk encountered, and risk heuristics.

This cockpit display shows the *probability of significant damage to your aircraft by the time you arrive at that place*. For the purposes of this experiment, assume the weather front is completely static. Therefore, this single snapshot is quite reliable, even over time.

A. Suppose fuel is its normal price and you're in no hurry. Draw a line showing the shortest flight path acceptable to you from Departure to Destination. Label it "A".
B. Suppose fuel is *twice as expensive* and you are *late to an important engagement*. Draw a second line showing the shortest flight path acceptable to you. Label it "B".

Lines "A" and "B" can be the same or different. There are no "right" or "wrong" answers, except according to your own standards of safety. Answers are completely confidential and will never be entered into your airmen record.

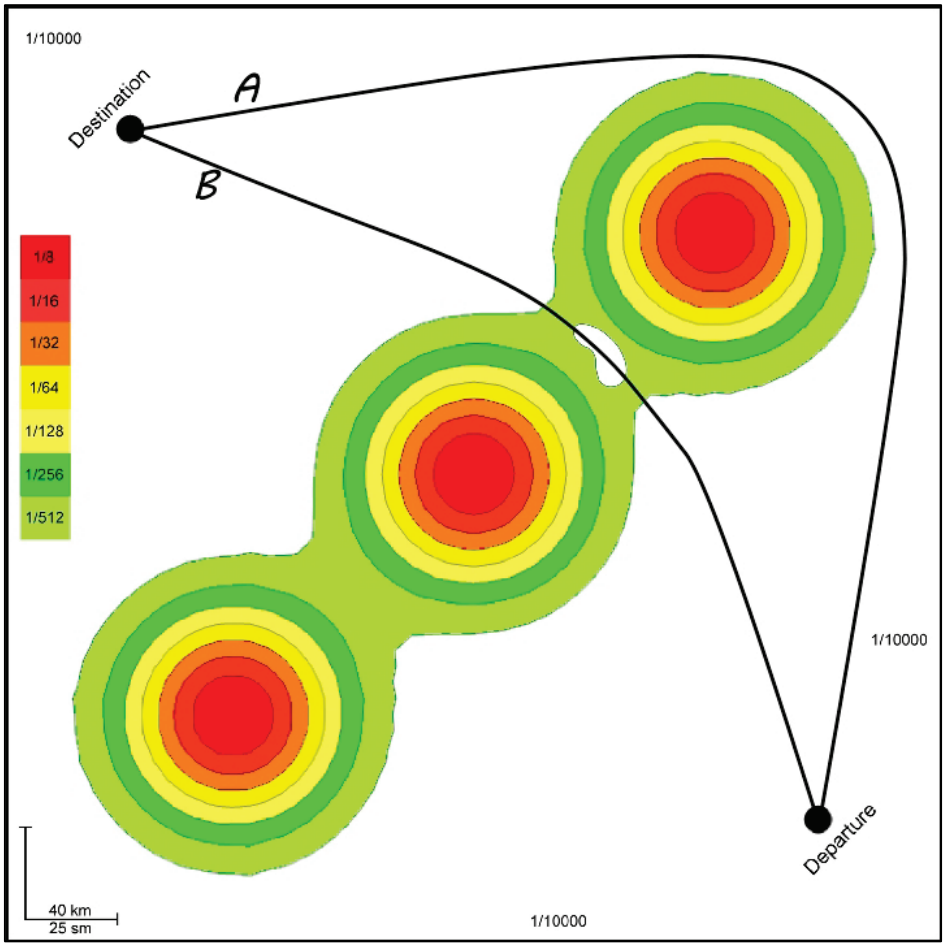


Figure A-1. Sample page from the test, showing the colored risk scale, its associated odds of serious damage to the aircraft, and the NEXRAD-like "topographic" risk gradient.

APPENDIX B
Demographic Questions

The demographic questions put to pilots were: a) (top, left) basic information on certificates and ratings, b) (top, right) a 7-point Likert scale representing pilots' self-rating of their weather flying skill and experience, c) (bottom, left and right) the scales and information used to construct the "availability metric" of Equations 2 and 3 in the **Results** section. To test the availability hypothesis, we asked pilots if they personally knew of anyone involved in an aviation accident. The assumption was that personal knowledge of such an accident would be more cognitively/emotionally "available" and, therefore, bias that pilot to be more cautious.

Please check all that apply

☐ Private pilot ☐ CFII ☐ Multi-engine
☐ Instrument-rated ☐ Commercial _____Age
☐ CFI ☐ ATP _____Total
flight hours

How would you honestly rate yourself in terms of weather flying skill and experience? (check one)						
☐	☐	☐	☐	☐	☐	☐
Lower 5%	Lower 15%	Slightly below average	About average (50%)	Slightly above average	Upper 15%	Upper 5%

Q: Has anyone *you know personally* been involved in an aircraft accident of any kind? ☐Yes ☐No

Q: If "Yes," about how many years ago? How serious was it? Use the table at right. ⇨ If multiple people were involved, write numbers in the boxes. You can list up to 2 separate accidents.

Accident 1	☐acquaintance ☐friend ☐relative (including yourself)						
Acc 1→ About how many years ago? Acc 2 →	☐	☐	☐	☐	☐	☐	☐
	No injuries	One minor injury	Multiple minor injuries	One serious injury	Multiple serious injuries	One fatality	Multiple fatalities
	☐	☐	☐	☐	☐	☐	☐
Accident 2	☐acquaintance ☐friend ☐relative (including yourself)						

APPENDIX C
Individual Differences

This appendix discusses *individual differences*—the analysis of individual pilot responses—in which we attempt to discover and quantify commonalities among pilots, thereby identifying heuristics and themes. This type of analysis is rarely presented in journal articles because it may lack the mathematical rigor of standard quantitative analysis, usually takes up far too much space, and can be painfully tedious to read. Nonetheless, at the very least, astute researchers recognize its value in generating hypotheses for future research. And, when conducted carefully and cleverly, it can lead us to conclusions we would have overlooked, had we not invested the effort.

One caveat does need to be made clear: This individual differences analysis will not use repeated measures methodology. Consequently, there is no statistical way to control for exaggeration or bravado in analyzing causes of risk tolerance. Therefore, one should interpret these data conservatively.

Color scheme. The test's color scheme and risk levels are detailed below in Figure A-2.

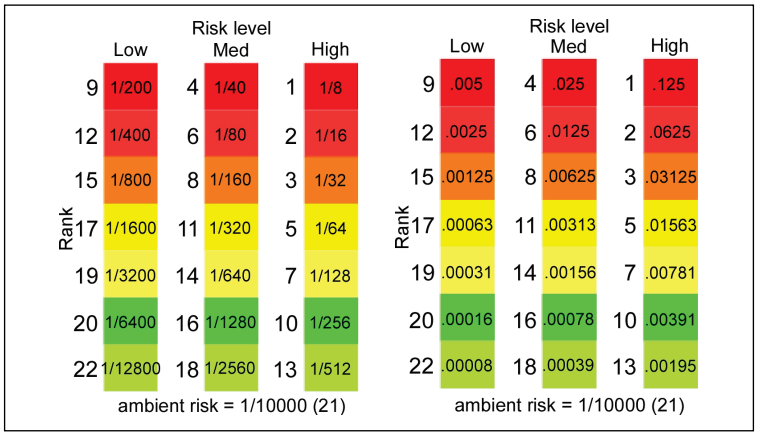
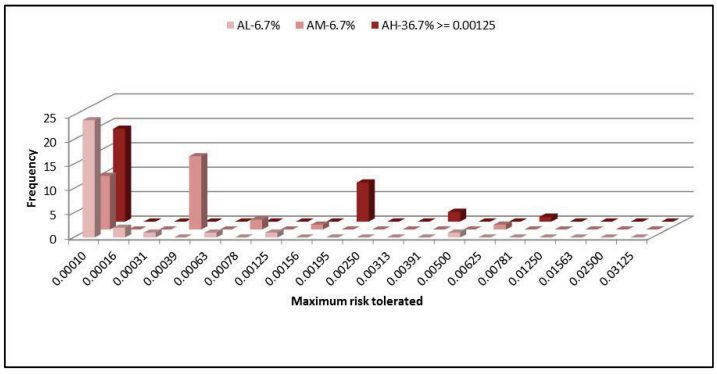
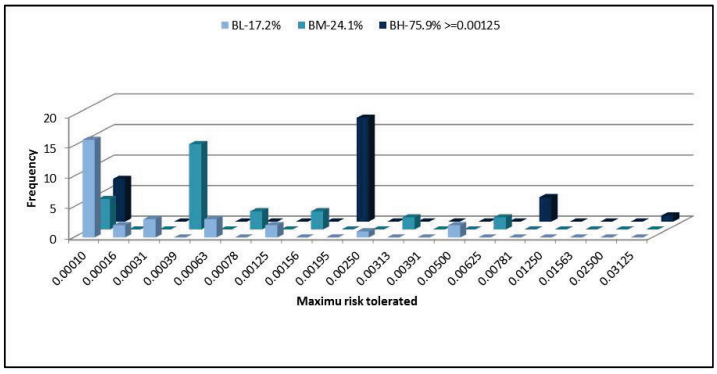


Figure A-2. Rank-ordered risk categories as odds and decimals. These ranks are used in the tables that follow, to represent maximum risk taken on a specified flight path, and can be used to visualize the participant's risk strategies.

Overall risk tolerance. The first thing evident in graphing out risk tolerance is the seemingly high degree of risk that some pilots appear to tolerate, particularly when motivation is elevated. Compare the maximum-risk-taken levels (the horizontal axes in Figures A-3a-b) for this study's pilots to the actual 2010 U.S. non-commercial GA total (fatal+non-fatal) accident rate of about 0.000063 accidents per flight hour (AOPA, 2011). Pay particular attention to the numbers of pilots (the vertical axes) stating risk tolerance greater than 0.00125 (1/800 chance of serious damage to the aircraft), which is 19.8 times greater than the actual 2010 accident rate ($p_{odds\ ratio} < .00001$).



a



b

Figure A-3. Frequency histograms of stated maximum risk tolerance for a) low-motivation (baseline) A-condition. (AL=lowest risk, AM=medium risk, AH=highest risk), b) higher-motivation B-condition. The horizontal (x) axis represents risk level, the vertical (y) axis represents numbers of pilots.

Clearly, something is happening when this many pilots state they would take on this much risk. We saw earlier, in the context of repeated measures, that part of the effect involves motivation level. But, what else can explain the rest of this seemingly high risk tolerance?

Unfortunately, we cannot easily dismiss exaggeration. But, we can argue that its effect is minimal in the Low (L) and Medium (M) scenario risk level conditions, because Table 4 shows those percentages to be relatively small. Only in the High (H) risk condition is tolerance inordinately high.

Over the course of the next few pages, we will argue that this seemingly high risk tolerance is actually mainly due to *misunderstanding of relative versus absolute risk*.

Understanding relative risk. Most of these pilots seemed to understand the basic color-coding of *relative levels of risk within a single scenario*. In other words, red is more dangerous than yellow; yellow is more dangerous than green, and so forth. We can infer that by noting that only one pilot drew purely straight paths from Departure to Destination (which took him directly through red cells). Everyone else drew curved paths, as one would expect.

Curved flight paths would be more expensive, in terms of time and money. Given the reasonable assumption that people value time and/or money, then this near-universal willingness to trade these for safety implies that these pilots understood the basics of *risk differentiation* (by color) and *risk prioritization* (by internal cognitive construct of threat level).

In other words, they understood relative risk.

We can further support this by looking at how motivation affected their maximum levels of acceptable risk (Appendix C, Table A-4), although the data are occasionally confusing.

Logic says that risk tolerance should increase (or, at least, not decrease) with motivation. This proved generally, but not completely true. In 80 of the 90 scenarios, pilots accepted as much or more maximum-risk-taken for the high-motivation (path "B") condition than for the low-motivation (path "A") condition $p_{Wilcoxon} < .000001$). In contrast, in nine of those 90 scenarios, a total of seven of the 30 pilots (23%) "flipped" and accepted less risk in H than in L (which one could argue shows misunderstanding of relative risk).

So, what of those nine troublesome reversals? Appendix C shows us exactly who those seven pilots were, and details the risk and motivation levels of the scenarios (Tables 7-9).

We conclude that these seven pilots were simply not paying close attention to all the odds on the scale.

Confusion over absolute risk. How well did pilots understand absolute risk, from one scenario to the next? Did pilots pay close attention to the actual numbers stated on each page (the odds)? Did they understand those odds with a firm-enough grasp of probability to apply them uniformly from one scenario to the next?

If we have an absolute idea what odds mean—as opposed to just a relative idea—then, for a given level of motivation, we should generally pick pretty much the *same maximum acceptable risk from one scenario to the next*.

The data do not support that picture at all. In only one case did a pilot (#22) specifically state his acceptable risk criterion, overtly showing a grasp of probabilities. This single individual consciously accepted any risk less than 1/100, and made that clear by writing "less than .01" on the test sheets.

All other evidence of probability awareness has to be inferred by us indirectly from the data. A few pilots showed relatively uniform risk tolerance across scenarios. But, in most cases, their tolerance varied considerably from scenario to scenario, indicating misunderstanding of absolute risk.

To better visualize the entire group's range of risk understanding, we can create a small "uniformity metric." Let us first divide the data into low-motivation (path "A") and high-motivation (path "B") conditions, because we know that motivation modulates risk tolerance. Next, to use "A" trials as an example, let us calculate a ratio between the *maximum* acceptable risk each pilot was willing to take in those three "A" trials, divided by the *minimum* acceptable risk for the same three trials. Then, this ratio (Eq. 4) can directly measure uniformity across all three "A" trials.

As detailed in Appendix E, we can define an "ideal" ratio of $R_{Max/Min} < 2$, because of the way the test is constructed. And, we can use that "ideal" of risk tolerance uniformity as a measure of absolute risk understanding.[1] The greater $R_{Max/Min}$ rises above 2, the more the pilot may be confused about probability.

$$R_{Max/Min} = \frac{odds_{Max}}{odds_{Min}} \qquad (4)$$

Figure A-4 shows the frequency histogram for $R_{Max/Min}$.

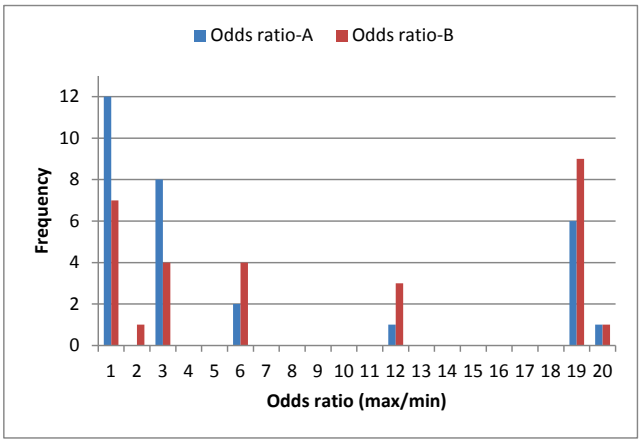


Figure A-4. Numbers of pilots (y-axis) scoring at various levels of $R_{Max/Min}$ (Eq. 4, x-axis) for the low-motivation (A) condition and higher-motivation (B) condition. Any odds ratio score > 2 suggests pilot confusion about odds.

Figure A-4 shows 19 instances of Equation 4 being at or below the ideal score of 2, versus 40 above it ($p_{binomial} = .004$). This suggests significant confusion about what absolute odds mean and how one would go about setting up three separate flight paths with similar risk. From Table 4, we can see that this confusion appears to be *particularly strong when colors represent very high risk levels*.

Now, let us expand $R_{Max/Min}$ into a "index of confusion" (I_c) by adding points each for instances where pilots either

a. in the lowest-risk gradient, failed to traverse light green, which was safer than ambient risk (1 pt),
b. in the lowest-risk gradient, in traversing light green, intersected the white, bean-shaped "island," which actually represented *greater* risk (1/10000) than the surrounding light green (1/12800, ½ pt), or
c. in the medium and high gradients, in traversing light green, *failed* to intersect the white, bean-shaped "island," which now represented *lower* risk than the surrounding light green (½ pt).

[1] Because each risk is itself an odds, Equation 1 is technically an "odds ratio." However, we cannot use it to determine statistical significance (as we can a standard odds ratio) since its components are not based on actual, repeated numbers of Bernoulli trials.

In the tables that follow, cells meeting condition a, b, or c will be superscripted with the matching letter (Appendix E, Table A-6 contains the full listing of such errors).[2]

Metrics such as I_c, while admittedly somewhat arbitrary, will allow us to better construct a compelling argument for notions such as the one that many pilots are coming pre-biased with internal affective and cognitive constructs of what colors represent within a risk context. We can further argue that trying to shoehorn existing color schemes into representing extremely high risk levels may not make much sense.

Table A-1. Categorization of pilot behavior.						
		Risk level				
Category	**Heuristic**	**Low**	**Med**	**High**	**N_A**	**N_B**
W	"Always pick white"	21	21	21		
G	"Always pick green"	22 OR 20	18 OR 16	13 OR 10		
Gw	"Low & Med=green, High=white"	22 OR 20	18 OR 16	21		
gW	"Low=green, Med & High=white"	22 OR 20	21	21		
R	High risk-taking	<19	<18	<21		
E	Evidence of a probability error					
	True Probability Calculator	r	≈r	≈r		
X	Refused to fly					

Table A-2. Maximum risk accepted for **low**-motivation scenario ("A").

Risk level	**Pilot ID**																												
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Low	22i	20	15	21	21	22i	22i	22i	19	21	22i	22	22	21	22	17	22i	22i	21	21	22	9	21	21	22i	21	22i	22i	20	22i
Med	18i	18	11	21	21	18i	18i	18i	16	21	18i	18i	21	21	21	18i	18i	18i	18	21	18	8	21	18i	21	21	21	18i	16	18
High	13i	13i	10	21	21	13i	21	13i	13	21	21	21	21	21	21	21	21	21	13i	21	21	7	21	21	21	21	21	13i	10	13i
	N			N	N	N	N	N		N	N		Y	N	Y		N	N	N	N					N	N	N	N	N	N
Cat.	G	G	R	W	W	G	Gw	G		W	Gw	Gw	gW	W	gW		Gw	Gw		W	Gw	R	W		gW	W	gW	G	G	G
"I" denotes the white, bean-shaped "island" within the lightest-green scale (ranks 13, 18, and 22).																														

Table A-3. Maximum risk accepted for **higher**-motivation scenario ("B").

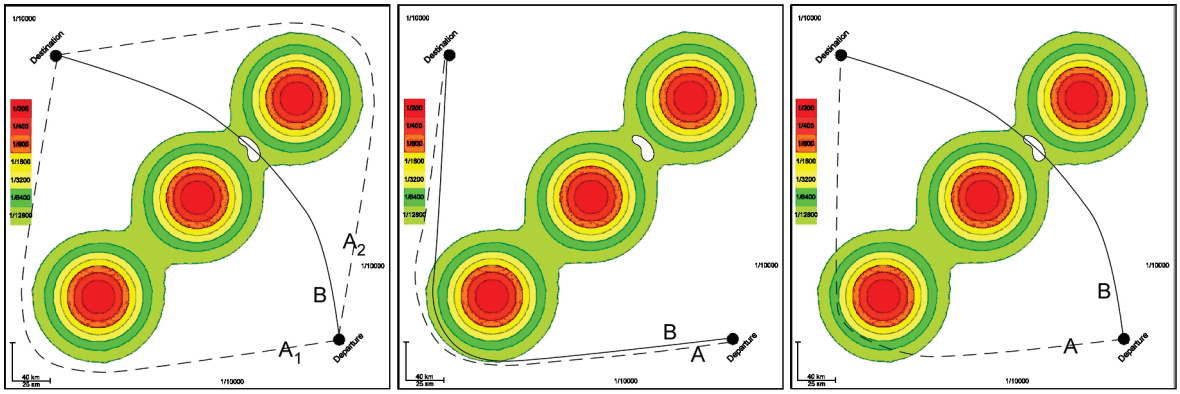
Risk level	**Pilot ID**																												
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Low	22i	20	15	22i	22i	9	19	22i	17	22i	12	19	22	22	22	15	22	17	20	21	17	9	21	22i	X	22i	22i	22i	22i	19
Med	18i	18	11	21	21	8	18	18i	11	18i	11	16	21	18	18i	16	18	14	18	21	16	8	21	18	X	18i	18i	18i	18i	14
High	13i	13i	7	21	21	3	13i	13i	7	13i	13i	13i	21	13	13i	13i	21	13	13i	21	13i	7	21	13i	X	13i	21	13i	13i	7
	N			N	N		N	N		N			Y		Y					N			N	N		N		N	N	
Cat.	G	G	R	gW	gW	R		G	R	G			gW	G	G		Gw	R	G	W		R	W	G	--	G	Gw	G	G	

Table A-4. Consistency between low-motivation ("A") and higher-motivation ("B") scenarios.

Risk level	**Pilot ID**																												
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Low	=	=	=	<	<	>	>	=	>	<	>	>	=	<	=	>	=	>	>	=	>	=	=	<	X	<	=	=	<	>
Med	=	=	=	=	=	>	=	=	>	>	>	>	=	>	>	>	=	>	=	=	>	=	=	=	X	>	>	=	<	>
High	=	=	>	=	=	>	>	=	>	>	>	>	=	>	>	>	=	>	=	=	>	=	=	>	X	>	=	=	<	>
"<" means "Took more risk with high-motivation (n=33), "<" means the opposite (n=9), "=" means A and B were identical (n=40). "X" means "Pilot refused to fly."																														

Misunderstanding of relative risk for seven pilots. Seven pilots misread the lowest-risk scenario, where the lightest green actually represented a risk level lower-than-background. Shown in Figure A-5a, pilots 4, 5, 10, 24, and 26 all took one of the two long ways (path A_1 or A_2) around the entire weather system in the low-motivation condition, while cutting through the "valley" containing the bean-shaped "island," *and* cutting through the island in the higher-motivation (path B) condition. Meanwhile, pilot 14 (Figure A-5b) took the long way around in both A and B conditions, but "cut the corner" into the light green in B, resulting in a slightly shorter path length. Finally, pilot 29 took the long way around in A, cutting through the dark green in the process, while, in B, cutting through the valley and traversing the island, just like 4, 5, 10, 24, and 26.

[2] We fully realize the arbitrary nature of this point assignment, and acknowledge that its validity and reliability have not been established. Nonetheless, there is merit in attempting to understand the degree to which pilots misunderstood (or ignored) the odds presented here. And, having even a crude system of prioritizing misunderstanding is arguably better than no system at all.



a b c

Figure A-5. Lowest-risk scenario flight paths of pilots a) 4, 5, 10, 24, and 26, b) 14, and c) 29

This does not necessarily reflect anything dire. It probably only means that these seven pilots were not paying close attention to all the odds on the scale. No one told them there was going to be an anomaly here. So, if they only attended to, say, the odds associated with red and yellow, we should not read too much into this particular error. What it may indicate is simply that not everyone thinks about probabilities in terms of absolute numbers. Many of us may code these concepts relatively. If so, that is an interesting, important finding, and we will see if we can find more support for that notion as we proceed in our analysis.

APPENDIX D
Preliminary Inspection of Data

This appendix details an inspection of the dependent variable (DV) data. Statistics rest on assumptions, for instance about how DV response variation is distributed. Violation of these assumptions can lead to incorrect use of statistics. It is vital, therefore, to check certain DV characteristics, such as frequency distributions, to ensure that the choice of statistical methods is correct.

Path length frequency distributions are bimodal. Pilots could fairly safely avoid risk by taking a "short way through the valley," avoiding "risk mountains." Or, they could always take the "long way around" the entire system. We therefore expected path lengths to be bimodal, and that was clearly the case. Figure A-6 shows a typical histogram for path lengths.

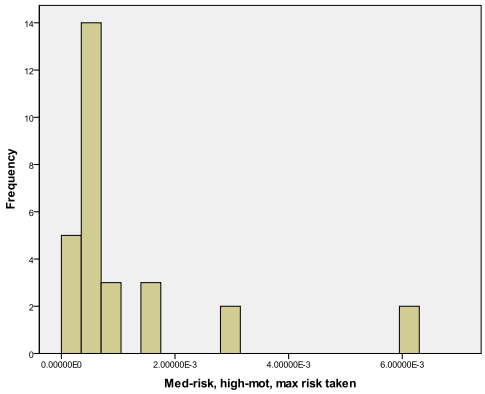
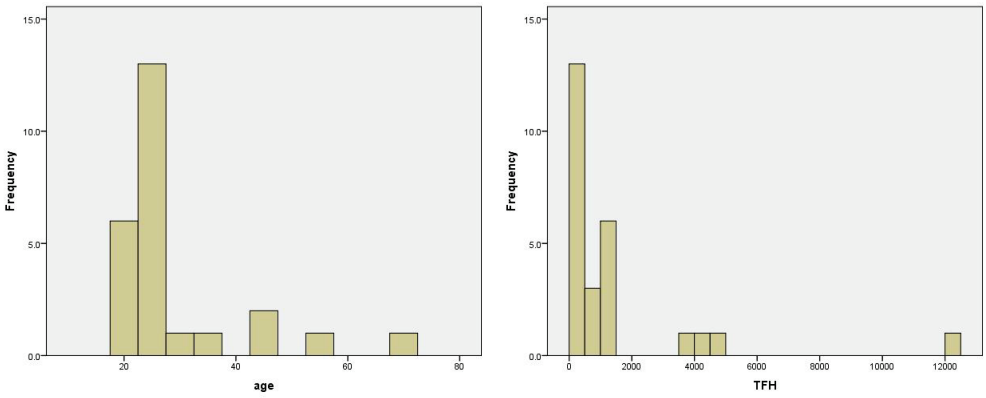


Figure A-6. Frequency histogram of path lengths for the high-risk, low-motivation condition. The y-axis represents numbers of pilots. The x-axis shows binned path lengths, in cm.

Notice how paths less than 26 cm fall cleanly into one category, while those greater than 26 fall cleanly into another.[1] This means we should analyze paths not as normally distributed, with standard parametric statistics, but rather nonparametrically, with statistics based on rank order.

Maximum risk frequency distributions are skewed. Figure A-7 shows the histogram for the medium-risk gradient, higher-motivation condition for the DV of maximum risk taken.



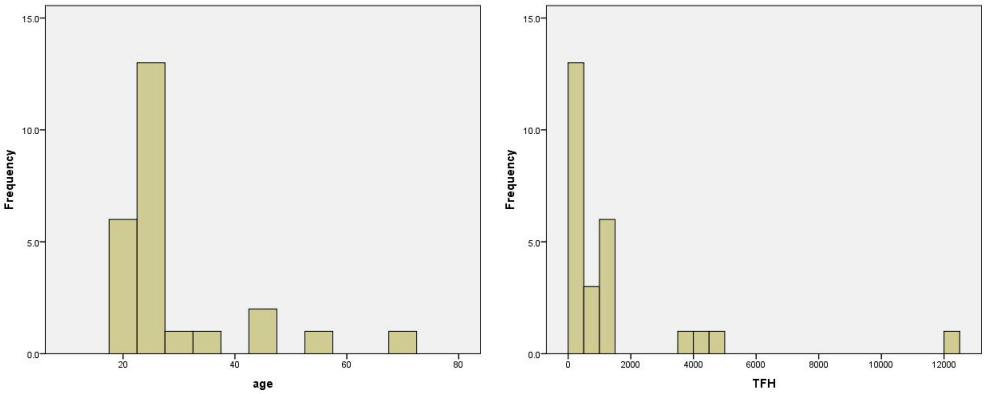
a b

Figure A-7. Frequency histogram of maximum risk taken for the medium-risk, higher motivation condition.

[1] As shown later in Table A-6, there was only one individual (Pilot 25) who declined to fly in the three higher-motivation scenarios, resulting in path lengths of zero for those three scenarios.

This non-normal shape is typical of the others for that DV. Notice the long right-hand tail and how the overall distribution resembles a gamma or Weibull function.

Outliers. Figure A-8 shows the histograms for age and total flight hours (TFH). These distributions are also non-normal ($skew_{age}$ = 2.284, $SE_{skew,age}$ = .464, $p_{z,skew,age}$ < .00001; $skew_{TFH}$ = 3.375, $SE_{skew,TFH}$ = .464, $p_{z,skew,TFH}$ < .00001).



a b

Figure A-8. Frequency histograms for pilot a) age, b) total flight hours.

Therefore, again, most analysis should be done with nonparametric statistics.

Treatment order effects. During testing, one would not want to see average risk tolerance increasing or decreasing merely as one particular risk gradient followed another (gradient-order effect) or merely as the test progressed (temporal-order effect). That would imply that risk tolerance is an unstable trait, influenceable by factors such as priming, fatigue, boredom, or learning effects.

As stated earlier, to guard against gradient order effect, six versions of the test were created. These were handed out randomly as pilots walked in the door. One still has to check for temporal order effect, however, since that cannot be counterbalanced. Every 3-page test unavoidably has a first, second, and third page, no matter what happens to be on each page.

Statistical analysis shows no evidence of significant temporal-order effect. At first blush, the mean *ranks* of the data seem to decrease monotonically by page (Table A-5), which might imply an order effect—namely, pilots taking increasing risk as time went on. However, a nonparametric repeated-measures Friedman test was not significant. Nor are the mean path lengths themselves uniformly monotonic.

Table A-5. Test of page-order effects.

Motivation level	Page	Mean path length (cm)	Mean rank	$p_{Friedman}$
Low	1	25.2	2.22	.085
	2	26.2	2.10	
	3	24.9	1.68	
High	1	23.1	2.21	.203
	2	23.1	2.03	
	3	22.5	1.76	

Interrater reliability. Given manual path length scoring, we must demonstrate accuracy by having high interrater reliability between two scorers. That does appear excellent, with $r_{Pearson}$ = .998 and the nonparametric $r_{Spearman}$=.958 (p<.0001). These extremely high rs were due to a) the measuring wheel's high accuracy (±0.25%), b) the large number of pairs correlated (30×3×2=180), and c) reliably high consistency between scorers (the mean difference score between pairs was only 1.00 mm, SD 2.13 mm).

APPENDIX E
Analysis of Individual Risk Tolerance

This appendix describes the very finest-grained analysis of each pilot's risk tolerance. By color-coding the numerical risk-taken, patterns become much easier to spot. Please refer to Figure A-2 in Appendix C for the detailed color-coding scheme used in the tables below.

Maximum risks taken. Table A-6 shows the maximum risk accepted by each pilot for each trial. These are color-coded to reflect the maximum-risk color intersected by each flight path.

Table A-6. Maximum acceptable risk level, by scenario.

ID	IR?	Motivation "A" (baseline) Scenario risk level			Motivation "B" (higher) Scenario risk level			Tot "A" ≥ 0.00125	$R_{Max/Min}$ "A"	$R_{Max/Min}$ "B"
		Low	Med	High	Low	Med	High			
1	N	0.00010[b]	0.00039	0.00195	0.00010[b]	0.00039	0.00195	1	19.5	19.5
2	Y	0.00016	0.00039[c]	0.00195	0.00016	0.00039[c]	0.00195	1	12.5	12.5
3	Y	0.00125	0.00156	0.00391	0.00125	0.00156	0.00781	3	3.1	6.3
4	N	0.00010[a]	0.00010	0.00010	0.00010[b]	0.00010	0.00010	0		
5	Y	0.00010[a]	0.00010	0.00010	0.00010[b]	0.00010	0.00010	0		
6	N	0.00010[b]	0.00039	0.00195	0.00500	0.00625	0.03125	1	19.5	6.3
7	N	0.00010	0.00039	0.00010	0.00031	0.00039	0.00195	0	3.9	6.3
8	Y	0.00010[b]	0.00039	0.00195	0.00010[b]	0.00039	0.00195	1	19.5	19.5
9	Y	0.00031	0.00078	0.00195	0.00063	0.00313	0.00781	1	6.3	12.5
10	Y	0.00010[a]	0.00010	0.00010	0.00010[b]	0.00039	0.00195	0		19.5
11	Y	0.00010[b]	0.00039	0.00010	0.00250	0.00313	0.00195	0	3.9	
12	Y	0.00010	0.00039	0.00010	0.00031	0.00078	0.00195	0	3.9	6.3
13	Y	0.00010	0.00010	0.00010	0.00010	0.00010	0.00010	0		
14	Y	0.00010[a]	0.00010	0.00010	0.00010[a]	0.00039[c]	0.00195[c]	0		19.5
15	N	0.00010	0.00010	0.00010	0.00010	0.00039	0.00195	0		19.5
16	Y	0.00063	0.00039	0.00010	0.00125	0.00078	0.00195	0	6.3	2.5
17	N	0.00010[b]	0.00039	0.00010	0.00010	0.00039	0.00010	0	3.9	3.9
18	N	0.00010	0.00039	0.00010	0.00063	0.00156	0.00195	0	3.9	3.1
19	Y	0.00010[a]	0.00039[c]	0.00195	0.00016	0.00039[c]	0.00195	1	19.5	12.5
20	N	0.00010[a]	0.00010	0.00010	0.00010[a]	0.00010	0.00010	0		
21	Y	0.00010	0.00039[c]	0.00010	0.00063	0.00078	0.00195	0	3.9	3.1
22	N	0.00500	0.00625	0.00781	0.00500	0.00625	0.00781	3		
23	Y	0.00010[a]	0.00010	0.00010	0.00010[a]	0.00010	0.00010	0		
24	Y	0.00010[a]	0.00039	0.00010	0.00010[b]	0.00039[c]	0.00195	0	3.9	19.5
25	Y	0.00010[b]	0.00010	0.00010	refused to fly			0		
26	N	0.00010[a]	0.00010	0.00010	0.00010[b]	0.00039	0.00195	0		19.5
27	Y	0.00010[b]	0.00010	0.00010	0.00010[b]	0.00039[c]	0.00010	0		3.9
28	Y	0.00010[b]	0.00039	0.00195	0.00010[b]	0.00039	0.00195	1	19.5	19.5
29	Y	0.00016	0.00078	0.00391	0.00010[b]	0.00039	0.00195	1	25.0	19.5
30	Y	0.00010[b]	0.00039	0.00195	0.00031	0.00156	0.00781	1	19.5	25.0

[a]Pilot failed to recognize that light-green risk < ambient (baseline) risk within the low-risk gradient.
[b]Pilot intersected the white, bean-shaped "island" representing *greater* risk within the low-risk gradient.
[c]Pilot *failed* to intersect the white, bean-shaped "island" representing lower risk within the medium- or high-risk gradient.

Note that, in the low-risk scenarios, a cell can be colored light green, which has an associated risk level of .00008 (1/12800), yet still have a risk level of .00010 (1/10000, associated with "white space"). This is because, even though the flight path traversed light green, it still flew through white to get there.

Calculation of $R_{Max/Min}$. Recall Equation 4

$$R_{Max/Min} = \frac{odds_{Max}}{odds_{Min}}$$

To illustrate from Table A-6, for instance, for pilot #1A, $R_{Max/Min}$ = Max(0.00010, 0.00039, 0.00195)/Min(0.00010, 0.00039, 0.00195) = 0.00195/0.00010 = 19.5.

Calculation of pilot risk tolerance stated in the Executive Summary. In the Executive Summary, we reported that "in more than half the flights tested here, pilots appeared to exhibit risk tolerances in excess of formal national policy goals." That figure was estimated by counting the number of flights (95) in Table A-6 (Appendix E) having maximum-risk-accepted of greater than the baseline rate 0.0001, divided by the total number of flights made (30*6 = 180). 95/180 = 52.8%. This figure can be deemed conservative, since the FAA's stated do-not-exceed GA accident rate was actually only 1.05 per 100,000 flight hours (0.0000105, FAA, 2014), and our hypothetical flight would have lasted less than 2 hr at typical GA flight speeds.

Why the ideal $R_{Max/Min} \leq 2$. $R_{Max/Min}$ *is a measure of* uniformity *in risk tolerance.* The baseline (ambient) risk is 1/10,000 (0.00010). So, for instance, if a pilot "always stays in the white," $R_{Max/Min}$ will exactly equal (1/10,000)/(1/10,000)=1.0. If a pilot choosing to fly within colored zones never varies more than one color across the three risk gradients, because each gradient's colors are on a log-2 scale, $R_{Max/Min}$ will never exceed 2.0. Note that most large values of $R_{Max/Min}$ come from transitions from white to color, particularly when the pilot flies through color on the high-risk (H) gradient. That often denotes failure to attend to the odds that the colors represent.

APPENDIX F
Underlying Pilot Heuristics and Themes

This appendix attempts to distill and summarize everything learned from the data into the final product of heuristics and themes—the ultimate goal of individual-differences analysis. Again, refer to Figure A-2 in Appendix C for the detailed color-coding scheme used in the tables below.

"Green means 'safe,' red is 'unsafe,' and yellow means 'exercise caution.'" There is very likely a strong cultural bias to cognitively code these colors, perhaps because of lifelong experience with traffic lights.

"Avoid all risk, all colored areas." There were only two pilots (20,23) who always deviated around the risk gradient, regardless of its numerical value or the stated motivation level. This was probably not due to chance ($p_{binomial} < 4.3\times10^{-7}$, assuming a 50% chance of Y/N monolithic decision style for two of 30 pilots).

Of these two, pilot 20 was not instrument rated (IR), which makes good sense, since non-IR pilots are taught to avoid weather. Oddly, pilot 23 was IR, certified commercial, rated for multi-engine, and a certified flight instructor, including instrument (CFI/CFII). So, we seem to have that one potentially suspicious data point that we should "red-flag" and perhaps discount.

Recall that uniform avoidance of all colored areas on the test implies that these pilots did not pay full attention to the lowest-risk scenario, where light green meant risk below ambient level. In Table A-7, cells superscripted "a" reflect that error.

Table A-7. Maximum acceptable risk level, by scenario.

S	IR?	Motivation "A" (baseline) Scenario risk level			Motivation "B" (higher) Scenario risk level			Tot "A" ≥ 0.00125	$R_{Max/Min}$ "A"	$R_{Max/Min}$ "B"
		Low	Med	High	Low	Med	High			
20	N	0.00010[a]	0.00010	0.00010	0.00010[a]	0.00010	0.00010	0	1	1
23	Y	0.00010[a]	0.00010	0.00010	0.00010[a]	0.00010	0.00010	0	1	1

Tot. # comparisons = 4. $\Sigma R_{max/Min}$ = 4. $\mu_{max/Min}$ **= 6.9**. 4 "confusion points" are added for cells superscripted with [a]. Therefore, ΣConfusion = 4+4 = 8. $\mu_{Confusion}$ =8/4**= 2.0.**

These two pilots probably represent fairly risk-averse individuals, so focused on safety that they missed the anomaly here. But, there is also a dimension of caution we should consider, and that has to do with the fact that, even though this was a static test with reassurances that the colored cells would not close in or expand, pilots may have still applied dynamic mental models to construct their own risk estimates. A written comment from one pilot illustrates:

> *"From my experience with weather, no risk is worth the reward. I've heard of too many accidents where a plane is flying through a 'no-risk' area that closed upon them."*

Technically, this is known as *proactive interference*, where what has been learned previously interferes with new learning (Mackintosh, 1974, pp. 478-81).

"Top-rated pilots take risks." As we just saw, uniform weather-avoidance was not a popular heuristic, even with the 10 non-IR pilots. Table A-8 shows these non-IR's age and total flight hours (TFH) where reported,[1] and risk tolerance.

Note that, since the lowest-risk scenario was designed to encourage pilots to fly through the lightest-green area, those instances should not be considered errors but, rather, possible successes.

[1] We could not require pilots to report any potentially identifying information they felt uncomfortable reporting.

Table A-8. Maximum acceptable risk, non-IR pilots only.

ID	Age	TFH	Motivation "A" (baseline) Scenario risk level			Motivation "B" (higher) Scenario risk level			Tot "A" ≥ 0.00125	$R_{Max/Min}$ "A"	$R_{Max/Min}$ "B"
			Low	Med	High	Low	Med	High			
1	24	120	0.00010[b]	0.00039	0.00195	0.00010[b]	0.00039	0.00195	1	19.5	19.5
4		100	0.00010[a]	0.00010	0.00010	0.00010[b]	0.00010	0.00010	0		
6	27	15	0.00010[b]	0.00039	0.00195	0.00500	0.00625	0.03125	1	19.5	6.3
7			0.00010	0.00039	0.00010	0.00031	0.00039	0.00195	0	3.9	6.3
15	35	200	0.00010	0.00010	0.00010	0.00010	0.00039	0.00195	0		19.5
17	45	150	0.00010[b]	0.00039	0.00010	0.00010	0.00039	0.00010	0	3.9	3.9
18			0.00010	0.00039	0.00010	0.00063	0.00156	0.00195	0	3.9	3.1
20			0.00010[a]	0.00010	0.00010	0.00010	0.00010	0.00010	0		
22	26	180	0.00500	0.00625	0.00781	0.00500	0.00625	0.00781	3		
26	20	92	0.00010[a]	0.00010	0.00010	0.00010[b]	0.00039	0.00195	0		19.5

Tot. # comparisons = 20. $\Sigma R_{max/Min}$ = 137.9. **$\mu_{max/Min}$ = 6.9**. 3 "confusion points" added for the a-superscripted cells, 3 for the b-superscripted. ΣConfusion = 137.9+6 = 143.9. **$\mu_{Confusion}$ = 7.2.**

At first glance, Table A-8 might disturb us, because it implies that non-IR pilots might seem willing to fly into actual adverse weather. Yet, consider that, except for pilot 6, these were all private pilots working on their instrument rating. They were students at a top flight school, actively pursuing advanced training, preparing themselves to deal with future risks. Theoretically, people pursuing a higher status may think and act a lot like those already holding it (Kolman, 1938).

Moreover, their risk tolerance is statistically indistinguishable from the IR pilots on any of the six risk gradient/motivation combinations (range of $p_{Mann\text{-}Whitney\ U}$ .588-.982, NS). So, any lessons to be drawn from them are no different than those to be drawn from the entire group. The lesson here is just that students working on an "advanced degree" can be expected to think and act a lot like those already holding that degree.

Only pilot 6 remains a "person of interest" from the individual differences perspective. Young, low-experience, non-IR, relatively high risk tolerance—we cannot say how many there are in the general population, but we can argue that there are probably a small number.

"Avoid risk unless there's a compelling reason." Eight pilots (6,9,11,12,16,18,21,30) always accepted greater risk under all three high-motivation "path B" conditions than they did under the three low-motivation ("path A") conditions.

Table A-9. Maximum acceptable risk level, by scenario.

ID	IR?	Motivation "A" (baseline) Scenario risk level			Motivation "B" (higher) Scenario risk level			Tot "A" ≥ 0.00125	$R_{Max/Min}$ "A"	$R_{Max/Min}$ "B"	$Risk_A/Risk_B$		
		Low	Med	High	Low	Med	High				Low	Med	High
6	N	0.00010[b]	0.00039	0.00195	0.00500	0.00625	0.03125	1	19.5	6.3	50.0	16.0	16.0
9	Y	0.00031	0.00078	0.00195	0.00063	0.00313	0.00781	1	6.3	12.5	2.0	4.0	4.0
11	Y	0.00010[b]	0.00039	0.00010	0.00250	0.00313	0.00195	0	3.9		25.0	8.0	19.5
12	Y	0.00010	0.00039	0.00010	0.00031	0.00078	0.00195	0	3.9	6.3	3.1	2.0	19.5
16	Y	0.00063	0.00039	0.00010	0.00125	0.00078	0.00195	0	6.3	2.5	2.0	2.0	19.5
18	N	0.00010	0.00039	0.00010	0.00063	0.00156	0.00195	0	3.9	3.1	6.3	4.0	19.5
21	Y	0.00010	0.00039[c]	0.00010	0.00063	0.00078	0.00195	0	3.9	3.1	6.3	2.0	19.5
30	Y	0.00010[b]	0.00039	0.00195	0.00031	0.00156	0.00781	1	19.5	25.0	3.1	4.0	4.0

Tot. # comparisons = 16. $\Sigma R_{max/Min}$ = 127.6. **$\mu_{max/Min}$ = 8.0**. 1.5 "confusion points" added for b-superscripted cells. ΣConfusion = 127.6+1.5 = 129.1. **$\mu_{Confusion}$ = 8.1.**

For these eight, a "compelling reason" was high fuel price plus being late to an engagement. Both are common reasons for taking some additional risk. The question is how much. Table A-9 shows this in both absolute and relative numbers (highlighted in gray and yellow, respectively). The medium-risk (M) gradient arguably represents the best to focus upon, given that L contained anomalies, and H appeared confusing to pilots.

Resist small motivations. The exact opposite of the "compelling reason" heuristic is that some pilots appear "motivation-resistant." Table A-4 in Appendix C indicates that, of the 90 A-B comparisons, 40 times (44%) pilots accepted the same stated risk regardless of motivation level (represented by an "=" sign). Nine pilots were uniformly motivation-resistant across all three scenarios (1,2,8,13,17,20,22,23,28). An additional five (3,4,5,19,27) were resistant across two scenarios. Was this invariant response the result of considered reason, or was it merely an automatic, trained response or heuristic at work?

Table A-10. Maximum acceptable risk level, by scenario.

ID	IR?	Motivation "A" (baseline) Scenario risk level			Motivation "B" (higher) Scenario risk level			Tot "A" ≥ 0.00125	$R_{Max/Min}$ "A"	$R_{Max/Min}$ "B"
		Low	Med	High	Low	Med	High			
1	N	0.00010[b]	0.00039	0.00195	0.00010[b]	0.00039	0.00195	1	19.5	19.5
2	Y	0.00016	0.00039	0.00195	0.00016	0.00039	0.00195	1	12.5	12.5
3	Y	0.00125	0.00156	0.00391	0.00125	0.00156	0.00781	3	3.1	6.3
4	N	0.00010[a]	0.00010	0.00010	0.00010[b]	0.00010	0.00010	0		
5	Y	0.00010[a]	0.00010	0.00010	0.00010[b]	0.00010	0.00010	0		
8	Y	0.00010[b]	0.00039	0.00195	0.00010[b]	0.00039	0.00195	1	19.5	19.5
13	Y	0.00010	0.00010	0.00010	0.00010	0.00010	0.00010	0		
17	N	0.00010[b]	0.00039	0.00010	0.00010	0.00039	0.00010	0	3.9	3.9
19	Y	0.00010[a]	0.00039[c]	0.00195	0.00016	0.00039[c]	0.00195	1	19.5	12.5
20	N	0.00010[a]	0.00010	0.00010	0.00010[a]	0.00010	0.00010	0		
22	N	0.00500	0.00625	0.00781	0.00500	0.00625	0.00781	3		
23	Y	0.00010[a]	0.00010	0.00010	0.00010[a]	0.00010	0.00010	0		
27	Y	0.00010[b]	0.00010	0.00010	0.00010[b]	0.00039[c]	0.00010	0		3.9
28	Y	0.00010[b]	0.00039	0.00195	0.00010[b]	0.00039	0.00195	1	19.5	19.5

Tot. # comparisons = 28. $\Sigma R_{max/Min}$ = 208.3. **$\mu_{max/Min}$ = 7.4**. 7 "confusion points" added for a-cells, 5.5 for b-cells, 1.5 for c-cells. ΣConfusion = 208.3+14 = 222.3. **$\mu_{Confusion}$ = 7.9.**

For the most part, Table A-10 shows that this heuristic encourages safety but does not guarantee it. In the case where the low-motivation risk tolerance was high, it means that the higher-motivation risk tolerance is also high, because the heuristic is not based on any particular understanding of probability, as evidenced by the high levels of $R_{Max/Min}$ for those cases.

"Always pick green." Another intriguing heuristic involved pilots 1,2,8,28,29, who always flew through green areas, regardless of what numerical risk was actually associated with those areas.

Based on their high $R_{Max/Min}$ ratios and a tendency to misunderstand the lowest-risk condition, we can argue that these were the pilots most confused about what the risk odds meant. In fact, they constituted the right-hand tail in Figure A-8. They apparently meant not to take high risk, but did so accidentally in not fully understanding the risk gradients.

Table A-11. Maximum acceptable risk level, by scenario.

ID	IR?	Motivation "A" (baseline) Scenario risk level			Motivation "B" (higher) Scenario risk level			Tot "A" ≥ 0.00125	$R_{Max/Min}$ "A"	$R_{Max/Min}$ "B"
		Low	Med	High	Low	Med	High			
1	N	0.00010[b]	0.00039	0.00195	0.00010[b]	0.00039	0.00195	1	19.5	19.5
2	Y	0.00016	0.00039[c]	0.00195	0.00016	0.00039[c]	0.00195	1	12.5	12.5
8	Y	0.00010[b]	0.00039	0.00195	0.00010[b]	0.00039	0.00195	1	19.5	19.5
28	Y	0.00010[b]	0.00039	0.00195	0.00010[b]	0.00039	0.00195	1	19.5	19.5
29	Y	0.00016	0.00078	0.00391	0.00010[b]	0.00039	0.00195	1	25.0	19.5

Tot. # comparisons = 10. $\Sigma R_{max/Min}$ = 186.5. **$\mu_{max/Min}$ = 18.7.** 3.5 "confusion points" added for b-cells and 1 for c-cells. ΣConfusion = 186.5+4.5 = 191.0. **$\mu_{Confusion}$ = 19.1.**

The high-risk gradient was confusing. If we set the threshold slightly more liberally at 1/640 (0.00156), the only pattern that emerges is a lot of stated risk tolerance on the high-risk H scenario (11 cases on the baseline "A" motivation level and 22 cases on the higher "B" motivation level). Again, this supports the notion that the H scenario was confusing to pilots, with the risk values (the odds) probably conflicting with their prior biases of what colors *should* mean.

Reflectivity range (dBZ)	<30 dBZ	30-40 dBZ	>40-50 dBZ	>50 dBZ
WSR-88D terminology	Light	Moderate	Heavy	Extreme

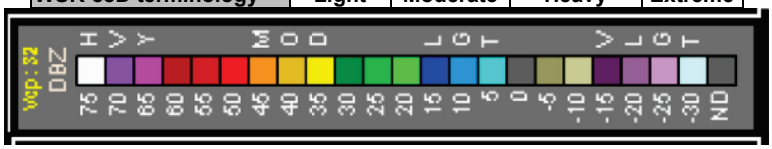


Figure A-9. NWS WSR-88D Weather Radar Precipitation Intensity scale and Weather Radar Echo Intensity Legend.

Figure A-9 shows the latest intensity color scheme, based on the latest FAA Advisory Circular on thunderstorm avoidance (AC 00-24C, FAA 2013) available at the time of this writing

We can hypothesize that the WSR-88D Intensity Legend constituted another example of proactive interference. Pilots already familiar with "safe colors" versus "unsafe colors" probably had difficulty replacing their existing notions for the arbitrary new ones we asked them to temporarily learn.

No risk too great. High risk tolerance is a tricky category to assess because it could reflect confusion, not true risk tolerance; 22 pilots who accepted at least one risk equal to or exceeding 1/512 (0.00195).

Increasing our threshold to 1/256 (.00391) eliminates most of this "noise," leaving just two pilots (6,22) with areas of uniformly high risk tolerance.

Table A-12. Maximum acceptable risk level, by scenario.

ID	IR?	Motivation "A" (baseline) Scenario risk level			Motivation "B" (higher) Scenario risk level			Tot "A" ≥ 0.00125	$R_{Max/Min}$ "A"	$R_{Max/Min}$ "B"
		Low	Med	High	Low	Med	High			
6	N	0.00010[b]	0.00039	0.00195	0.00500	0.00625	0.03125	1	19.5	6.3
22	N	0.00500	0.00625	0.00781	0.00500	0.00625	0.00781	3	1.6	1.6

Tot. #comparisons = 4. $\Sigma R_{max/Min}$ = 29.0. **$\mu_{max/Min}$ = 7.3.** .5 "confusion point" added for b-cell.
ΣConfusion = 29+.5 = 29.5. **$\mu_{Confusion}$ = 7.4.**

And only one was uniformly high across both motivation conditions (22). Both happened to be non-IR, but that could be coincidental ($p_{binomial}$ = .11, NS).

Overt risk calculators. As stated earlier, pilot 22 (who keeps turning up in a number of our heuristics) was the only one to write his stated risk threshold (.01) on the test sheets. He was therefore a high risk tolerator and, we also saw earlier, a private pilot working on his instrument rating.

Table A-13. Maximum acceptable risk level, by scenario.

ID	IR?	Motivation "A" (baseline) Scenario risk level			Motivation "B" (higher) Scenario risk level			Tot "A" ≥ 0.00125	$R_{Max/Min}$ "A"	$R_{Max/Min}$ "B"
		Low	Med	High	Low	Med	High			
22	N	0.00500	0.00625	0.00781	0.00500	0.00625	0.00781	3		

Tot. #comparisons = 2. $\Sigma R_{max/Min}$ = 3.2. **$\mu_{max/Min}$ = 1.6.** **$\mu_{Confusion}$ = 1.6.**

The idea of high risk tolerance coupled with high understanding of probabilities is certainly interesting, although we should not try to generalize on the basis of a single case.