DOT/FAA/AM-05/7

Office of Aerospace Medicine
Washington, DC 20591

The Influence of Visibility, Cloud Ceiling, Financial Incentive, and Personality Factors on General Aviation Pilots' Willingness to Take Off Into Marginal Weather, Part I: The Data and Preliminary Conclusions

William Knecht
Howard Harris
Scott Shappell
Civil Aerospace Medical Institute
Federal Aviation Administration
Oklahoma City, OK 73125

April 2005

Final Report

U.S. Department
of Transportation

**Federal Aviation
Administration**

NOTICE

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-05/7		

4. Title and Subtitle	5. Report Date
The Influence of Visibility, Cloud Ceiling, Financial Incentive, and Personality Factors on General Aviation Pilots' Willingness to Take Off Into Marginal Weather, Part I: The Data and Preliminary Conclusions	April 2005
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Knecht W, Harris H, Shappell S	

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

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

15. Supplemental Notes

Work was accomplished under approved task AM-HRR-522-04.

16. Abstract

Adverse weather is the leading cause of fatalities in general aviation (GA). In this research, influences of ground visibility, cloud ceiling height, financial incentive, and personality were tested on 60 GA pilots' willingness to take off into simulated adverse weather. Results suggested that pilots do not see "weather" as a monolithic cognitive construct but, rather, as an interaction between its separate factors. This was supported by the finding that the multiplicative statistical effect of visibility and ceiling could better predict takeoff than could the linear effect of either variable considered separately. Also found was a statistical trend toward financial incentive being able to predict takeoffs. However, none of the 10 personality tests (incorporating over 500 separate response items) could predict takeoff.

17. Key Words	18. Distribution Statement
General Aviation, VFR-Into-IMC, Adverse Weather, Marginal Weather, Accidents, Fatalities, Pilot Personality, Financial Incentive	Document is available to the public through the Defense Technical Information Center, Ft. Belvior, VA 22060; and the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	44	

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

The Influence of Visibility, Cloud Ceiling, Financial Incentive, and Personality Factors on General Aviation Pilots' Willingness to Take Off Into Marginal Weather, Part I: The Data and Preliminary Conclusions

INTRODUCTION

In the 1934 movie *Bright Eyes*, a six year-old Shirley Temple plays a brave little orphan, ever-cheerful and indomitable in the face of fell circumstance. Her godfather/pilot, James Dunn, desperately wants to adopt the stalwart girl, but lacks the money for legal fees. So he accepts a job no other pilot will dare, delivering cargo through a hair-raising storm. In the end, love conquers adversity, and everyone lives happily ever after.

This movie is far more than fluff. It makes a point critical to aviation psychology, namely that people take risks that seem to defy logic. They have their reasons; we may just not see them right away.

We know that some pilots take risks with weather, the question is why. U.S. general aviation (GA) typically averages far higher accident and fatality rates than both commercial aviation and automotive transport. Both in relative and absolute numbers, more people die from accidents in GA than in commercial aviation. Yet, GA safety research invariably receives less attention than its higher-profile commercial aviation counterpart. Figure 1 illustrates the relative fatality rates.

In this sample, GA averages a fatality rate 223 times higher per-passenger-mile than commercial aviation. Data were derived by us from National Transportation Safety Board (2002, Tables 10, 6) and U.S. Bureau of Transportation Statistics (2001, Tables 1-31, 2-1). Fatalities were simply divided by official estimates of total miles flown. NTSB data exclude fatalities due to terrorism and include those due to collateral damage (e.g., bystanders hit by debris).

What are the factors that contribute to this higher fatality rate for GA? A leading candidate is adverse weather, technically known as flight into instrument meteorological conditions (IMC). Data derived by us from National Transportation Safety Board statistics (1995-1997) support this assertion, showing IMC implicated in approximately 32% of GA fatalities (Appendix A details the derivation). Many of these inadvertent violations happened during flight-into-IMC when pilots rated only for visual flight rules (VFR) found themselves caught in IMC (Goh & Wiegmann, 2002).

A number of authors have reviewed and investigated external (environmental) and internal (perceptual/cognitive) factors leading to weather-related decision making errors (Adams, Koonce, & Hwoschinsky, 2002; Hunter, 2002a,b; O'Hare, 1990; O'Hare & Owen, 1999; O'Hare, Chalmers, & Scuffham, 2003; Wiegmann, Goh, & O'Hare, 2002). To summarize cogent points:

1) Weather-related situation risk may escape perceptual or cognitive appreciation
   a) Perceptual:The physical situation of risk may literally not be perceived.
   b) Cognitive: It may be physically perceived but psychologically underestimated for some reason(s).
      i) The base rate of risk for that situation may be unknown.
      ii) Knowledge of weather-related flight rules may be inadequate.
      iii) The probability of disjunctive events may be underestimated (explained below).
      iv) The probability of conjunctive events may be overestimated (explained below).

2) Weather risk may be appreciated—and knowledge of flight rules adequate—but cognitive or emotional factors may overrule apparent good judgment.
   a) A pilot may overestimate his or her piloting ability.
   b) Sunk-cost effect may promote flight continuance (explained below).
   c) Fractional anticipatory goal response may promote continuance (explained below).
   d) A mental risk/benefit equation may be operating, in which case the estimated benefit of continuing into IMC may outweigh the estimated risk.
      i) Objective risk is the net probability of and severity of loss, given all parameters of the situation.
      ii) Perceived risk is one's cognitive/affective estimate of objective risk.
      iii) Benefit may be positive reinforcement (R^+) or negative reinforcement (R^-)[1].

To illustrate Point 1a, consider the difficulty of determining ground visibility during fog. In the absence of reference objects of known size, there are simply no reliable perceptual cues on which to base an accurate visibility estimate. Accurate risk assessment fails because of missing physical information. In some cases, the information may be there, but the cues may simply be missed due to lack of knowledge of what they are or inexperience at judging them. This kind of situation can often be handled with training (Wiggins & O'Hare, 2003).

In Point 1bi, *base rate* means the statistical likelihood of some accident happening under given circumstances with an average individual (O'Hare, 1990). For instance, how many of us know the average odds of crashing, given some complicated flight situation? Often, the only information we have is very crude, such as "fatalities per million miles." This tells us little about the average risk involved with any given *kind* of weather (Oppe, 1988). Because there are so many kinds of situations, it becomes nearly impossible to truly know the base rates for every one. Therefore, base rate estimation often devolves into an educated guess based on inadequate information.

Point 1bii involves inadequate pilot knowledge of weather-related flight rules. Rules are the "yardstick" defined by knowledgeable authority, by which individuals can gauge the risk of their own behavior, even though they may never have experienced the catastrophes that motivated those rules in the first place. And, if we deeply consider the meaning of rules, we realize that good rules are actually statements embodying prior knowledge about accident base rates. If clear knowledge of those rules is missing, so is the yardstick—the accumulated statistical knowledge about how risky a given situation is likely to be—and risk assessment is reduced to guessing based on lack of appropriate knowledge.

Point 1biii refers to a highly common psychological bias. In *disjunctive* risk assessment, we have a bias to *under*estimate the global chance of something happening *at least once* when a chance process is repeated over and over (Kahneman & Tversky, 1982, p. 15). For instance, what is the actual chance of crashing at least once, if we fly six times, each time having a base rate of 1/30 for the crash risk? The answer is $1 - (29/30)^6 = .184$. Many people would estimate it as less probable than that, even when given the true base rate. The point is that most of us do not understand (much less calculate) true probabilities. Instead, we do seem to "guesstimate" odds, particularly the odds of repeatedly taking risks without getting caught. And, included in this guesstimation is some kind of human tendency to be somewhat optimistic with regard to certain kinds of risk.

Point 1biv refers to a similar bias to *over*estimate the global chance of something happening *all the time* when a chance process is repeated over and over (Kahneman & Tversky, 1982, p. 15). For instance, what is the chance of always flying successfully, if we fly six times, each time having a 1/30 chance of crashing? Notice that this example is the logical complement to the previous one. The answer is $(29/30)^6 = .816$. Many people would estimate it as more probable than that, indicating a bias toward optimism with respect to this particular kind of risk.

Point 2 has to do with risk that is appreciated but is mentally overridden or overruled for some reason. Shappell and Wiegmann (2003) estimated that GA fatal accidents are four times more likely to be the result of rule violations than are non-fatal accidents. Some of these violations are unconscious, due to the factors listed in Point 1. The rest are conscious. If one knows the yardstick but chooses to ignore it, then we are faced with a much different situation, one calling for a very different strategy to correct.

One simple reason that people ignore rules may be 2a, often called *confidence calibration*. In other words, how well does one's own confidence in oneself truly match what one is capable of doing? The plain fact is that most people put more faith in their own abilities and judgments than actual performance warrants (Baranski & Petrusic, 1994). Pilots are no exception (Wilson & Fallshore, 2001). Most people tend to idealize their own skill level, for one thing. For another, we often tend to think we can put forth our best effort at any time, ignoring the fact that instantaneous skill is affected by ever-changing factors such as physical health, level of sleep or rest, nutrition, and so on. Nobody functions 100% effectively 100% of the time.

Sunk-cost effect (Point 2b) is the degree of investment we have in some situation, literally what we have "sunk into it." The importance of sunk cost has to do with prospect theory (Kahneman & Tversky, 1979). Prospect theory states that, everything else being equal, certainty carries more mental weight than uncertainty. For example, if one has ten hours invested in a flight, and encounters bad weather ten minutes from destination and decides to divert, the risk of crashing in that last ten minutes is uncertain—but the loss of getting home, and of all that work invested, is 100% certain. The sunk costs are ten hours' work plus the positive expectations associated with arrival. And, the greater this sunk cost, the more likely one may be to press on to the destination. VFR-into-IMC could be due to sunk-cost effect, although the research is divided on this point (Goh, Wiegmann, & O'Hare, 2002).

A simpler, older proposal, put forth by behavioral theory, is Point 2c, "fractional anticipatory goal response." This is the behaviorist's explanation for why a rat tends to run faster the nearer it approaches the goal. It is a stimulus-response principle thought to result from pre-goal cue stimuli associated by generalization with past goal rewards (Hull, 1932, 1934). Apparently there is no research relating VFR-into-IMC and fractional anticipatory goal response. This is probably due to the almost-total ascendancy of cognitive theory in aviation psychology versus a steep decline in behavioral theory. That situation is unfortunate, given that many of the fundamental principles of animal behavior remain universal and unchanged. Human attitudes, motivations, and inhibitions impact behavior in addition to rational, analytical decision making.

Point 2d, the risk/benefit equation, is a plausible theory as well. The basic notion is that our brain, in combination with learning and emotions, instantiates some kind of "computational algorithm" that attempts to compute situational risk, then weighs that risk (or cost) and benefit, and finally stimulates us to make a choice between, say, behavior A or B:

If $(benefit + noise_{benefit}) > (risk + noise_{risk})$ Then A
Else B

"Benefit" here is an intangible mental construct based on tangible gain. Similarly, "risk" is an intangible mental construct based on tangible loss. More specifically, benefit and risk are based on the estimated success probabilities of a specific behavior or course of action. Both risk and benefit are person-specific and situation-specific. The two kinds of noise represent various sources of error, both about benefit and risk.

One explicit goal of pilot training is to correctly calibrate risk-taking behavior. This means to fly when we should fly and to not fly when we should not. It is not simply about avoiding all risk. After all, we could eliminate all accidents by eliminating all flight. But the practical goal of learning to be proactive about safety is rather to learn how to sense the cues to risk, along with rules teaching us how to compare our subjective sensations and estimations to some objective standard of risk. Finally, we also have to consider the recuperative aspect, to learn skills to get ourselves out of trouble, should we find ourselves inadvertently caught in some risky situation.

The current study concerns this problem of what motivates pilots to risk VFR-into-IMC flight. We chose to ignore the recuperative aspect of risk, focusing instead on how pilots do get into trouble in the first place. So far, this has usually been studied in situations where the pilot is airborne. Here, we examined a complementary situation where, given marginal weather right from the onset, we measured whether or not the pilot would take off.

This paradigm not only looked at another aspect of the overall problem but also simplified the baseline motivational situation in three ways. First, it minimized perceptual miscalculation (Point 1a), since we actually gave the pilots a weather report containing more accurate current information than what they could derive through their own senses. Second, it minimized the issue of disjunctive and conjunctive events (Points 1biii, 1biv), since the experiment involved only one trial per subject. Third, it simplified the sunk-cost motivation (Point 2b), since zero time or effort was invested prior to the adverse weather appearing. Any baseline sunk costs had primarily to do with elapsed time after the experiment began. Pilots might arguably have had a sunk cost if they showed up expecting to fly and had to give up that expectation, should they choose to stay on the ground. However, we avoided that cost by telling them they would be allowed to fly after the experiment, no matter what their experimental decision. The net result of all these simplifications was to let us focus more on Point 2d, the "risk-benefit equation."

METHOD

Participants and Demographics

Sixty GA pilots participated in this study (53 males and seven females, aged 18-69; median 23.5, mean 26.0, SD 8.4). All subjects were instructed that participation was voluntary, and each signed an informed consent form after a formal briefing of the study parameters. In addition, each subject completed a basic demographic questionnaire before the study began and a debriefing form when finished (see Appendix B).

Pilots had a median flight experience (GA plus commercial experience) of 2.3 years (mean 4.2, *SD* 7.5, range .25-48.25). Median flight hours (GA plus commercial) were 183.5 (mean 755, *SD* 2604, range 35-20000). Exactly one-half of the pilots were instrument-rated.

Means and standard deviations (SD) were skewed by the presence of a relatively small number of older pilots with a great deal of experience (see Appendix C). Therefore, medians were sometimes the more informative estimates to consider in discussion and distribution-insensitive statistics the more appropriate to use during analysis. Appendix D details all variables examined. Additionally, a data-conditioning technique (winsorization) was applied to correct for outliers during the analysis. (Appendix E).

Experimental Design, Independent and Dependent Variables

The aim of this study was to examine some of the perceptual, cognitive, and affective processes used by GA pilots when deciding whether or not to take off into marginal weather. This necessitated finding correlates of takeoff behavior (predictors). This, in turn, required manipulation of conditions likely to induce takeoff.

Therefore, three independent variables (IVs) were set up to assess how each (plus their interactions) would subsequently influence a single dependent variable (DV)—*Takeoff Decision*.

These three independent variables were all factors external to the pilot, namely

1) Three levels of marginal ground visibility
   a) 1 statute mile (sm)
   b) 3 sm
   c) 5 sm

2. Two levels of marginal cloud ceiling
   a) 1000 ft
   b) 2000 ft

3. A financial incentive for takeoff
   a) Straight salary ($17 per hour)
   b) Straight salary plus a $200 sliding-scale bonus (described below)

Visibility and cloud ceiling are two familiar and important components of weather normally influencing pilot decision making. Financial incentive is a real-world commercial factor that, to our knowledge, has not been closely examined as a modulator of risk behavior in the GA setting. The goal here was to cover a wide enough range of weather conditions to have some in which almost no pilots would take off, and some in which practically all would. To that end, our selection of weather values was heavily influenced by the modal values for personal local-flight daytime minima reported in Hunter's survey of 6,735 GA pilots (1995, pp. 22-23, Table 22, Q75 and Q79). This gave us plausible reference points for our two weather variables.

The financial variable was set to be appealing to our modal pilot. Demographics from past experiments showed that our usual subject's age was circa twenties, with income modest, since a considerable number were students from local flight schools. Two levels of financial incentive were chosen. Low Financial Incentive consisted of a base salary of $17/hr for all 60 participants, whether they decided to take off or not, and was the only financial reward in that condition. In contrast, High Incentive pilots were offered the base salary, again regardless of takeoff, plus an additional bonus of a $200 for immediate takeoff. This bonus was engineered to decrease by $50 every 30 minutes, being billed as an offer made by an oil company, which supposedly needed a critical drill bit delivered as soon as possible to get interrupted production back on-line. Subsequent payments were made by check and mailed to each participant.

A potential confound arose, in the sense that the reinforcement value of money varies in relation to the participant's need for it. Unfortunately, there was no good way to control for this ahead of time, other than by random assignment to groups. Trying to assign subjects to groups based on either real-time or post hoc sample-matching sometimes runs as much risk of introducing bias as it does of eliminating it, so we elected to assign group membership as a random function of whenever a given person happened to walk through the door.

Certain personality factors might predict takeoff into adverse weather. Specifically measured here were:

- Prior aviation risk-taking behavior
- Anxiety
- Impulsivity
- Risk orientation
- Sensation-seeking

because each has previously shown to bear some relation to risk-taking in other domains. Table 1 shows a brief list of the personality instruments administered, what they measured, and primary references to each in the open literature.

One of the planned statistical methods (logistic regression) would allow treating the scores on each of these instruments as an IV (technically a "predictor" in this context, because the variable was not actually subject to experimental manipulation). The strength of the statistical relation between each pilot's predictor scores and subsequent yes-or-no takeoff could then be measured and used to infer relations between decision making and the putative factors involved.

Apparatus and Procedures

All participants were given instructions to imagine themselves as pilots working for a private air cargo company currently contracted to transport a drill bit from Page Municipal, an uncontrolled airfield near Oklahoma City, OK, to an oilfield near Amarillo, TX. Because the field was uncontrolled, no weather condition pilots faced was a true technical violation of FAA rule or procedure.

Pilots were instructed to consider themselves working on a per-hour basis with their base salary to be paid whether they chose to fly or not. The company itself was described as being in good financial shape, with corporate attitudes favoring the pilot as the ultimate tactical decision

maker. Pilots were instructed that their flight would have to be made under VFR because the aircraft was not instrument-certified. They were strongly encouraged to treat their situation as close to reality as possible. They were also told that, no matter what their takeoff decision, their identities would be kept strictly in confidence.

Each pilot received a single trial containing one combination of *Visibility x Ceiling x Incentive*. The net result was a between-subjects, 3x2x2 design with 12 cells, each containing five pilots. Cells had either two or three VFR-rated pilots, with the remainder being instrument-rated.

Each pilot received written instructions and an 8.5 x 11-in roadmap-style map showing major regional cities and roads, with the weather front overlaid. At the bottom of this map was a Meteorological Aerodrome Report (METAR)-style text weather report. Pilots also saw a visual simulation of the stated weather conditions shown from taxiway level using the high-fidelity Advanced General Aviation Research Simulator (AGARS) of the Civil Aerospace Medical Institute. The simulator was configured as a Piper Malibu, a high-performance, single-engine aircraft. Finally, pilots received a standard sectional map showing airports, terrain features, and obstacles.

Pilots were given as much time as they required to complete their flight planning before countdown was initiated on the financial incentive. This typically took 5-10 minutes. Once finished planning, they were asked to give their initial go/no-go decision. Pilots choosing to fly were then asked to take off in their assigned weather and fly for a period of time equivalent to their clearing the weather front (about 20 minutes). Those choosing to stay grounded were allowed to wait and see if the weather would change and were given an updated situation report every 30 minutes until a total of two hours had elapsed. However, during this time, the weather did *not* change, and this "opportunity" was primarily put in place to see if boredom and/or frustration would exert any sort of influence. Particularly in the High Incentive case, the time-dependent loss of the anticipated takeoff bonus (a secondary reinforcer) was expected to produce considerable frustration (Amsel, 1958; Hull, 1932) which could be avoided by taking off.

The experiment was terminated either after two hours for pilots electing not to take off, or after 20 minutes of flying for those airborne. The assumption was made for high-incentive pilots that, if they resisted takeoff until the bonus went to zero at two hours, they would probably continue to resist indefinitely. Low-incentive pilots electing to remain on the ground were therefore cut off at 2 hours to equilibrate them with their high-incentive counterparts.

RESULTS

The strategic aim of this study was to examine weather-related decision making. By exposing pilots to conditions of marginal visibility, we hoped to separate a large group of pilots into two groups—ones who would fly into bad weather versus ones who would not—and then study those two groups for critical differences. Those differences, both in personality factors and in reaction to specific environmental factors, might grant insight into the mind of the pilot.

Two analytical techniques were used, chi-square and logistic regression. Given appropriate comparisons, both give similar results. Chi-square has the advantage in simplicity and universality of use. Logistic regression has the advantage of being able to interpret statistical interaction terms. However, extreme care has to be exercised because logistic regression is a complicated procedure with great room for error (Appendix F gives details).

Regression involves the search for predictors—measurable factors that predict some outcome of interest. Appendix D gives the complete list of predictors used in this study. The tactical aim of regression is to create and test models—purposeful simplifications of reality used on a higher level to infer causes of behavior.

To summarize how we judged a model's performance or quality, our major numerical criteria for judging regression model quality were a) Wald p; b) predictivity; and c) Nagelkerke R^2. In logistic regression, model reliability is estimated with the Wald statistic. Wald p is analogous to regular p values in other statistics and estimates how likely we would be to get different results, were we to test a new group of pilots. Predictivity is expressed, first of all, by a raw percentage of cases successfully predicted (number of cases correctly predicted divided by the total number of cases). Second, predictivity is also expressed by R^2, an estimate of explained variance. SPSS gives two kinds of R^2 in logistic regression, Cox and Snell R^2 and Nagelkerke R^2. The Cox and Snell estimate addresses a model's total theoretical deviation from perfect data fit, while the Nagelkerke estimate addresses the fact that certain models, in theory, cannot explain all the variance, even given their best fit. Therefore, the Nagelkerke estimate tries to address "explained variance divided by total explain*able* variance." This usually results in a higher, more liberal numerical value for R^2 but also one more arguably pragmatic. For this reason, we present only Nagelkerke R^2 in this report.

Finally, we needed to assign a higher credibility to simple models—and ones based on theory—than just to models with high numbers for p, predictivity, and R^2. There are both logical and statistical reasons for doing

this. We all know Occam's Razor, the old adage that, all things being equal, the simplest explanation is usually the correct one. Highly complex theories often turn out to be wrong. This applies to regression in the sense that, given two models with equal *p*, predictivity, and R^2, the model based on sounder theory AND containing fewer predictors generally proves more stable on retest (Tabachnick & Fidell, 2000, ch. 12).

Analysis of the Full Data Set (N=60)

This experiment began with a total of 83 measurements on various environmental, motivational, and psychological aspects of both the situation and the pilots, some of which would reasonably be better than others at predicting takeoffs. Two of these predictors were independent variables involving the physical environment (*Visibility, Ceiling*) and a third was psychological/motivational (*Financial Incentive*). That left 80 descriptive variables, both demographic and from personality tests. Most involved numbers which could vary continuously (or reasonably so). A few involved measurement of discrete categories, usually the binary choice yes/no. Admittedly, categorical data can be confusing, requiring careful thought and attention in setting up proper contrasts (Appendix E). Nonetheless, all 83 candidates started out as plausible predictors for the discrete (yes/no) dependent variable of *Takeoff*.

Before speaking to specifics, four potential criticisms must be addressed, namely boredom, instrument rating, multicollinearity, and accidental group differences. First, during a two-hour experiment in which pilots had to essentially sit and do nothing while waiting for weather reports, we might expect boredom to interact silently and uncontrollably with the financial bonus to induce the high-incentive pilots to take off more quickly than the rest. However, boredom demonstrated no evidence of being a factor in this experiment. In fact, among pilots who chose to take off, those who were offered the takeoff bonus actually took *longer* to make their decisions than did low-incentive pilots (mean decision latency was 31.6 vs. 22.9 minutes respectively, $t = -.61$, $p = .55$, two-tailed, unequal variance). From this we surmised that, if boredom did have any effect, it was distributed about equally between the two financial incentive groups.

A second concern could be the questionability of testing instrument-rated pilots in a situation demanding VFR flight. While the experimental groups were set up to distribute the two types as equally as possible among the various treatment conditions, the question could still be reasonably asked. If instrument rating were a contaminating factor, then we would expect the two groups to differ in takeoffs when sorted by rating. But no such effect was evident. Instrument-rated pilots had slightly fewer takeoffs, but not significantly so ($t = 1.29$, $p = .20$, two-tailed, unequal variance). From that we surmised that using instrument-rated pilots was acceptable for this study.

The third concern had to do with multicollinearity. Multicollinearity is basically about two or more predictors measuring the same experimental factor. This can artificially inflate a model's performance (Tabachnick & Fidell, 2000). Given the large number of candidate predictors examined here, it made sense to look at their correlations, to ensure that they were not merely measuring the same factor more than once. Appendix G shows correlation matrices for the small number of significantly correlated predictors. These show nothing surprising. Factors that should be correlated were correlated, specifically those measuring related aspects of the same construct. Likewise, factors that should not be highly correlated were not. Armed with that information, we then simply needed to follow standard procedure and avoid models that contained highly correlated factors.

The fourth concern (accidental group differences) was certainly valid. Certainly, one cell in the analysis might end up with, say, a significantly higher mean number of flight hours than other cells. This might exert some unknown, unwanted effect on takeoffs. However, little could be done in a study of this type to avoid this problem because there were simply too many predictors being examined. Typically, with a small predictor set, one might counterbalance subjects on values for every predictor and every cell. But, with dozens of predictors, that kind of counterbalanced subject assignment to cells is impossible. So standard operating procedure in a case like this is to rely on random assignment to wash out most unwanted cell-mean differences and, for the most part, it does.

Descriptive Statistics—Demographics and Debrief Data. Table 2 summarizes the demographic data. As is evident from the medians in Table 2, most of these pilots were fairly new. Most were males (53/60), and recall that exactly half were selected to be instrument-rated.

The primary statistical problem in the demographic data was the presence of outliers, a small number of extremely deviant scores (defined here as values greater than 3.0 SD above or below the mean). This was immediately evident from looking at the means and standard deviations. When the two are nearly equal, this implies logical absurdities such as the possibility of having negative flight hours. In reality, this was an artifact of outliers. Outliers are common in aviation studies, particularly in the demographic data. In our case, the outliers were older pilots with a large amount of flight experience. As mentioned, Appendix C details the probabilities of skew

normality (p_{zskew}) and kurtosis normality (p_{zkurt}), far less than .001 in most cases, indicating severe non-normality of these particular distributions.

Fortunately, there are well-accepted ways to deal with outliers. These data responded well to winsorization (Appendix E). Virtually all outliers were eliminated with only minimal change to the data's ratio-scale structure.

Regression analysis can be problematic when a large number of predictors are examined. First, there has to be a limit to the number of predictors allowed in each model. Generally between 3-10 cases (here meaning pilots) are needed per model predictor (R. A. Foster, personal communication, Jan. 15, 2004). Second, the sheer number of candidate predictors present at the beginning of the analysis can be a problem (Foster & Stine, 1998). Finally, a number of widely different models may show roughly the same predictivity and explained variance (R^2). Under such "shallow model" conditions, it may be hard to logically defend any one model as best.

Rather than dwell on complex statistical issues at this time, we address them in a separate technical report entitled *Pilot willingness to take off into marginal weather. Part II: Antecedent overfitting with forward stepwise logistic regression*. For the purposes of this discussion, the reader can be assured that the results presented here (Part I) were coordinated with, and subjected to the rigorous scrutiny detailed in, the Part II report. We now turn to the analysis of the full data set.

Individual Predictors. In this analysis, each predictor was first run separately in a logistic regression model including a constant. This was done to get a feel for the performance of each predictor before doing more complex modeling. One reason for including the constant was pragmatic. SPSS tends to stall without it unless a model has at least one extremely strong predictor to start with. The other reason was theoretical. The constant-only model is basically a guess based on the after-the-fact success ratio (i.e., takeoffs/total runs). The importance of this will become increasingly clear as we discuss the models themselves.

Appendix H contains the complete results of single predictors for the full data set. At this point, it was unnecessary to test both conditions of binary categoricals because single-predictor analysis yields the same result no matter which contrast is tested. Summary results for our three primary IVs are listed in Table 3.

In this particular case, chi-square and logistic regression gave identical p values. These revealed that the effect of our three primary IVs was not remarkable across the full data set. *Financial Incentive* was the strongest, yet its nominal Wald p was only .07, R^2 = 7%, and the *Financial Incentive* term only increased predictivity 3.4% (two cases) above performance yielded by the model's constant alone. In other words, none of our three IVs, when examined by itself, performed much better than an educated guess. Among the remaining 80 predictors, only the *Venturesomeness* scale of the Multidimensional Personality Questionnaire (MPQ) was even remotely notable (p=.088, R^2 = 8%). Moreover, these numbers were uncorrected for the number of comparisons made (experimentwise Type I error).[2] If they had been corrected, the effective p-values would have been even less significant.

Interactions. Interactions in regression reflect synergy between predictors. Logically, they ask "What factors working *together multiplicatively* affect takeoffs more than the same factors would if merely added together?" Table 4 lists the results for two- and three-way interactions between our IVs in the full data set.

A significant interaction implies that the effect of one variable is different at different levels of the other variable(s). To better visualize this, Figure 2 shows the actual raw takeoff frequencies split by *Financial Incentive*. In the Low Incentive group, note the cell at 5 miles visibility and 2000-ft ceiling. This has four takeoffs. This illustrates how interactions work. It as if *Visibility* and *Ceiling* team up to produce an outsized effect at that particular combination.

At the level of main effects, trends emerged here as expected. Fewer pilots generally elected to take off as visibility, ceiling, and financial incentive decreased. However, as Figure 2 shows, inter-cell variability kept these trends from being statistically significant—particularly the presence of four pilots flying at the very worst conditions of visibility and ceiling.

One thing especially interesting about these interactions was the notion that *Visibility x Ceiling* very probably constituted a cognitive "chunk." In other words, it was logical that many pilots cognitively considered *both factors simultaneously and synergistically*, suggesting that "the whole was greater than the sum of the parts." This probably denoted a key underlying mental representation driving some of these statistical interactions. Inside the minds of pilots, "weather" was probably not just visibility, not just ceiling—but both at the same time (along with other factors).

As for the personality variables, examining all possible interactions between all predictors would have been tedious and unwise, inflating the Type I error rate ever higher. There were 83(83-1)/2 = 3403 possible two-way interactions alone. Clearly, it made good theoretical sense to limit examination mostly to interactions involving the IVs. So, operating on the assumption that *"Weather"* ≅ *Visibility x Ceiling*, we therefore inspected weather-by-everything-else, plus financial incentive-by-everything-else.

This turned out to be surprisingly uninformative. Thirty-one models of the type *Vis x Ceil x Predictor Y* and *Financial Incentive x Predictor Y* had nominal significances of < $p = .10$ (uncorrected). This implied four things. First, there were obviously a large number of models, roughly equal in reliability. Second, after correcting merely for multiple comparisons, none of these would be traditionally significant. Third, most of the predictivity was coming from the IVs themselves ("*Weather*" and *Financial Incentive*)—not from *Predictor Y*—because the reliabilities and R^2s were not much better than the IVs alone (plus constant, of course). In many cases, the additional predictor *decreased* model performance. Fourth, even the very most-reliable model without missing values, *Weather x MPQ Aggression* ($p = .008$), only showed predictivity of 70% (versus 65% for *Weather* alone, and 58.3% baseline for constant alone) and R^2 of .186 (versus .101).

Preliminary Conclusions for Full Data Set. So what did these full-set models mean? In all candor, we must not place too much stock in them. While predictivities in the 80% range might seem impressive, less than half the outcome variance was explained. We have to keep in mind that the model constant alone is always capable of predicting at better than a chance (50%) level. Nor was correction made for the number of models tested (which would have pulled significance levels down even further). Finally, there were quite a few other models having results almost as good as the ones examined. So we ought to interpret these results conservatively and stick to relatively high-level conclusions that express commonalities across different models.

On the surface, the results looked confusing. There were many models, often with similar performance. But, in truth, it was not hard to understand the big picture: a) There were undoubtedly predictor effects going on, but they were multiple effects, pulling in different directions; b) Pilots have multiple, simultaneous motivations, and the effects of environment are filtered through each pilot's unique, multi-dimensional personality; c) So the net effect on behavior is complex. Not altogether unpredictable—just complex.

Another way of saying this is that pilots were apparently not just being swayed by the independent variables. *They were interpreting the visibility, ceiling, and financial incentive in various combinations through the medium of their own individual personalities, which were highly variable and themselves comprised of interrelated components, some of which acted synergistically.*

Another high-level realization was that the more reliable model predictors could be usefully categorized into either "go" or "no-go" factors. The beta (β) term in the SPSS regression output expresses the direction of influence for its associated factor. A $\beta > 0$ means that pilots with higher scores on that factor are more likely to take off. That is a go factor. Conversely, $\beta < 0$ denotes a no-go factor. For no-goes, the higher the predictor score, the *less* likely is takeoff.

This underscores a lesson regarding the modeling process itself. As our experience with modeling became deeper, it became clear that stronger models typically had at least one go predictor and at least one no-go predictor. Part of this pairing effect may just be an artifact of how the logistic prediction equation is used (Appendix F, Equation 1). Optimal modeling may mathematically *require* either a mix of go and no-go predictors, or else a constant that can serve as either positive or negative. This makes some sense from the point of view of having to minimize an error function during data fitting.

But, mathematics aside, what could also be happening here is that people actually may be motivated by two fundamentally different cognitive processes. There could be a positive go-factor process assessing and weighing what is good about a situation versus a negative no-go process assessing and weighing what is bad about the exact same situation. This certainly makes intuitive sense and may be one operating principle underlying these results. At this point, it would be premature to argue the precise nature of these go/no-go factors. But we do need to keep this in mind as a heuristic for future use.

To summarize, without a clearly front-running model for the full data set, it ultimately seemed reasonable to break the data into Low Incentive versus High Incentive groups, and conduct separate analyses. That would control for the effect of the most reliable single IV of *Financial Incentive*, and it would allow the Low group to serve as the control and the logical contrast for the High group.

Analysis of the Low Financial Incentive Group Only

Simple Models With a Constant. Following the same basic process used for the full data set, the Low Incentive data were separately winsorized, rather than merely breaking the winsorized full set in two. This separate winsorization moderated outliers, while ensuring that data were treated as if they had been from a separate experiment.

Appendix I shows the full single-predictor-plus-constant analysis for this group. *Visibility* was the only single predictor to even approach traditional, uncorrected significance (Wald $p = .064$, $R^2 = .177$, predictivity boost over the constant-alone model [76.7 – 70] = 6.7% = 2 extra cases).

One of the major problems confronting this analysis was precisely the high base rate of pilots refusing to take off (70%). And, while this did say quite a bit all by itself, it also guaranteed that any model with a constant

could always be at least 70% correct simply by "guessing" non-takeoff. This left all other model predictors to fight over the remaining variance, almost guaranteeing modest performance no matter how clever the model. It also suggested looking at models without a constant, even if they might be harder to bootstrap.

Two Outliers Are Dropped. Close examination of the pilot debrief data ultimately suggested re-running the analysis minus two outliers—two pilots who chose to take off into the absolute worst weather for no apparent reason. Statistical analysis of the two's demographic and personality data at first failed to show any measurable significant differences from the rest of the pilots on any predictor. On the surface, their personalities and motivations seemed normal, yet their takeoff behavior seemed a mystery. Finally, their debrief sheets provided a very plausible story:

Pilot 1 (age 29)

1. Zero latency (i.e., impulsive takeoff).
2. Had never had a traumatic weather experience before.
3. Had never flown a Piper Malibu before and was anxious to try out the simulator.
4. Said "I would never try that again."
5. On retest a month later with the same scenario, he stayed on the ground.

Pilot 2 (age 22)

1. 59 minutes latency (very hesitant).
2. Had never had a traumatic weather experience before.
3. Had flown a similar situation in real life.
4. Went aloft to scout the situation. Found I-40 (highway leading straight to destination)
5. Next said he was afraid to turn around or divert, for fear of losing sight of I-40.
6. Said "I probably wouldn't do it again."
7. On retest, he also stayed on the ground.

This turned out to be a classic case where quantitative data failed to uncover the root causes of behavior, but the qualitative story was crystal clear. Pilot 1 made an impulsive takeoff decision, based on his lack of any prior truly bad experience with weather, coupled with a youthful desire to fly a million-dollar simulator in which he knew he could not really get hurt. Pilot 2 took a far more conservative tack but, ironically, one leading to the same takeoff decision. He was quite hesitant, taking nearly an hour to make up his mind to fly. Yet he, too, had never had a truly bad experience with weather and so decided to go aloft and scout the conditions on the assumption that he could always turn around and come back if the situation simply proved too risky. However, once aloft, he quickly found Interstate 40, the major road leading straight to the destination. Once on I-40, he said he was literally too scared to turn back, for fear of losing sight of the road. From the weather map he knew that the front would break in about 20 minutes—and, from that knowledge, he felt it actually *less* risky to keep going than to abort the flight.

The question at hand here was whether or not dropping these two pilots from the analysis would be justified. We felt it was. First, this was an exploratory study, so extensive reporting of results was critical. Low Type I error always comes at the expense of inflated Type II error. So, completely rigid standards of reporting would mean failing to report potentially valuable preliminary analytical information useful to others in the field. Second, we had at least anecdotal information supporting the idea that neither pilot would have taken off, had they had more experience with weather. Both pilots essentially said they regretted their decision (see, above, the final quotation made by each). To check their sincerity (and the hypothesis of weather-experience effect), we re-ran both pilots a month later, using the exact same scenarios. Neither chose to fly the second time around.

This was priceless empirical evidence: Inexperience enables foolish behavior. Inexperienced pilots lack key mental constructs, namely how difficult it is to fly through certain kinds of weather. This is a root problem. Moreover, this root problem is correctable. It can be fixed with a very small amount of simulated weather experience. Our two errant young pilots aptly demonstrated that experience can change behavior, since this experiment, itself, was precisely a small amount of simulated weather experience. This was good news for pilot training, and made sense to report, whether based on anecdotal evidence or not.[3]

With two outliers excluded from the analysis, the Low Incentive data now told a stronger, richer, more compelling story. Forward stepwise regression (LR method) was now able to produce 3-4-factor models with predictivity as high as 92.9%, and R^2 as high as .921. In perfect honesty, arcane-but-valid statistical considerations did suggest that many of these models may have been artifacts of the regression procedure.[4] So in the end, rather than defending stellar-but-suspect models, it made more sense to retreat to a single, simpler, theoretically defensible model, namely *Visibility x Ceiling* (+ constant). As Table 5 summarizes, this model's statistical effect might be lower, but it would certainly be far more reliable because it was based on straightforward logic.

This implied compactly that "*Weather*"—when defined as *Visibility x Ceiling*—was a reasonable predictor of takeoff in the Low Financial Incentive group. The constant term embodied an overall group tendency to regard *all* the weather conditions as bad, while the *VxC* interaction

term represented a kind of finer-grained distinction being made between the six conditions of visibility and ceiling themselves after the average effect of weather had been accounted for.

To be circumspect, modelwise predictivity (85.7%) was only 10.7% higher than the constant-only baseline (75%). It was also true that we did not correct for experimentwise error. But keep in mind here that the constant, itself, undeniably reflected an extremely powerful weather effect here. The easiest way to understand this is to realize that virtually all pilots would have taken off in perfect weather. As it was, 21 of 28 stayed on the ground. The odds against that happening by chance were extremely remote.[5]

This model was simple and made a great deal of sense. It contained one factor essentially saying "None of this weather looked good" plus a second factor saying "We can represent levels of 'weather acceptability' by multiplying visibility times ceiling." Together, both factors accounted for about half the explainable variance in takeoffs. The rest could be attributed to individual differences, that is, individual pilot logic reasoning and affective states.

In short, pilots appeared to have many, varied, and individual motives for their risk behavior here. They did show some commonalities. No pilot thought any of this was good weather. But they also displayed considerable individual reaction to specific weather combinations. There is no such thing as the "average" pilot when it comes to weather.

Analysis of the High Financial Incentive Group Only

Simple Models With a Constant. Like the Low Financial Incentive group, the High Incentive data were separately winsorized, rather than merely breaking the winsorized full set in two.

Appendix J shows all single-IV models (+ constant). To summarize, very little stood out.

Testing the VxC+Constant Model on the High Incentive Data. Given the good performance of the *VxC+constant* Low Incentive model, it made sense to try this on the High Incentive (n=30) data as well. Table 6 summarizes the result.

Note that 53.3% of pilots did take off (versus 30% in the Low Incentive group). The non-significant *VxC* interaction here implies that the High Incentive group failed to show the same degree of fine-grained weather discrimination as the Low Incentive group (explained below).

This model did show evidence of an overall gross-weather effect, because the constant alone could predict a base rate 53.3% of takeoffs. Since we can assume that nearly all pilots would take off in perfect weather, something had to explain all those non-flying pilots, and nothing was plausible other than weather.[6] As before, the base rate represented an overall group tendency to regard all the weather conditions as bad.

A second, more startling observation was the complete and utter lack of *VxC* interaction in the high-incentive group (p=.396). This term represented how the individual pilot judged the specific weather in a given cell—fine-grained weather discrimination, in a manner of speaking.

Given our assumption of weather discrimination being the *VxC* "cognitive chunk," it was logical to ask if some kind of fine-discrimination effect might have been hidden elsewhere, for example inside a *VxCxPredictor Y* interaction. Appendix K shows this was not supported. Only three of 86 interaction contrasts fell below the uncorrected $\alpha = .06$ level (this was even fewer than expected by pure chance).

This was extremely meaningful and important information. Whereas specific weather conditions probably did influence groupwise takeoffs under low financial motivation, once a financial bonus entered the picture, the same differences in weather ceased to matter. This point cannot be stressed too much. As soon as money entered the picture, the focus of pilot rationality shifted away from attention to weather details and toward something else. And the most logical candidate from the pilot point of view, of course, was how to successfully acquire the bonus. This finding should be especially relevant to commercial GA pilots, for whom salary is frequently tied to whether or not a particular flight is made at a particular time.

Detailed Analysis. It is one thing to make this assumption, though, and quite another to support it against all challenges. For one thing, groups are made up of individuals. What constitutes a groupwise effect may not equally apply to all individuals within that group. We knew we had at least some pilots for whom the $200 bonus was financially irrelevant. In fact, one was recently retired from commercial service at a final salary of $250,000/yr. He came right out and told us the bonus was irrelevant. So did these "bonus-immune" pilots differ in takeoffs to their "bonus-susceptible" counterparts?

With that in mind, the High Financial Incentive data were sorted into a 2x2 matrix (Table 7). There had been one predictor, *buck_mot*, which was a very straightforward debrief question: "If you were in the 'high-incentive' condition, did this affect your willingness to take off?" (yes/no).[7]

A "Yes" answer would supposedly indicate a bonus-susceptible pilot, with "No" indicating a bonus-immune pilot. These answers could then be compared with actual yes/no takeoff behavior.

Bonus-susceptible pilots should have been more likely to take off than bonus-immune pilots. Table 7 shows that a trend did emerge as predicted (chi square = .062, uncorrected). This was also supported by logistic regression (Appendix J, *buck_mot* Wald *p*=.*071*). Those not strongly desiring the bonus tended to stay on the ground ("actual frequency" column, gray cell, *n*=7). Those desiring the bonus tended to take off (*n*=12).

This was encouraging, but an alternate hypothesis could have explained the results. Could the average weather just have been different in the four cells? After all, pilots had not been assigned to any particular weather conditions by *buck_mot*. So, for instance, perhaps the bonus-susceptible pilots had merely had better-than-average weather, and maybe their stated financial desire had had nothing to do with their higher takeoff rate.

To investigate this, we created an "average weather" score for each cell by multiplying *Visibility x Ceiling* for each pilot,[8] summing values cell-wise, and dividing by the cell *n* (Table 7). By this method, larger numbers represented better average cell-wise weather. The overall average was 4476.

The result was complex, but not impossible to interpret. First, bonus-susceptible pilots (buck_mot="Y") had (by pure chance) worse-than-average weather (2333 and 4333). Yet they still took off at a higher rate than bonus-immune pilots (12/18 takeoffs vs. 3/10 non-takeoffs, or 67% vs. 30%). So we could rule out the idea that the bonus-susceptible pilots had merely had better-than-average weather, because they did not.

Second, Appendix K shows a split-significance interaction. The *VxCx buck_mot* interaction was nearly significant for the "N" group (.052, uncorrected, meaning that pilots who answered "No" to buck_mot were coded as "0" in the logistic regression equation, so the analysis was focusing on the "Yes" group). Given the yes-group β of .450, this implied that the better the weather, the more bonus-susceptibles tended to take off. It was as if their internal "threshold of acceptability" for weather had been lowered by the presence of the bonus. This cleared up the picture quite a bit. Now it looked like we were dealing with two sub-populations. In a loose sense, money behaved like an infection. Some seemed more susceptible to it than others.

Moreover (although we did not specifically test it), we could speculate that the degree of susceptibility would be a function of how much money was involved. Logically, more pilots should take off for $1,000,000 than for the $200 offered here. But the "infection threshold" is probably less important than the higher-level principle that money influences people to take risks they normally would not take. From a theoretical viewpoint we could say money is a go factor that skews the cost-benefit equation.

Buck_mot seemed to measure the net effect of that equation—that is, the perceived benefit minus the perceived cost. Further evidence of the potential of *buck_mot* can be found in Appendix L. A total of 29 of 172 (17%) *p*<.05 (uncorrected) 3-way predictor interactions involved *buck_mot*. This was a greater proportion of low *p* values than found in any other 3-way predictor combination tested.

Case Study of Two High Incentive Outliers. Like the Low Financial Incentive group, the High Incentive group also had two outlier pilots who took off in the very worst weather condition (1 mile/1000 ft). In the Low group, elimination of those outliers cleared up the analysis greatly. Was a similar approach appropriate for the High group?

It appeared that the data did not support that approach, mainly because of the debrief information. The first High Incentive pilot told us he had taken off because he a) had flown a similar situation before, b) knew he could not get injured in the simulator, c) was an instructor with 2500 flight hours experience, and, d) said the bonus was very significant to him (5 on a scale of 5). The second pilot said that he a) also had flown a similar situation before, b) also knew he could not get injured in the simulator, c) had a low *crashsig* score (2 out of 5, meaning that, had he crashed during the subsequent flight, he would have felt only "a little" embarrassed by it), and, d) also said the bonus was "fairly" significant to him (3 out of 5).

In other words, like the two Low Incentive pilots, neither of these High Incentive pilots seemed particularly intimidated by the simulator. But, unlike the Low Incentive pilots, these two knew precisely what they were getting into because they had done it before in real life. Both flew calmly and methodically, executing the flight without incident. And both told us they would repeat their decision, given the opportunity. Their systematic approach was quite unlike the first Low Incentive pilot who took off on a lark, or the second who took off and was too scared to turn back once he found I-40. The Low pilots both said they had made a mistake and refused to take off when retested a month later. The High pilots gave every impression of knowing precisely what they were doing. This gave little rationale to eliminate their data.

Summary of Results for the High Financial Incentive Group. In summary, we concluded that both weather and money had some role in predicting takeoff in the High Financial Incentive group. However, there was more to it than met the eye. Money probably changed the effect of weather most for the sub-group of individuals who wanted it most. In the absence of the takeoff bonus, most pilots seemed to judge weather rather loosely, as "acceptable" or "unacceptable." In the presence of the bonus, things got more complex. Those pilots seemed,

first, to have a lowered overall threshold for what they considered "acceptable." Second, they also seemed far more likely to take off when given better combinations of visibility and ceiling—as if they became more sensitive to the *VxC* interaction.

In short, pilots could be sorted into two groups according to stated financial motivation—bonus-susceptible versus bonus-immune. Bonus-immune pilots seemed more sensitive to weather, since only 30% took off, whereas 67% of susceptible pilots took off. Money is apparently capable of shifting the "acceptable weather threshold" for a substantial number of pilots (but not all pilots).

However, not all High Incentive pilots behaved predictably, merely according to money. Nine of 28 (32%) made decisions opposite to their stated financial motive. Obviously, they had other, varied motivations beyond just weather or money. Finally, we can assume that some of the six bonus-susceptible non-takeoffs might have taken off, had they not serendipitously received worse-than-average weather. This would have further increased the predictive value of stated financial motivation.

Is the Simulator a Valid Environment to Test Financial Motivation?

An important issue to explore here concerned a criticism some make of in simulo risk studies, namely why not take a chance if the payoff is big enough and the risk of dying in the simulator is zero? This concern has to be addressed, since it implies that simulator studies may not be valid under certain conditions.

Two debrief questions were asked to try to quantify this issue. An affective predictor *crashsig* was derived from the question "If you were to crash in the simulator, how embarrassed would you be?" This was rated ordinally (rank-ordered) on a scale of 1 ("not at all") to 5 ("extremely embarrassed").[9] Similarly, a cognitive predictor *simmot* was derived from the question "Did the fact that this was a simulation (and not reality) affect your willingness to take off?" This was scaled ordinally as "decreased willingness," "no effect," and "increased willingness." If either, or both, of these scores correlated negatively with takeoff, that would provide support for the notion that the simulator task was invalid. For example, if pilots low in embarrassment were more likely to take off, that could arguably imply a lack of "appropriate fear" of, or respect for, the simulation.

Neither item was reliable at predicting takeoff (*crashsig* Wald $p = .135$, lowest *simmot* $p = .260$ by Deviation method). Nor was the interaction of these two significant ($p = .199$). This was as we would prefer. However, integrity demands claiming only modest support for simulation validity. First, we *knew* we had at least two individuals who did not take the risk terribly seriously because they openly admitted it (see above, High Incentive condition). Second, self-reports may or may not be trustworthy. There could have been individuals who misrepresented themselves during debrief. Third, the *crashsig* data represented only 40, and the *simmot* data 51, of the 60 pilots (the idea to ask these questions came in mid-experiment). Finally, in formal logic, the absence of support for something's *in*validity is necessary, but insufficient, to confirm its validity.

In the end, perhaps the sensible thing to conclude is, once again, that "The action is in the interaction." Simulator realism probably interacts with other individual personality factors as well as external environmental factors to influence the decision of some pilots more than others. In the final analysis, it is noise—an uncontrolled source of variance. We can probably only hope to control it by random assignment of participants to groups and conditions because no modern Institutional Review Board is likely to permit use of electroshock or any truly punitive consequence for a simulated crash. The only other alternative (counterbalancing) is rarely a viable option. Given a limited subject pool plus a host of other factors equally worthy to counterbalance, that becomes virtually impossible, leaving random assignment the only truly viable option for exploratory studies such as this.

DISCUSSION

This was an extremely difficult analysis, for three main reasons. First, there were no easily identifiable effects for the independent variables (*Visibility, Cloud Ceiling, Financial Incentive*), or for the rest of the takeoff predictors examined one at a time. Whatever was influencing takeoff was far more subtle and hidden.

There was certainly one enormous statistical effect for "weather" as a whole. Over half the pilots chose to stay grounded, whereas close to 100% would have normally taken off in perfect weather. The trouble came in trying to determine what "weather" meant in the minds of the pilots, given that it did not seem to mean visibility by itself or ceiling by itself. The main logical alternative left was that visibility and ceiling were somehow interacting, and that takeoff behavior was possibly also being influenced by factors such as individual pilot personality, experience, knowledge, and financial incentive. These other influences could have acted separately, or interactively in complex combinations, possibly unique to each pilot. So the way to get at the "Go/No-go equation" was to examine more complex, multi-factor models.

After exhaustive search, there proved to be dozens of roughly equally well-performing multi-factor models. These had marginal significance, indicating that multiple factors were exerting a concerted effect, often in the form

of interactions. Plainly stated, pilots do not normally make important decisions based on a single factor alone. They consider multiple factors simultaneously. Moreover, different pilots are motivated by different combinations of internal and external circumstances. This is not a simple conclusion, but it is logical and defensible.

The second main difficulty with this analysis was the side effect of looking at so many different models. This greatly increased the probability of Type I error (the chance of appearing to have an effect when none is truly present in the population at large). On the one hand, in such a preliminary study we wanted to explore a large number of candidate predictors to look for effective ones. On the other hand, had we used anything as statistically rigorous as a Bonferroni-type correction, there would be no results to report[10]. Since this was a preliminary study, it made sense to report preliminary results, as long as it was perfectly clear that the values reported needed to be considered strictly tentative, and that we followed through with a Part II report on statistical issues, which we have.

The third reason this analysis was difficult concerned the nature of causation itself. The exegesis of these data has to be both quantitative *and* qualitative. Neither way of looking at it is complete all by itself. Just looking at numbers often fails to uncover a pilot's full motivations. Sometimes the simplest, most reliable thing to do is to just come out and ask them why they did what they did. And yet, we certainly respect the standard arguments against qualitative data, namely that each person's story is anecdotal, and that people occasionally dissemble, or may not even be conscious of their full motivations. Therefore, the most complete solution we can offer is to cross-check numbers with stories and look for concurrence and confluence of logic. A "good story" usually can be supported by numerical data.

High-Level Conclusions

"Go Factors" and "No-Go Factors" Modulate Risk Behavior. Go factors can be defined as ones that increase the chance of a pilot's proceeding with some risk-laden course of action. No-go factors are then factors that decrease the likelihood of taking that risk. Go/no-go factors can be external to the pilot (e.g., cloud ceiling) or internal (e.g., anxiety). We can hypothesize that each pilot has a unique go/no-go "cognitive/affective equation" whose outcome we can model, based on some weighted, summed combination of factors.

In fact, this was just the assumption underlying our logistic regression analysis. In each model, β terms defined go factors when $\beta > 0$, and no-go factors when $\beta < 0$. Predictivity and R^2 were taken as indicators of a factor's relative strength, and Wald p indicated reliability, the likelihood that our results were due to chance. As it turned out, estimation of reliability was more complicated than indicated by the SPSS output. But that did not change the basic, high-level conclusion that go factors and no-go factors probably do capture much of the way pilots actually think about marginal weather (as well as many other risk factors). That, by itself, is a very useful way to think about risk-taking behavior because it does seem to reflect how people think. Usually, part of us wants to do one thing while another part wants to do the opposite.

Individual Demographic and Personality Factors Were Not Good Predictors of Weather-Related Risk Taking. Demographic factors such as pilot age and flight hours did not produce reliable, highly predictive risk models all by themselves here. The exception was models with interaction terms (discussed below). The same was true of such personality factors as sensation seeking, aggression, venturesomeness, and anxiety. Essentially, this said that no one factor or group of individual factors belonging to pilots themselves could explain most takeoff decisions. Whatever was shaping pilots' weather risk decisions was not simple, but complex.

Pilots Probably Look at Groups of External Weather Factors as "Cognitive Chunks." What did produce a large number and variety of interesting models was statistical interactions. Scores of models contained interactions demonstrating marginal reliability. Some of these were undoubtedly artifacts because false-positive results do occur whenever many, many models are tested. However, the large number of "almost significant" models found makes it unlikely that all were meaningless.

Statistical interactions imply cognitive/affective factor-grouping (chunking) because an interaction essentially means that the whole has more effect than just the sum of its parts. For example, the *Ground Visibility x Cloud Ceiling* interaction found in the Low Financial Incentive group implied that the statistical effect of $V*C > V+C$.[11]

What this means in terms of how pilots think is that multiple weak factors, taken together, can sway a pilot's decision. It depends on the factors involved and the relations between those factors.

Pilots Have Varied, Complex Motivations for Takeoff Into Marginal Weather. The lack of one single, simple, definitive model said there was no "average pilot" here. Each pilot was unique. Each pilot filtered the external IVs (*Visibility, Ceiling, Financial Incentive*) through his or her unique personality and experience. And—because so many personality and situational dimensions have demonstrated influence on risk taking in other venues—determining exactly which factor combinations resulted in risky weather behavior here turned out to be difficult. This difficulty was exacerbated by the interaction effects referred to earlier. This may be an "untidy" conclusion but, again, it is logical and defensible.

Specific Lower-Level Conclusions

With caveats fully in mind, the "best" quantitative models are summarized below (Table 8). These represent a reasoned balance between numerical performance, theory, and logic.

However, to be perfectly clear, of these three, only the Low Incentive model approaches traditional reliability due to the large number of models tested. Refer to the Part II report for details. All full-set models had to be heavily discounted. First, the predictivity and R^2 were roughly equal to what we could expect by chance from random number simulations (conducted in Part II). Second, those particular models were less theory-based and more just the output of stepwise regression. Stepwise regression produces maximum statistical significance, but operates mechanically without regard to experimental theory and logic. Third, the "best" full-set model was complex, with one 3-way interaction, one 2-way interaction, and three single factors. Complex, weakly theoretical models often fail to hold up on retest. Finally, there were numerous other models, not shown here, with practically equal performance.

So, if the Full Set models said anything at all, they supported the assertions stated previously. Specifically, a lot was going on inside the heads of our pilots.

Sorting the data by financial incentive had a very beneficial effect on the analysis. In the Low Incentive group, differences in weather did seem to affect takeoff behavior. At least some pilots seemed to mentally operationalize "weather" as a synergistic interaction of variables[12] (*Visibility x Ceiling*) whose net effect was greater than just the sum of parts. It is reasonable to extend this conclusion to include other factors as well (although specifically naming those other factors and their interactions would not be justified by these data alone).

Turning to the High Incentive group, we did not see many simple, straightforward effects. Money apparently did change some cognitive/affective processes, but it was not clear which were affected. Weather still mattered, but the effect was weaker than in the Low Incentive group. Moreover, as a group, pilots made very little distinction between the various weather conditions themselves. The best regression models involved the financial incentive. This probably meant that financial incentive interfered with the cognitive process of weather discrimination. However, an equally plausible hypothesis is that weather discrimination, itself, was unaffected and was just consciously overridden by the motive of money.

Whatever the cause, according to their self-report, 25% of the pilots appeared to be immune to the financial bonus. Others appeared quite susceptible (43%). The rest were unpredictable, at least their stated financial motives did not match their final behavior. However, keep in mind that the relative proportions were specific to this group, and would not generalize to all pilots under all circumstances. What might generalize is the rough principle of risk susceptibility.[13] But even that should be approached with caution. First, the results of this particular study suggest but do not firmly support it. Second, while there are clearly pilots who adhere more to a set of internal standards than others, we did not test all circumstances. Temptation is certainly domain-specific *and* reward-specific. For instance, suppose the bonus been $100,000 instead of $200? Or life versus death? The point is that predictions generally hold true only for very specific sets of circumstances. Change the circumstances, and you may completely change the ability to predict what a given pilot will do.

RECOMMENDATIONS

Immediate Conclusions

We agree with Oppe (1988) and others that risk tolerance depends on situation specifics and a large number of factors. Theoretically, these results supported the idea that most pilots were aware of some degree of weather risk, but some overrode that assessment because of one or more other factors, for example, money.

Weather risk-susceptible pilots consisted of two groups, a) the small group who were not aware of the risk in the first place and, b) the small group who chose to fly but may have lacked the skill to do so flawlessly under every circumstance.

The first group would be fairly easy to amend. A small amount of PC-based weather training could be very helpful in dealing with pilots who have never flown in various types of bad weather. Such training would be relatively inexpensive in time and money, and could directly translate into lives saved.

The second group presents a somewhat greater problem, mainly having to do with identification of individuals at risk for having too much self-confidence, given the skill they possess. We obviously do not have to be concerned with highly skilled pilots whose abilities always exceed the risks they take. The ones we worry about are those who have more confidence than actual skill. Yet overconfidence is plainly hard to measure, at least on paper. Pilots tend to be highly confident people to start with. So how can we tell which ones "talk the talk" better than they "walk the walk?" Paper-and-pencil methods are obviously incomplete. And other methods (e.g., flight simulators), introduce their own set of problems. For example, a flight simulator is a non-lethal environment. Since one cannot actually die, exactly what risk are we measuring? Some pilots take such experiments dead seriously, others less so. Some try to please the experimenter, while

others respond based on natural behavioral instincts. Finally, what we really wish we could measure is not risk taking per se, but *inappropriate* risk taking. So how do you penetrate the fiber of individuals to know exactly what their real skill level is? And, if that were possible, the number of pilots taking inappropriate risks is going to be small, so how can that be pulled out of a large pool of otherwise normal research participants? This a signal-to-noise ratio problem.

These are problems that need to be addressed. The optimistic viewpoint is that we do learn valuable, practical things from this kind of research. For instance, we now can guess that inexpensive, garden-variety PCs must soon play a pivotal role in pilot training because they can safely deliver certain critical kinds of flight experience for a relatively small investment in time and money. This suggests development of a personal computer-based program that can be used in conjunction with inexpensive home flight simulators such as Microsoft Flight Simulator®.

Future Research Directions

We are still mainly at the factor-identification stage. We need to keep developing this comprehensive list of factors shown to influence inappropriate weather-related risk taking. As this study shows, this list of factors needs to include the possibility that some factors can synergistically interact with others to produce an effect greater than each would have separately. To restate the mantra, "The action is in the interaction."

Our next weather-risk study will focus on the role of financial incentive, since that showed the greatest reliability of the three individual IVs tested in this study. We hope to settle the reliability issue by a relatively quick look at just that one predictor. Beyond that, we intend to explore the role of social factors (e.g. peer pressure) since this is also a plausible study area, well researched in other fields, but far less so in aviation.

One of the major obstacles to developing a Critical Factors List is statistical reliability. Given what we now know from this study about weak, multi-factor, interactional pilot mental processes, this means we could be wrong about the influence of some of our predictor factors. The best way around this reliability problem is to have other researchers repeat certain studies. Replication enhances reliability because, if two different studies are merely 90% sure of a given result, together, the results are 100 – (100-90)*(100-90) = 99% reliable. Moreover, replication is cost-effective when done by university-based research teams using talented (inexpensive) graduate students. This serves both the purpose of ensuring reliability of results and also trains up the next generation of aviation researchers.

REFERENCES

Adams, R.J., Koonce, J.M., & Hwoschinsky, P.V. (2002). Decision-making styles associated with accidents: Defining the high-risk pilot. *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting, 2002*, 948-52.

Amsel A. The role of frustrative nonreward in noncontinuous reward situations. *Psychological Bulletin,* 1958*,* 55, 102-18.

Baranski, J.V., & Petrusic, W.M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics, 55,* 412-28.

Barratt, E.S. (1975). *Barratt's Impulsiveness Scale*. Princeton, NJ: Educational Testing Service

Bureau of Transportation Statistics (2001). *National transportation statistics, 2001.* Washington, DC: U.S. Department of Transportation.

Eysenck, H.J., & Eysenck, S.B.G. (1964). *Eysenck Personality Inventory.* San Diego, CA: Educational and Industrial Testing Service.

Eysenck, H.J., & Eysenck, M.W. (1985). *Personality and individual differences: A natural science approach.* New York: Plenum.

Eysenck, H.J., & Eysenck, S.B.G. (1975). *Manual of the Eysenck Personality Questionnaire.* London: Hodder & Stoughton.

Eysenck, S.B.G., & Eysenck, H.J. (1977). The place of impulsiveness in a dimensional system of personality description. *British Journal of Social and Clinical Psychology, 16,* 57-68.

Foster, D.P., & Stine, R.A, (1998). Honest confidence intervals for the error variance in stepwise regression. Retrieved July 23, 2003 from http://www-stat.wharton.upenn.edu/~bob/research/honests2.pdf. Unpublished manuscript, University of Pennsylvania, Wharton School, Department of Statistics.

Goh, J., & Wiegmann, D.A. (2002). The effects of distance traveled and pilot experience on pilot's decisions to continue visual flight rules into adverse weather. *Aviation, Space, and Environmental Medicine, 73*(3), 252-53.

Goh, J., Wiegmann, D.A., & O'Hare, D. (2002). Human factors analysis of accidents involving visual flight rules flight into adverse weather. *Aviation, Space, and Environmental Medicine, 73*, 817-22.

Hull, C.L. (1932). The goal gradient hypothesis and maze learning. *Psychological Review, 39*, 25–43.

Hull, C.L. (1934). The concept of the habit-family hierarchy and maze learning: Part I. *Psychological Review, 41*, 33-54.

Hunter, D.R. (1995). *Airman research questionnaire: Methodology and overall results.* (Report DOT/FAA/AM-95/27). Washington, DC: Federal Aviation Administration.†

Hunter, D.R. (in press). Measurement of hazardous attitudes among pilots. *International Journal of Aviation Psychology.*

Hunter, D.R. *Measuring the safety-related attitudes of pilots. Twenty-fifth Conference of the European Association for Aviation Psychology.* Warsaw, Poland. September 2002a.

Hunter, D.R. (2002b). *Risk perception and risk tolerance in aircraft pilots*, (Report no. DOT/FAA/AM-02/17), Washington, DC: Office of Aerospace Medicine.†

Keppel, G. (1982). *Design and analysis: A researcher's handbook.* (2nd Ed.). Englewood Cliffs, NJ: Prentice-Hall.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, *47*, 263-91.

Kahneman, D., & Tversky, A. (1982). Judgment under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic, & A. Tversky (Eds.). *Judgment under uncertainty: Heuristics and biases* (pp. 3-20). Cambridge, UK: Cambridge University Press.

Kreyszig, E. (1972). Advanced engineering mathematics. (3rd Ed.). New York: Wiley.

Lejuez, C.W., Read, J.P., Kahler, C.W., Richards, J.B., Ramsey, S.E., Stuart, G.L., Strong, D.R., & Brown, R.A. (2002). Evaluation of a behavioral measure of risk-taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*, 75-84.

National Transportation Safety Board. (1995-1997, 2002). *Annual review of aircraft accident data, U.S. general aviation*. Washington, DC.

Norušis, M.J. (1999). *SPSS regression models 10.0*. Chicago: SPSS, Inc.

O'Hare, D. (1990). Pilots' perception of risks and hazards in general aviation. *Aviation, Space, and Environmental Medicine*, 61(7), 599-603.

O'Hare, D., Chalmers, D., & Scuffham, P. (2003). Case-control study of risk factors for fatal and non-fatal injury in crashes of civil aircraft. *Aviation, Space, and Environmental Medicine, 74*(10), 1061-66.

O'Hare, D., & Owen, D. (2002). Cross-country VFR crashes: Pilot and contextual factors. *Aviation, Space, and Environmental Medicine, 73*(4), 363-66.

Oppe, S. (1988), The concept of risk: A decision theoretic approach. *Ergonomics*, *31*(4), 435-40.

Patrick, C.J., Curtin, J.J., & Tellegen, A. (2002). Development and validation of a brief form of the Multidimensional Personality Questionnaire. *Psychological Assessment*. *14*(2), 150-63.

Peterson R.A, & Reiss S. (1994). *Anxiety Sensitivity Index Manual.* (2nd Ed.). Worthington, OH: International Diagnostic Systems.

Reiss S., Peterson R.A., Gursky D.M, & McNally R.J. (1986). Anxiety sensitivity, anxiety frequency and the prediction of fearfulness. *Behavioral Research Therapy 24*, 1-8.

Rohrmann, B. (2002). Risk attitude scales: Concepts and Questionnaires. Retrieved January 29, 2004 from http://www.psych.unimelb.edu.au/staff/br/racreport.pdf. Unpublished manuscript, University of Melbourne, Australia.

Shappell, S. & Wiegmann, D. (2003). Reshaping the way we look at general aviation accidents using the human factors analysis and classification system. Presented at the 12th International Aviation Psychology Symposium, Dayton, OH.Spielberger, C. (1983) State-Trait Anxiety Inventory (Form Y). Palo Alto, California: Mind Garden.

SPSS, Inc., (2004). Logistic regression (algorithms). Retrieved February 19, 2004 from http://www.spss.com/tech/stat/Algorithms/11.5/logistic_regression.pdf.

Tabachnick, B.G., Fidell, L.S. (2000). *Multivariate statistics* (4th Ed.). Needham Heights, MA: Allyn & Bacon.

†This publication and all Office of Aerospace Medicine technical reports are available in full-text from the Civil Aerospace Medical Institute's publications Web site: http://www.cami.jccbi.gov/aam-400A/index.html

Weber, E.U., Blais, A.R., & Betz, N. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making, 15,* 263-90.

Wiegmann, D.A., Goh, J., & O'Hare, D. (2002). The role of situation assessment and flight experience in pilots' decisions to continue visual flight rules flight into adverse weather. *Human Factors, 44*(2), 1890-97.

Wiggins, M., & O'Hare, D. (2003). Weatherwise: Evaluation of a cue-based training approach for the recognition of deteriorating weather conditions during flight. *Human Factors*, *45*(2), 337-45.

Wilson, D.R., & Fallshore, M. (2001). Optimistic and ability biases in pilots' decisions and perceptions of risk regarding VFR flight in IMC. *Proceedings of the 11th International Biennial Symposium on Aviation Psychology*, Columbus, Ohio, March 5-8, 2001.

Winer, B.J. (1971). *Statistical principles in experimental design* (2nd Ed.). New York: McGraw-Hill.

Zuckerman, M. (1994). *Behavioural expressions and biosocial bases of Sensation Seeking.* Cambridge: Cambridge University Press.

ENDNOTES

[1] R^+ is receipt of some reward. R is avoidance of some punishment.

[2] We will use the abbreviation "uncorrected" to stand for "uncorrected for experiment-wise Type I error."

[3] Some critics would argue it unfair to draw such a strong conclusion based on anecdotal evidence taken from only two subjects. The logical response to that criticism is: a) Exploratory studies ought to make suggestions; b) Anecdotal evidence is usually more dependable when it comes from a relatively controlled experimental setting such as this.

[4] Once again, refer to the Part II study for details.

[5] Assuming a highly conservative base rate of 26/28 takeoffs for perfect weather, the estimated chance of getting the real takeoffs actually observed would be $p=1184040*.07^{21}*.93^{7}$ by expansion of the binomial—about 4 in 10 billion billion.

[6] Expected $p=145422675*.07^{14}*.93^{16} \cong 3*10^{-9}$ or about 3 in a billion.

[7] Two pilots failed to answer the buck_mot question on their debrief sheet, making n=28 for that analysis.

[8] This was exactly the same way logistic regression operationalized interactions, by multiplying the two scores together.

[9] Logistic regression is appropriate for ordinal data as well as ratio-scale or categorical data.

[10] To (over)simplify, Bonferroni correction basically involves dividing each α significance criterion value by the number of comparisons made (Keppel, 1982). The more stringent the α, the lower that p and X^2 values have to be to achieve "significance." Given too many models, α can become so stringent that literally nothing is "significant."

[11] Technically, the statistical effect of $\beta_1 V*C >$ that of $\beta_2 V + \beta_3 C$, but that does not change the basic argument.

[12] It needs to be said that an "average" weather effect (the base rate, modelable with a constant) would also emerge as an artifact from logistic regression if all pilots discriminated solely on the basis of a VxC process alone. However, to be conservative, we say "some pilots" and not "all pilots" mentally operationalize weather as an interaction.

[13] Risk is known to be domain-specific (Weber, Blais, & Betz, 2002). That is, a pilot who takes risks playing poker may be highly conservative in the air (or vice versa). This makes it virtually impossible to globally diagnose a "risk-taking personality." Aviation risk-taking is a separate problem and has to be assessed separately.

FIGURES

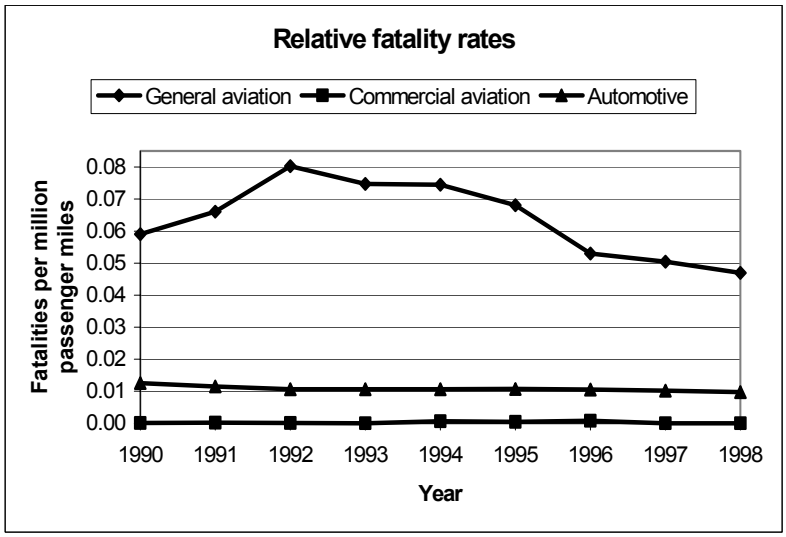


Figure 1. Fatality rates per million passenger-miles for U.S. general aviation, domestic commercial aviation, and automotive transport, sample years 1990-98.

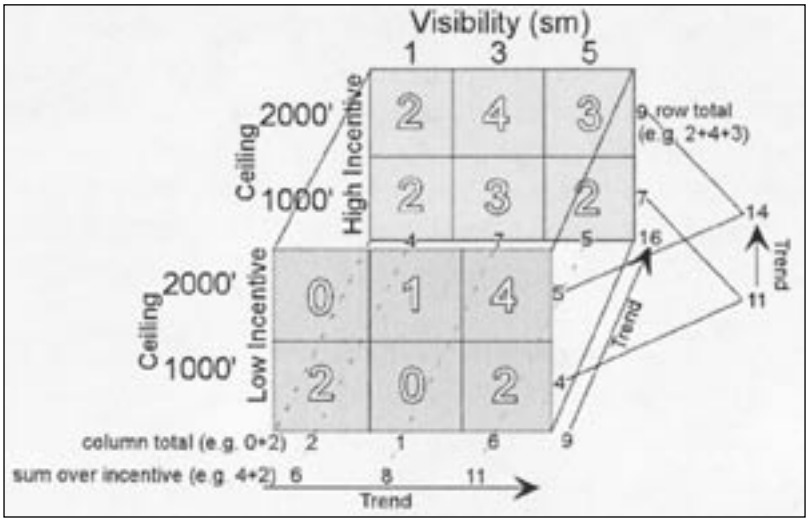


Figure 2. Summary of raw data–the number of pilots choosing to take off, given that cell's triple combination of ground visibility, cloud ceiling, and financial incentive. Each of the 12 cells contained five pilots.

TABLES

Table 1. Personality instruments used, their meanings, and a key reference to each (see References section).

Instrument	High score implies	Reference
Aviation Safety Attitude Scale	high history of aviation risk behavior	Hunter, 1995, 2002a, 2002b, in press
Anxiety Sensitivity Index	high scores indicate high anxiety	Peterson & Reiss, 1994; Reiss et al., 1986
Balloon Analog Risk Task	high risk tolerance	Lejuez et al., 2002
Barratt Impusiveness Scale V10	high impulsivity	Barratt, 1975
Eysenck Impulsivity Scale	high impulsivity	Eysenck & Eysenck, 1964,1975,1977,1985
Hazardous Events Index	high history of aviation risk behavior	Hunter, 2002
Multidimensional Personality Questionnaire	high degree of specified trait	Patrick, Curtin, & Tellegen, 2002
Risk Orientation Questionnaire	high risk tolerance	Rohrmann, 2002
Sensation Seeking Scale	high desire for stimulus-seeking	Zuckerman, 1994

Table 2. Demographic data for the full data set.

	Years flying	age	Flight hrs total	FH past 12 mo.	High-performance hrs, total	Simulated instrument hrs, total
Median	1.8	23.5	183.5	80.0	5.0	25.0
mean	4.3	26.1	753.0	159.9	37.9	69.3
S.D.	7.6	8.5	2604.7	239.6	88.5	256.8

Table 3. Summary of regression analysis for the three primary independent variables. Each of the three separate models tested contained one IV plus a constant.

		% prediction	Nagelkerke
Independent Var.	*p*	increase > constant	R^2
Visibility	0.11	3.4	0.06
Ceiling	0.43	0.0	0.01
Financial Incentive	0.07	3.4	0.07

Table 4. Summary of interactions between the three primary independent variables.

		% prediction	Nagelkerke
Interaction	*p*	increase > constant	R^2
Visibility x Ceiling	0.037	6.7	0.10
Visibility x Incentive	0.090	6.7	0.07
Ceiling x Incentive	0.053	5.0	0.09
Vis x Ceil x Incent	0.063	5.0	0.08

Table 5. Theory-based *Weather* model (*Visibility x Ceiling + Constant*, uncorrected values) for the Low Financial Incentive group minus two outliers (*n*=28).

	Wald p	Predictivity	R^2
VIS x CEIL	0.008	Base: 75.0%	0.521
Constant	0.003	Model: 85.7%	

Table 6. Effect of "Weather" (*Visibility x Ceiling* interaction) on the High Financial Incentive group.

	Wald p	Predictivity	R^2
VIS x CEIL	0.396	Base: 53.3%	0.033
Constant	0.604	Model: 56.7%	

Table 7. Frequency counts for High Financial Incentive pilots, grouped into 4 cells by *Takeoff* versus individual pilots' self-stated personal financial motivation (*buck_mot*).

	Actual frequency				X^2 expected freq				ave. weather	
	buck_mot				buck_mot				buck_mot	
Takeoff	N	Y			N	Y			N	Y
N	7	6	13		4.6	8.4	13		5571	2333
Y	3	12	15		5.4	9.6	15		5667	4333
	10	18	**28**		10	18	**28**		average	
					X^2	**0.062**			4476	

Table 8. Best models for full-set, Low, and High Financial Incentive sub-groups. "Best" is defined by a combination of high reliability, predictivity, and R^2, in combination with support from logic and theory.

Data set	Best model found	Wald *p*	Predictivity	R^2_{Nagel}	Comments
Full set	*Visibility x Ceiling x MPQ Aggression*	.001-.003			Two models, depending on
N=60	*Actual instr. time (90 d) x $ Incentive*	.046-.072	75-80%		whether the referent was
	Anxiety Sensitivity Index	.001		.527-.488	"instrument-rated" or "non-
	Hazardous Events Index	.044-.092	(base 58.3)		instrument-rated". Models
	Instrument Rating (VFR vs. IFR)	.029-.106			show some effect of weather
					but suggest breaking the data
					into Low vs. High $ Incentive
Low $ Incentive	*Visibility x Ceiling*	.008			Model shows that "Weather"
N=28	*Constant*	.003	85.7%	.521	can be defined as (*Vis x Ceil*)
High $ Incentive	*Financial Motivation (buck_mot)*				Many shallow models
N=30	*x Predictor P*	≈ .04	≈ 75%	≈ .28	Models do not exceed chance
	Constant				No sig. effect of weather

APPENDIX A

Method of Calculation for Rate of Weather Involvement in GA Fatalities

In the Introduction of this report, we stated "Data derived by us from National Transportation Safety Board statistics (1995-1997) support this assertion, showing IMC implicated in approximately 32% of GA fatalities." Here is how we calculated that figure:

Year	% GA acci-	% in	% of IMC acci-	% of VMC
	dents in VMC	IMC	dents fatal	accidents fatal
1995	90	10	68	16
1996	90	10	63	15
1997	90	10	69	16
		average	66.7	15.7
		10%*66.7=	6.7	6.7 / 20.8 =
		90%*15.7=	14.1	0.32
		total % fatal	20.8	

We started with data from the National Transportation Safety Board (NTSB) *Annual review of aircraft accident data, U.S. general aviation*, calendar years 1995-97 (the latest available year was, indeed, 1997). The figure of 90% GA accidents occurring in VMC was taken from the text, which claimed "*More than* 90 percent of accidents occur in visual meteorological conditions" (emphasis ours, NTSB, 1997, p. 2 used to illustrate). So we assumed that 90% was an estimate, but a conservative one. We then followed these steps:

- The reports state that 90% of GA accidents occur in VMC (visual meteorological conditions).
- Therefore, by deduction, (100-90) = 10% must occur in IMC (instrument meteorological conditions).
- The reports state that 68, 63, and 69% of IMC accidents during their respective years involved fatalities (average = 66.7%), as opposed to 16, 15, and 16% of VMC accidents, respectively (average = 15.7%).
- If 10% of accidents involve IMC, and 66.7% of these are fatal, then (10% * 66.7%) = 6.7% of overall accidents therefore involve IMC PLUS fatalities
- If 90% of accidents involve VMC, and 15.7% of these are fatal, then (90% * 15.7%) = 14.1% of overall accidents therefore involve VMC PLUS fatalities.
- Therefore, ((6.7 / (6.7+14.1)) = 32%) is the ratio of (fatal accidents involving IMC / total fatal accidents), meaning that IMC is implicated in approximately 32% of GA fatalities.

APPENDIX B

Participant Debrief Form

S # _____

- What is your own normal personal minimum for VFR visibility? _____
- Your normal personal minimum for VFR cloud ceiling _____
- Are these minimums rock-solid, or do you adjust them a little, depending on the circumstances?_____
- Have you ever flown this particular route before (or a similar situation)? _____
- Did the distance you had to fly through bad weather affect your willingness to take off ? _____ (for example, if the distance had been greater, would you have been even less inclined to take off than you were?)
- If you were in the "high-incentive" condition, did this affect your willingness to take off? _____
- Do you think having passengers would affect your willingness to take off? (increase it ____, no change____, decrease it ____)
- If you had a lot more flight hours, would that have change your willingness to take off? (increase it ____, no change____, decrease it ____)
- If your flight mission had been critical (for example, delivering a human heart for surgery), would that change your willingness to take off? (increase it ____, no change____, decrease it ____)
- Have you ever flown a Piper Malibu before?_____ Did this affect your willingness to take off?
- It made me **more** willing because I was anxious to try it out ___,
- It didn't matter one way or the other ___,
- It made me **less** willing because I was afraid I'd make more mistakes ___
- Did the fact that this was a simulation (and not reality) affect your willingness to take off?
- It **increased** willingness because
  - (a) I wanted to fly the sim___ and/or
  - (b) I knew I couldn't really get injured in it___,
- No, it had no effect because
  - (a) it didn't matter to me one way or the other___
  - (b) there were positives and negatives but they cancelled each other out___
- It **decreased** willingness because
  - (a) I was unfamiliar with this particular simulator___
  - (b) I didn't want to make any mistakes in front of the experimenter___
- How economically significant was the money to you?
  1__not at all 2__a little 3__fairly significant 4__significant 5__very significant
- If you were to crash in the simulator, how embarrassed would you be?
  1__not at all 2__a little 3__fairly 4__significantly 5__extremely
- Have you ever had a bad flight experience related to weather?___If so, please describe briefly below.

APPENDIX C

	YrsFlg	age	gender	Type Lic	Inst rating	fhtot	fh12mtm	fh90dtm	HP Tot	HP 12m Tm	HP 90d Tm	SI time	SI 12m Tm	SI 90d Tm	AI Time	AI 12m Tm	AI 90d tm
N	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60
# Missing																	
Median	1.8	23.5	male	priv.	50/50	183.5	80.0	35.0	5.0	3.0	0.0	25.0	10.0	5.0	2.0	1.0	0.0
Average	4.3	26.1				753.0	159.9	49.1	37.9	16.9	6.9	69.3	20.7	14.4	52.7	4.7	2.2
S.D.	7.6	8.5				2605	239.6	54.1	88.5	40.8	18.8	256.8	30.1	26.5	322.3	11.0	5.2
Max	48.25	69				20000	1200	250	560	200	125	2000	175	160	2500	75	25
Min	0.25	18				35	0	0	0	0	0	0	0	0	0	0	0
Skew	4.2	2.8				7.1	2.9	1.9	4.2	3.6	4.9	7.4	3.4	3.8	7.7	4.9	3.1
SE Skew	0.31	0.31				0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31
z_{skew}	13.67	9.07				22.95	9.47	6.05	13.51	11.57	15.84	24.13	11.15	12.29	24.86	15.90	10.09
p_{zskew}	0.000	0.000				0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Kurtosis	21.1	10.6				52.8	9.0	3.5	21.0	12.4	27.4	56.8	14.5	17.3	59.2	29.1	9.5
SE Kurt	0.61	0.61				0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61
z_{kurt}	34.8	17.5				86.8	14.8	5.7	34.5	20.4	45.1	93.3	23.9	28.4	97.4	47.8	15.5
p_{zkurt}	0.000	0.000				0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Outlier(s)?	yes	yes				yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes

	Vismin	ceilmin	minadj	flnrt	dthruwx	buckmot	passmot	flthrmot	missnmot	malibu	malibsub	simmott	simmosub	bucksig	crashsig	trafsig	badwx
N	59	59	57	57	57	28	56	55	47	53	50	51	46	51	40	11	55
# Missing	1	1	3	3	3	32	4	5	13	7	10	9	14	9	20	49	5
Median	4.0	2000	1	0	1	1	-1	0	1	0	0	0	3	3	3	2	0
Average	4.1	2186	0.8	0.4	0.6	0.6	-0.5	0.3	0.7		-0.2	0.2		2.7	3.2	2.3	
S.D.	1.6	1213					0.7	0.6	0.5		0.6	0.5		1.3	1.2	1.3	
Max	10	8000	1	1	1	1	1	1	1	1	1	1	6	5	5	5	1
Min	1	1000	0	0	0	0	-1	-1	-1	0	-1	-1	1	1	1	1	0
Skew	1.2	2.4					1.0	0.0	-1.3		0.0	0.4		0.2	0.2	0.8	
SE Skew	0.31	0.31					0.32	0.32	0.35		0.34	0.33		0.33	0.37	0.66	
z_{skew}	3.75	7.77					3.08	0.02	-3.73		0.00	1.34		0.53	0.43	1.22	
p_{zskew}	0.000	0.000					0.004	0.399	0.000		0.399	0.163		0.347	0.363	0.190	
Kurtosis	2.5	8.8					-0.1	-0.4	0.7		-0.1	0.4		-1.1	-1.2	0.5	
SE Kurt	0.61	0.61					0.63	0.63	0.68		0.66	0.66		0.66	0.73	1.28	
z_{kurt}	4.0	14.4					-0.2	-0.6	1.0		-0.2	0.6		-1.7	-1.6	0.4	
p_{zkurt}	0.000	0.000					0.390	0.324	0.235		0.389	0.339		0.092	0.108	0.367	
Outlier(s)?	yes	yes							yes								

Standard z-tests (e.g. skew/[standard error of skew]) showed that the demographic data were greatly skewed by the presence of a small number of older pilots with, for instance, a great deal of flight experience. Winsorizing corrected virtually all this non-normality. Appendix D explains the factors examined.

APPENDIX D

Complete list of factors examined

Name	Description		Name	Description
subjnum	Order in w hich S w as run		asi	AIS Anxiety Sensitivity Index
idnum	S. ID number		bis	BIS-10 Barratt Impusiveness Scale total
takeoff	Takeoff (yes/no)		imp_plan	BIS Impulsive Planning
latcy	Latency (time elapsed before takeoff)		imp_motr	BIS Motor Impulsivity
vis	Ground Visibility (statute miles)		imp_cog	BIS Cognitive Impulsivity
ceil_k	Ceiling (in thousands of ft)		impuls	EIS Eysenck Impulsivity Scale Impulsivity
incent	Financial Incentive (bonus / no bonus)		ventur	EIS Venturesomeness
wxsevrty	Weather severity (1 / (Vis x Ceil))		empath	EIS Empathy
exptr	Experimenter (WK or HH)		hei	HEI Hazardous Events Index (Hunter)
yr_flg	Year started flying			Multidimensional Personality Questionnaire
yrs_flg	Years flying		wellbe	MPQ Wellbeing
age	Age		socpot	MPQ Social Potency
gender	Gender		achieve	MPQ Achievement
type_lic	License type		socclose	MPQ Social Closeness
inst_rtg	Instrument rating (yes/no)		stress	MPQ Stress Reaction
fh_tot	Flight hours total		alienate	MPQ Alienation
fh_12m	FH past 12 months		aggress	MPQ Aggression
fh_90d	FH past 90 days		control	MPQ Control
hp_tot	High-performance aircraft hours total		harmav	MPQ Harm Avoidance
hp_12m	HP past 12 mo		tradit	MPQ Traditionalism
hp_90d	HP past 90 days		absorpt	MPQ Absorption
si_tot	Simulated instrument hours total		roq_c	ROQ Risk Orientation Q'naire Cautiousness
si_12m	SI past 12 mo		roq_p	ROQ Risk Propensity
si_90d	SI past 90 d		sss	SSS Sensation Seeking Scale
ai_tot	Actual instrument hours total		anx_st	STAS State-Trait Anxiety Scale State
ai_12m	AI past 12 mo		anx_tr	STAS Trait
ai_90d	AI past 90 d			Balloon Analog Risk Task (Lejuez)
vis_min	Personal visibility minimum		durn_sec	BART Task Duration (seconds)
ceil_min	Personal ceiling minimum		pumpsavg	BART Average # of pumps
min_adj	Adjust mins. to match the situation?		pmpavglo	BART Average # of pumps (low incentive)
fln_rt	Flow n this route before?		pmpavgme	BART Average # of pumps (med incentive)
dthruwx	Did distance through the w eather matter?		pmpavghi	BART Average # of pumps (hi incentive)
buck_mot	Was the $ bonus a motivation? (yes/no)		pmpadjav	BART Adjusted Ave. # of pumps
pass_mot	Would passengers have been a motivation?		padjavlo	BART Adj. Ave (low incentive)
fhincmot	Would more flt hrs increase motivation?		padjavme	BART Adj. Ave (med incentive)
missnmot	Was the type of mission a motivation?		padjavhi	BART Adj. Ave (high incentive)
mal_sub	Was the type of flight simulator a motivation?		pay_tot	BART Total Payoff (cents)
sim_mot	Was fact of being a sim (vs. reality) a motvn?		pay_low	BART Total Payoff (low incentive)
simmotsb	Sub-categories of sim_mot		pay_med	BART Total Payoff (med incentive)
buck_sig	How significant w as the $$ to you?		pay_hi	BART Total Payoff (high incentive)
crashsig	Was w orrying abt crash a motivation?		bang_tot	BART Total Balloon Explosions
tx_mot	Was traffic a motivation?		bang_low	BART Balloon Explosions (low incent)
badwx	Ever had a bad w x experience? (y/n)		bang_med	BART Balloon Explosions (med incent)
asa	Aviation Safety Attitude Scale (Hunter)		bang_hi	BART Balloon Explosions (high incent)

Three of these factors were not predictors, namely *idnum*, *takeoff*, and *latcy*. *IDnum* was merely the numerical proxy for subject name. *Takeoff* was the dependent variable. *Latency* was a descriptor, and could not be used as a discriminative predictor because maximum latency (120 minutes) was always associated with takeoff, anything less, with non-takeoff. Therefore, there were 83 usable predictors.

APPENDIX E

Statistical Issues in Logistic Regression

Outliers. *Outliers* are defined for our purposes here as any score greater than 3 standard deviations above or below the mean. Outliers can sometimes exert an almost unbelievable effect on the statistical outcome of an analysis. Take, for example, a distribution of ones and zeros representing Financial Incentive, one of our predictors of Takeoff. For the full data set, N=60, our actual raw distribution yields the following result during SPSS logistic regression:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	INCENT	.981	.541	3.287	1	.070	2.667
	Constant	-.847	.398	4.523	1	.033	.429

a. Variable(s) entered on step 1: INCENT.

This result says that the probability of Incentive being a significant predictor of Takeoff is .070.

Now let us change *one single value* in the data distribution from a "0" to a "10" to represent, say, a typographical error during data coding. Changing just this *one value in 60* results in the following:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	INCENT	-.027	.204	.018	1	.894	.973
	Constant	-.319	.294	1.171	1	.279	.727

a. Variable(s) entered on step 1: INCENT.

Suddenly we have gone from $p = .070$ to $p = .894$ in one step—by turning a single data point into a gross outlier. Obviously, this says a lot about the need for accurate data coding. It also says quite a bit about how outliers can affect an otherwise normal data distribution. Now logistic regression does not have an underlying logical assumption of normality (Tabachnick & Fidell, 2000). You could, for instance, use data with any relatively symmetrical distribution. But it does have problems with outliers, as this clearly demonstrates.

The data in this study showed outliers in the demographics, where a small number of older pilots significantly skewed the distributions for predictors such as age, flight hours, and years flying. Without some kind of correction, therefore, the effect of outliers would have led us to seriously misinterpret the statistical analysis.

Applying a data transformation (such as a square root or logarithmic function) is a common way to deal with outliers. A somewhat less well-known, but equally respected treatment is *winsorization* (Winer, 1971, pp 51-54). 1971). In winsorization, the two most-extreme values in the distribution (the one highest and the one lowest) are replaced by a copy of the next most-extreme values. For example, in the distribution

0 1 1 1 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5 5 5 99 (mean 7.23, SD 20.54)

we would replace the "0" with a "1" and the "99" with a "5."

1 1 1 1 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5 5 5 5 (mean 3.00, SD 1.38)

Now this new distribution is still not normal because it is too flat. But it no longer has the gross outlier it once had. That extreme value of "99" is still represented by a relatively high value, which preserves the ordinality (rank order) of the scores. But notice that there was no actual change to most of the numbers. Only two values were changed, and one of those was a very modest change from a "0" to a "1." Whereas, if we had applied a mathematical function such as a square root to shrink the "99" closer to the mean, almost all of the values would have been affected. Here winsorization exerts its biggest effect on the greatest offender, which is exactly how data conditioning should work. This illustrates how this technique can sometimes preserve the spirit and actuality of a distribution much better than can some of the more routinely used methods. For this reason, it was the method of choice for our data.

If a distribution has more than one outlier, say

0 0 1 1 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5 5 47 99 (mean 9.09, SD 22.23)

we simply apply the winsorization procedure twice, to yield

1 1 1 1 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5 5 47 47 (mean 6.82, SD 13.06)

at stage one and

2 2 2 2 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5 5 5 5 (mean 3.18, SD 1.14)

at stage two. In this example, the two-stage winsorization affects 6 values, rather than just 2. For this reason we have to be careful in repeating this process too often, since it can lead to the antithetical problem of range restriction.

In this, study winsorization was limited to no more than 2 stages. For example, in the full data set (N=60) 16 demographic variables were seen to have outliers > 3 SD, and therefore received either a 1- or 2-stage winsorization, depending on what was needed to eliminate these outliers. After treatment, all 16 variables emerged corrected to tolerance.

A final point worth mentioning is that winsorization has a net result of making our statistical analysis more conservative. This happens precisely because the distributions' ranges and variances contract during conditioning, and any time variance contracts, *p*-values generally contract as well. This is not true with purely ordinal statistics, because these calculate their value based on nothing more than rank order. But both chi-square and logistic regression do not fall into that category. While logistic regression is often touted as being distribution-free, in fact, we have graphically illustrated that things are a bit more complex. Outliers skew its innermost calculation of likelihood ratios (SPSS, 2004). However, the data conditioning process employed here allowed us to successfully treat data and to present *p*-values representing useful-yet-conservative estimates of statistical reliability.

Correction for Familywise Error. Another important issue is the one of correcting *p*-values to account for the number of predictors examined. Most statisticians recommend some sort of correction for experimentwise Type I error (unwarranted rejection of the null hypothesis). Otherwise, if we do many tests, odds are that some will be "significant" simply by chance.

However, we consciously chose to deviate from that standard procedure because, in an exploratory study such as this, such rigor, while admirable in one sense, would most certainly have the net result of too much Type II error, that is, failure to detect a true effect where there was one. And, while the danger of inflated experimentwise Type I error was fully appreciated, we also felt it made more sense to report low *p*-values where found, because these really do represent the best guess we have regarding effect.

The ideal way to resolve the problem, of course, is to run Monte Carlo simulations to get estimates for mean predictivities and R^2s, given specific parameters of specific models. This was done in Part II of this report. Another accepted approach is to replicate studies or parts of studies, using different participants. That will be done in follow-up studies, whenever possible.

APPENDIX F

Brief Description of Logistic Regression

Logistic regression is a statistical technique specially constructed for use with discrete dependent variables, for example, Takeoff versus No Takeoff. It is a very useful technique, but it is also extremely easy to miscode, misunderstand, and misinterpret. The best way to understand it is through a combination of mathematics and example.

Regression is the search for factors that predict other factors. In this experiment, we wanted to predict the likelihood that an average pilot would take off into known marginal weather, given the added influence of financial incentive. Three of our predictive factors (*Visibility, Cloud Ceiling,* and *Financial Incentive*) were under experimental control; the rest reflected either demographic or personality characteristics of each individual pilot.

Logistic regression uses an equation to predict the outcome of an *event*, in this case Takeoff versus No Takeoff (Dreyszig, 1972; Norušis, 1999; SPSS, 2004). This equation is

$$p_{event} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_n X_n)}} \quad (1)$$

where e is the natural log (approximately 2.718), B_0 (beta-sub-zero) represents a constant, and B_i is the corresponding beta weight for the ith predictor, X_i score. Varying the values of the exponent of e produces a distinctive sigmoid (S-shaped) curve capable of representing probability of takeoff

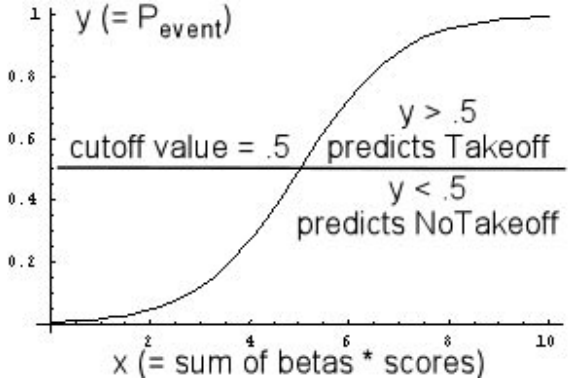


Figure F1. Sample plot of the sigmoid function $y=1/(1+ e^{(-5 + x)})$, showing how the overall value of the prediction equation lies between zero and one. In our case, this represents the probability of a pilot taking off, 0-1 (0-100% chance), given some particular combination of predictor scores X_1 through X_n. When a given pilot's calculated probability exceeds an predetermined cutoff level (for example, 0.5), we will predict "Takeoff," otherwise we will predict "No Takeoff."

Logistic regression has two very attractive advantages over competing statistics. First, as we mentioned, it allows us to make predictions. Second, it allows us to test statistical interactions between predictors. Equation 2 shows how this is typically implemented, showing the prediction equation with its constant B_0, one main variable X_1, plus one interaction term involving three factors B_2, X_2, and X_3. Notice that the interaction term literally involves multiplying together the separate predictors. This is an important point to which we will presently return.

$$p_{event} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2 X_3)}} \quad (2)$$

To illustrate this numerically, suppose we tested a model where X_1 represented a pilot's score of 34 on the Rohrmann Risk Orientation Questionnaire (subscale P), and where X_2 and X_3 represented the interaction of Visibility x Ceiling, 3 (miles) and 1 (feet, in thousands), respectively. In that case, the prediction equation for that individual would be

$$p_{Takeoff} = \frac{1}{1+2.718^{-(-3.396+.074(34)+.197(3)(1))}} = .428$$

Since .428 is less than the default cutoff value of .500, we would predict that this particular pilot would not take off.

When we run the SPSS analysis on the full data set, the program basically goes through a similar process for each individual, computing a set of guesses regarding each pilot's takeoff. Some guesses will be right, others wrong. Then the beta weights are shifted slightly, the analysis is repeated, and the results compared to the priors. If shifting the betas in that direction produces improvement, the direction of shift is repeated, otherwise it is reversed. After a certain number of iterations, the process halts and summary tables are produced. Here is one of the summary tables for an actual model:

The most important numbers in this table, as far as we are concerned, are the β weights, and the significance of the Wald statistic (Sig). What the βs here tell us is primarily the direction of the association between a predictor and the outcome. Take ROQ_P, whose β is positive. That tells us that an *increase* in the ROQ_P score predicts an *increase* in takeoff probability. If β had been negative, an increase in the ROQ_P score would have predicted a *decrease* in

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	ROQ_P	.074	.061	1.465	1	.226	1.077
	CEIL_K by VIS	.197	.095	4.325	1	.038	1.217
	Constant	-3.396	1.902	3.186	1	.074	.034

a. Variable(s) entered on step 1: ROQ_P, CEIL_K * VIS .

takeoff probability. The magnitude of β is also important, though its interpretation is more complicated. We have to take into account how the predictor was scored. Since β is multiplied by the predictor score, if the predictor scores are large (e.g. 1000 feet altitude) then even a small β can be very influential.

The Wald significance (Wald *p* value) works very much like a normal statistical *p* value. Wald *p* tells us the reliability of the measurement, estimating the proportion of times we would expect to find a different result, if we repeated the analysis a large number of times. In this particular instance, ROQ_P's Wald *p* is .226—too large to be considered reliable.

The Constant (β_0) in this analysis behaves somewhat like other predictors. However, the Constant is sometimes the most difficult term to interpret in a regression model. It can reflect the sample's base rate for the dependent variable. However, this depends on what other predictors happen to be in the model. If all the other predictors are "Go" predictors (ones with $\beta > 0$, where an increase in predictor score reflects an increase in the DV), then the Constant may take on a contrarian role and assume $\beta < 0$). Whereas, had all the predictors been "No-go," with $\beta < 0$, then the Constant may have a $\beta > 0$. In mixed models, with both Go and No-go predictors, things could go either way. Therefore, interpretation of the Constant has to be approached with skill and caution.

Categorical Variables and the Use of Contrasts

The analytic usefulness of logistic regression is a big plus. What is not a plus is the meticulous care that has to go into coding the data, setting up the analysis, and interpreting the results.

For one thing, the technique is susceptible to outliers, as we mentioned. Misentry of even a single data point can wreck an analysis.

Another serious difficulty lies in the use of categorical predictors. Although logistic regression is technically capable of handling both categorical and continuous variables, special care needs to be taken when using categoricals. As long as all variables are continuous, either ordinal or ratio-scale, no special care needs to be taken. But categoricals are different. This is because the program takes categoricals coded as letters and converts them internally into zeros and ones. For example, we had two experimenters involved in running the participants. Call them "H" and "B." During the SPSS analysis, experimenter "H" is internally converted by the program into either a zero or a one, in order to be plugged into equation 2. This conversion introduces the opportunity for serious conceptual errors to be made if we are not scrupulous in coding in the data, thinking out our analysis, and interpreting the results.

To drive this idea home, let us take this example further. If, during the analysis, we fail to specify the variable EXPTR as categorical (which requires bringing up a dialog box and making some adjustments), then we could be making a large mistake. That is because SPSS has automatic defaults and will change *any* letter into a number, whether or not we understand what it is doing. So look at the equation—trying to treat "H" as "nothing" and "B" as "one unit of something" makes sense only in a very limited context. And, say we run a model containing an in-

teraction. What the mathematics actually does is eliminate the effect of ALL the predictor scores in that interaction term whenever it calculates a data point involving "H," because it multiplies the other variables in that interaction term by zero for that data point. And this is something we might not have intended to do exactly that way. This is the way we do contrasts, but the point is that the program can be doing a contrast we do not know it is doing if we do not understand exactly what is happening mathematically.

Looking at some actual SPSS output will make this a little clearer. Below is some output for the simple model EXPTR + Constant. In the first case, "H" was set to internally code as "1" and "B" as "0." In the second case, those codes were reversed.

You can see that the statistical significance (Sig.) of EXPTR does not change, and that this particular variable did not produce a reliable effect (.593). The betas for EXPTR are the same, just with opposite sign. This is simply because we are logically testing one thing "A" against another thing "not A" and, because there are only two things, so "not A" has no degrees of freedom. But, as we would expect from the math and the iterative computational algorithm we talked about, the constants turn out to be different. This is because, in the first instance, "H" was exerting the

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	EXPTR(1)	-.288	.539	.285	1	.593	.750
	Constant	-.223	.335	.443	1	.506	.800

a. Variable(s) entered on step 1: EXPTR.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	EXPTR(1)	.288	.539	.285	1	.593	1.333
	Constant	-.511	.422	1.468	1	.226	.600

a. Variable(s) entered on step 1: EXPTR.

primary mathematical effect, since B = 0, whereas B = 1 in the second. But the constant is being updated always. Consequently, the two models are related, and similar, but the first one is really measuring "the effect on takeoffs of a pilot's being run by 'H'," as opposed to being run by 'not H' (i.e. 'B')." The negative beta in the first case means "pilots were less likely, on the whole, to take off if they were run by 'H' than by 'not H'" (although recall that p is not reliable, so we would not ultimately assert any difference) In this case, it so happens that 'not H' has to mean "B," but that was only because there were only two experimenters. Had there been three, we would have had to test a third contrast, and each would have tested primarily the effect of that one experimenter, set up consciously by us to code as "1."

Things get even more interesting when it comes to interactions involving more than one categorical variable. The essential logic remains the same, however: a) contrasts focus on whatever happens to be coded "1," and b) interactions go to zero whenever any single term in them becomes zero. The bottom line is that we cannot simply mindlessly run SPSS and hope to understand the data.

Problems Associated With Logistic Regression

Like all statistics, logistic regression is not a perfect technique (Tabachnick & Fidell, 2000). Some of its weaknesses include

1. *Correlation does not imply causation.* All regression techniques do is to establish a mathematical relation between the presence/absence of one thing and the presence/absence of another. But such correlation does not necessarily mean, for instance, that Factor A *causes* Factor B. The classic counterexample is the case where Factor A and Factor B are both caused by Factor C. In that case, A and B still show correlation, but there is no causation whatsoever between A and B.

2. *Outliers can greatly skew models and parameter estimates.* We demonstrated this clearly in Appendix E. Fortunately, this problem was easily overcome by winsorizing the data.
3. *Independence of samples is assumed.* Logistic regression is basically a between-subjects technique, not for repeated measures gathered over time. That was not a factor in this study, however.
4. *Absence of multicollinearity is assumed.* If predictors are highly correlated, they are probably measuring the same factor, and will not contribute much, if anything additional to a model, other than wrongly inflated significance. Fortunately, the models we present did not pose this problem (see Appendix G for the intercorrelation matrices).
5. *The ratio of cases to model predictors is important.* A common rule of thumb, seen in many textbooks, is that a model should contain no more than one predictor per 10 cases (e.g., per 10 pilots). If a constant is used, this should be counted as one predictor However, we noticed an ancillary problem during this analysis, namely
6. *The case-to-predictor ratio issue extends to the number of predictors measured <u>before</u> analysis is commenced.* This is discussed in greater depth below, and in the Part II report.

Problems Associated With Too Many Predictors in Forward Stepwise Logistic Regression

At some point, we had the intuition that simply trying to examine too many predictors in our primary technique of forward stepwise regression could introduce a combinatorial problem. That theoretical problem is easiest illustrated using our actual situation. We started with 83 candidate predictors, some of which were eventually eliminated due to reasons such as having missing values or being discrete (which often led to unwieldy combinations of contrasts). So, in the end, we looked at roughly 60 predictors.

Now, consider the following deductive logic: Suppose you were trying to model some data taken from 30 pilots, upon whom you had 60 measurements (predictors) each. This would correspond to, say, our Low Financial Incentive group. Then the rule of thumb we mentioned above in Point 5 suggests that all such models should have no more than 30/10 = 3 predictors. So far, so good.

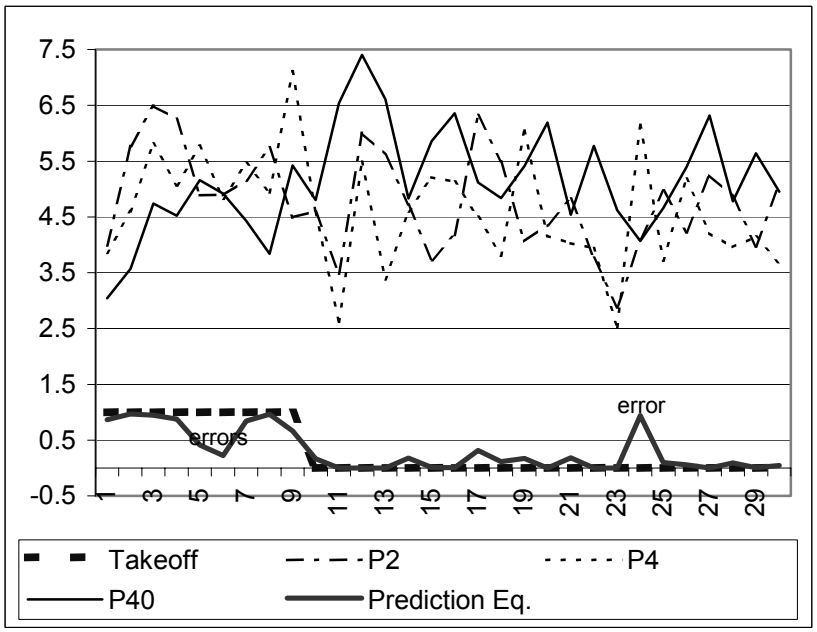
The problem comes when we consider random numbers. Suppose every one of our predictors was simply "noise," taken randomly from a Gaussian (normal, bell-shaped) distribution of numbers. Given that the logistic regression prediction equation is basically

$$p_{event} = \frac{1}{1+e^{-((\beta_0+)\beta_1 X_1+\beta_2 X_2(+\beta_3 X_3))}} \qquad (1)$$

notice how the exponent term $-((\beta_0 +) \beta_1 X_1 ...)$ is really a sum. It will be the sum of our predictors (each weighted). That means that, whatever the actual numbers are for each pilot's predictor scores, we are going to weight them, then add them up to form a total, which will then be plugged into Eq. 1. So what are the chances that, <u>given nothing but random numbers</u>, SPSS will ultimately end up finding the precise set of β weights such that the Equation 1 turns out greater than 0.5 for pilots who subsequently took off, versus a predicted score of less than 0.5 for those who did not?

Shockingly, the answer is that it is highly likely. We verified this by running Monte Carlo simulations, a standard technique in statistics. Using normal random number generation with μ (mu, mean) of 5 and σ (sigma, standard deviation) of 1, we were easily able to duplicate results such as the following:

This illustrates that SPSS essentially "made sense out of nonsense." It summed the three random pseudo-predictor scores for each pilot, shown by the three jagged curves, multiplying each score by the β weights it derived, inserted them into Equation 1 and came up with the much-more regular solid "Prediction Equation" line. Notice how closely that matched the thick, dashed "Takeoff" line representing a dependent variable score of 1 for a takeoff and 0 for a non-takeoff. The three points where those two curves did not closely correspond are labeled as "error." Since 27 of the 30 cases were "predicted" correctly, this model's predictivity was .90.



Why is this? Well, look at it from the model's point of view. In forward stepwise regression, the software first chooses one predictor to enter into the model, then a second, a third, and so forth. With 60 predictors, it has 60 candidates for the first choice, 59 for the second, 58 for the third, and so on. Thus, in a three-single predictor model, there are 60*59*58/(3*2*1) = 34,220 possible combinations,[1] not even counting interactions. What is happening is that, given such a huge number of combinations, at least one is highly likely to form a highly "predictive" sum, even though, purely taken individually, none of the individual scores has any meaning whatsoever.

We did 100 Monte Carlo simulations for each of our low- and high-incentive groups, with and without a constant in the model. While this was well under the usual standard of 1000-10000 or so simulations per condition, doing each simulation was quite tedious, and these 400 runs did have sufficient reliability to illustrate our basic points.

Here we see that the proportion of takeoffs matters. Noise models with a proportion of takeoffs close to .5 show lower predictivity and Nagelkerke R^2 than ones with a proportion of takeoff equal to .3. But, overall, predictivities were still in the 70-90% range, and R^2*s* in the 40-70% range for these random-number models.

Evaluation of the Meaningfulness of Our Data

So how reliable were the conclusions for this Part I report?

	Low Fin. Incentive group		High Fin. Incentive group	
	models with constant		models with constant	
	Predictivity	Nagel R^2	Predictivity	Nagel R^2
$\mu_{MonteCarlo}$	80.4	0.36	76.3	0.48
CI 95	**≅.89**	**≅.59**	**≅.85**	**≅.64**
$\mu_{ActualData}$	85.7	0.52	75	0.28
α estimated	0.16	0.08	NS	NS

The method used to derive these estimates is detailed in the companion report *Pilot willingness to take off into marginal weather, Part II: Antecedent overfitting with forward stepwise logistic regression.*

[1] The reason for the denominator is that the order of terms in the model makes no difference. SPSS logistic treats "ABC" the same as "ACB," "BAC," "BCA," "CAB," and "CBA"—three degrees of freedom for the first choice, two for the second, and one for the last.

To summarize, the .95 confidence intervals around the predicted means (CI .95) imply that any model exceeding these estimates for predictivity and R^2 is highly likely to be a better-than-chance model. Confidence intervals are a standard approach used in many statistics.

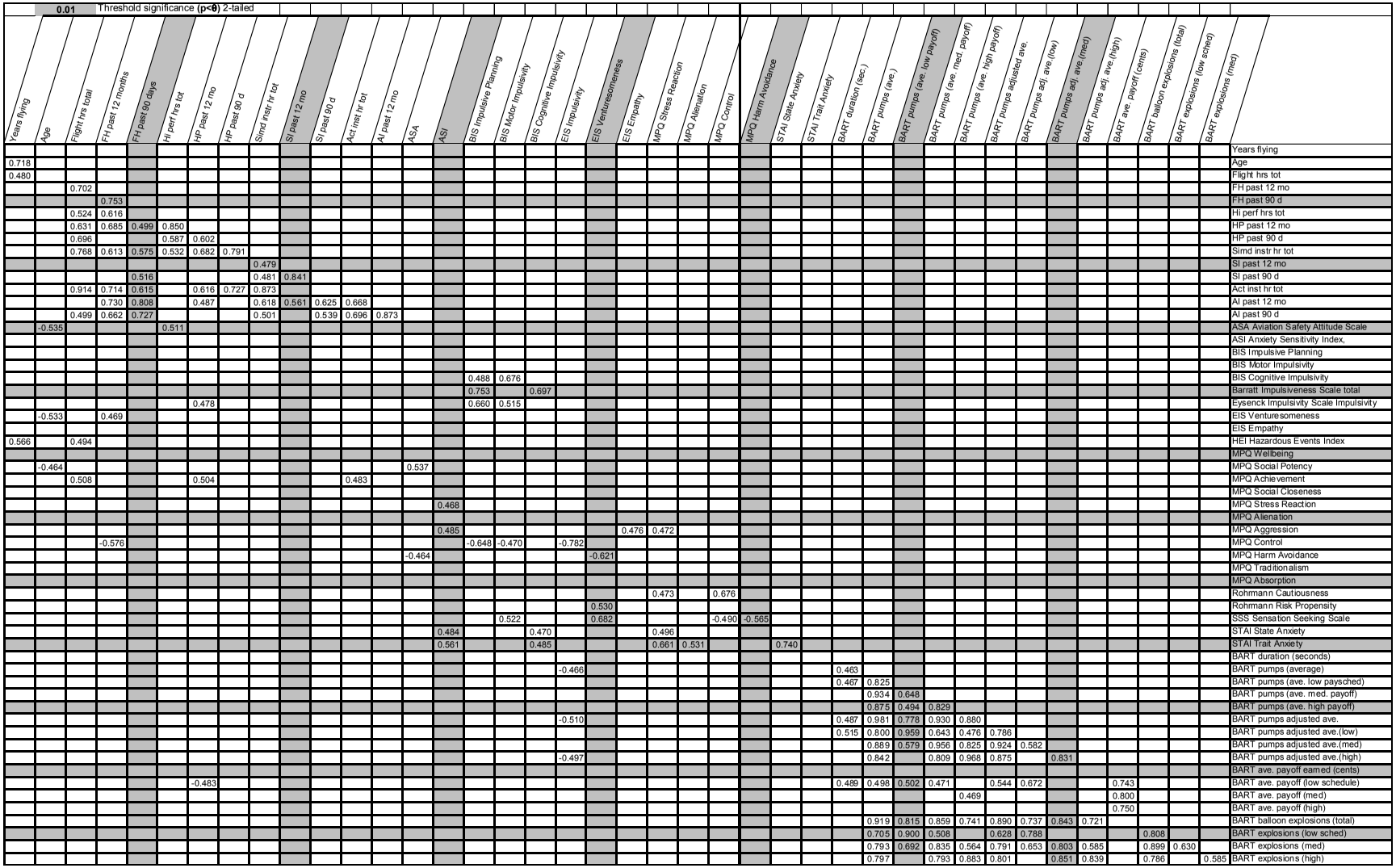
The full data set and high incentive models derived from real pilots' data in the current study did not differ significantly from what could be expected from random number simulations. That is why we limited our primary observations to high-level conclusions and the Low Incentive data. The real-pilots' low-incentive 85.7% predictivity did exceed the random-generated Monte Carlo mean of 80.4%, although it did not top the estimate of 89% for the .95 CI. Their Nagelkerke R^2 of .52 considerably bested the Monte Carlo mean of .36 and came close to meeting the .95 CI of .59. So, judging from the Monte Carlo scatterplots (shown in Part II), reliability for the low incentive n=28 experimental data was roughly α =.16 for predictivity and α =.08 for R^2.

As said previously, for the purposes of a preliminary report such as this, it is often wiser to be somewhat relaxed in reporting results than we would be later on in the research process. This is because of the Type I-Type II error tradeoff, that is, where excessive stringency in setting significance levels results in a lower number of false positive results but strictly at the cost of a higher number of missed results. In other words, at first the strategy involves going for breadth of findings. The small number of results that fail to be reliable will be discovered and eliminated as other studies cross check results presented here.

APPENDIX G

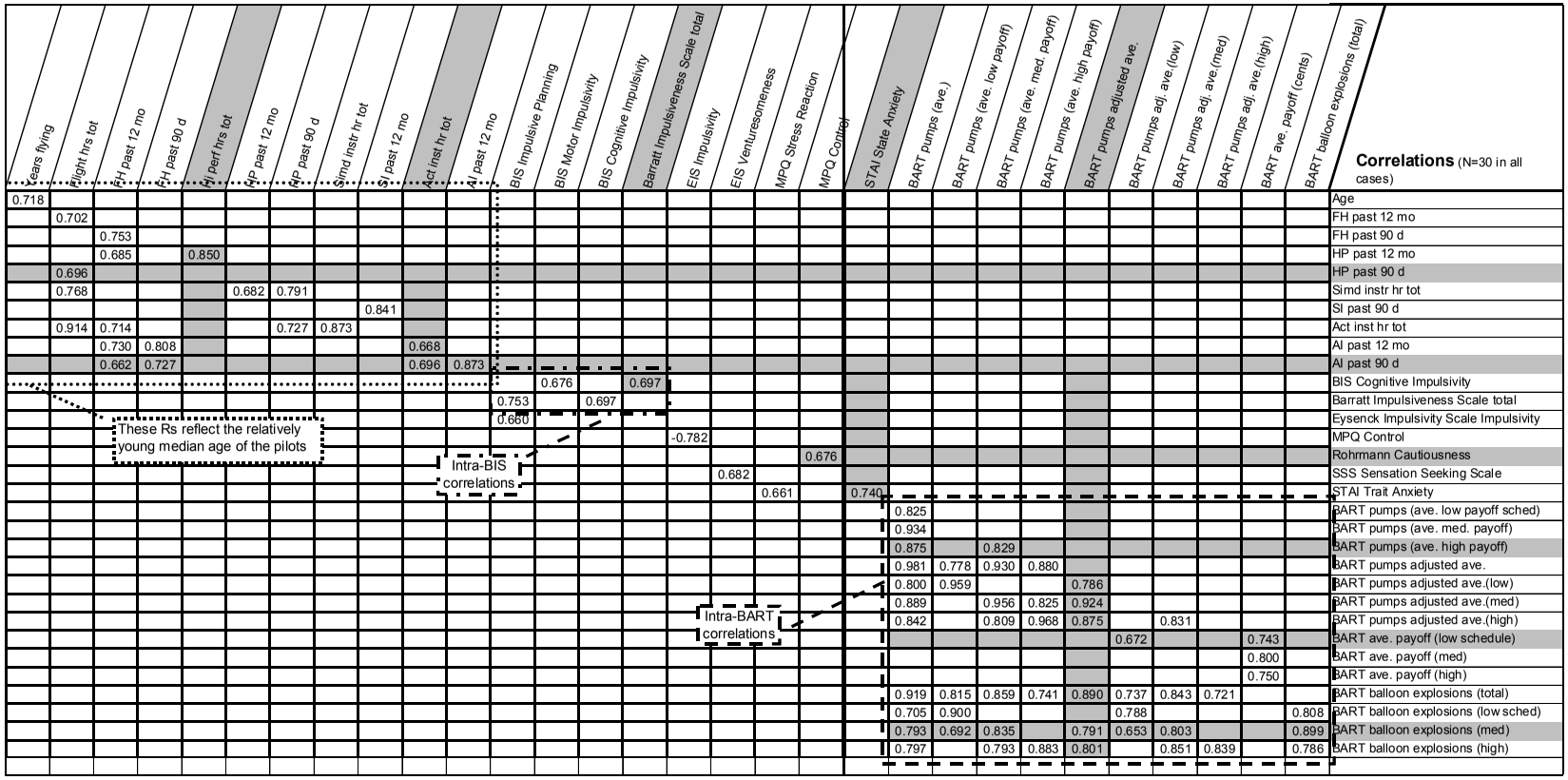
	Years flying	Flight hrs tot	FH past 12 mo	FH past 90 d	Hi perf hrs tot	HP past 12 mo	Simd instr hr tot	SI past 12 mo	SI past 90 d	Act inst hr tot	AI past 12 mo	AI past 90 d	Anxiety Sensitivity Index	BIS Impulsive Planning	BIS Motor Impulsivity	BIS Cognitive Impulsivity	BIS Total	EIS Impulsivity	EIS Venturesomeness	MPQ Stress Reaction	MPQ Control	MPQ Harm Avoidance	Rohrmann Risk Propensity	STAI State Anxiety	BART pumps (ave)	BART pumps (ave lo pay)	BART pumps (ave med pay)	BART pumps (ave hi pay)	BART pumps adj ave	BART pumps adj ave (lo)	BART pumps adj ave (med)	BART pumps adj ave (hi)	BART ave pay (cents)	BART explosions (tot)	BART explosions (lo)	Wald p (model w. constant)
Age	0.686	0.619																																		
FH past 12 mo			0.737																																	
FH past 90 d			0.591	0.482																																
Hi perf hrs tot				0.631	0.744																															
HP past 12 mo					0.541	0.767																														
HP past 90 d																																				
Simd instr hr tot			0.673	0.686	0.612																															
SI past 12 mo					0.520		0.567																													
SI past 90 d					0.517			0.828																												
Act inst hr tot			0.904	0.481		0.581	0.642																													
AI past 12 mo				0.693	0.656		0.559		0.500																											
AI past 90 d				0.623	0.605							0.915																								
BIS Cognitive Impulsivity													0.560	0.601																					
BIS (tot)													0.815	0.775																					
EIS Impulsivity													0.558	0.578			0.488																		
Hazardous Events Index			0.604		0.508				0.597				-0.576	-0.576		-0.488	-0.766																		
MPQ Control																		-0.567																	
MPQ Harm Avoidance																				0.532															
Rohrmann Cautiousness																																				
Rohrmann Risk Propensity																																				
Sensation Seeking Scale													0.487	0.532																					
STAI State Anxiety																																				
STAI Trait Anxiety													0.532																							
BART pumps (ave lo pay)													0.587			0.578								0.669											
BART pumps (ave med pay)																				0.648				0.845											
BART pumps (ave hi pay)																								0.937	0.725										
BART pumps adj ave																								0.829		0.732		0.831							
BART pumps adj ave(lo)																								0.982	0.810	0.924	0.831								
BART pumps adj ave (med)																								0.822	0.967	0.707		0.818							
BART pumps adj ave (hi)																								0.908	0.679	0.964	0.740	0.930	0.673						
BART ave pay (cents)																								0.793		0.701	0.977	0.813		0.726					
BART ave pay (lo)																								0.554		0.523	0.574	0.576			0.538				
BART ave pay (med)																								0.528	0.581	0.496		0.580	0.712			0.672			
BART ave pay (hi)																																	0.805		
BART explosions (tot)																																		0.512	0.697	
BART explosions (lo)																								0.913	0.837	0.858	0.684	0.879	0.777	0.844	0.655				
BART explosions (med)																								0.749	0.878	0.623		0.677	0.775	0.568		0.841			
BART explosions (hi)																								0.761	0.729	0.789		0.744	0.693	0.755		0.890	0.681		
																								0.753		0.705	0.832	0.759		0.769	0.780		0.735		

Intercorrelation matrix, Pearson R, all Ss (N=60), all variables with significance $p < .0001$ ($R^2 > .23$).



G2

Intercorrelation matrix, Pearson R, all variables with significance $p < .01$, Low Financial Incentive group (N=30).



Correlations (N=30 in all cases)	Years flying	Flight hrs tot	FH past 12 mo	FH past 90 d	AI perf hrs tot	HP past 12 mo	HP past 90 d	Simd instr hr tot	SI past 12 mo	Act inst hr tot	AI past 12 mo	BIS Impulsive Planning	BIS Motor Impulsivity	BIS Cognitive Impulsivity	Barratt Impulsiveness Scale total	EIS Impulsivity	EIS Venturesomeness	MPQ Stress Reaction	MPQ Control	STAI State Anxiety	BART pumps (ave.)	BART pumps (ave. low payoff)	BART pumps (ave. med. payoff)	BART pumps (ave. high payoff)	BART pumps adjusted ave.	BART pumps adj. ave.(low)	BART pumps adj. ave.(med)	BART pumps adj. ave.(high)	BART ave. payoff (cents)	BART balloon explosions (total)
Age	0.718																													
FH past 12 mo		0.702																												
FH past 90 d			0.753																											
HP past 12 mo			0.685		0.850																									
HP past 90 d		0.696																												
Simd instr hr tot		0.768				0.682	0.791																							
SI past 90 d									0.841																					
Act inst hr tot		0.914	0.714				0.727	0.873																						
AI past 12 mo			0.730	0.808						0.668																				
AI past 90 d			0.662	0.727						0.696	0.873																			
BIS Cognitive Impulsivity													0.676		0.697															
Barratt Impulsiveness Scale total												0.753		0.697																
Eysenck Impulsivity Scale Impulsivity												0.660																		
MPQ Control																-0.782														
Rohrmann Cautiousness																		0.676												
SSS Sensation Seeking Scale																	0.682													
STAI Trait Anxiety																		0.661		0.740										
BART pumps (ave. low payoff sched)																				0.825										
BART pumps (ave. med. payoff)																				0.934										
BART pumps (ave. high payoff)																				0.875		0.829								
BART pumps adjusted ave.																				0.981	0.778	0.930	0.880							
BART pumps adjusted ave.(low)																				0.800	0.959			0.786						
BART pumps adjusted ave.(med)																				0.889		0.956	0.825	0.924						
BART pumps adjusted ave.(high)																				0.842		0.809	0.968	0.875		0.831				
BART ave. payoff (low schedule)																									0.672			0.743		
BART ave. payoff (med)																												0.800		
BART ave. payoff (high)																												0.750		
BART balloon explosions (total)																				0.919	0.815	0.859	0.741	0.890	0.737	0.843	0.721			
BART balloon explosions (low sched)																				0.705	0.900				0.788				0.808	
BART balloon explosions (med)																				0.793	0.692	0.835		0.791	0.653	0.803			0.899	
BART balloon explosions (high)																				0.797		0.793	0.883	0.801		0.851	0.839		0.786	

Intercorrelation matrix, Pearson R, all variables with significance $p < .0001$ (equivalent to $R^2 > .44$), Low Financial Incentive group (N=30). Note that most of the highly significant correlations have very simple explanations

BIS Impulsive Planning	EIS Impulsivity	EIS Venturesomeness	MPQ Stress Reaction	MPQ Control	STAI State Anxiety	**Non-Evident Correlations** where $R^2 > .44$ (N=30 in all cases)
0.660						Eysenck Impulsivity Scale Impulsivity
	-0.782					MPQ Control
				0.676		Rohrmann Cautiousness
		0.682				Sensation Seeking Scale
			0.661		0.740	STAI Trait Anxiety

Pearson Rs, variables with significance $p < .0001$ (equivalent to $.44 \leq R^2 \leq .61$) whose explanation is not obvious simply because they are correlated by their very nature (e.g. the various measures calculated from BART). The upshot here is that a) Each of these correlations is perfectly logical, and; b) Even this small number of correlations involves less than half the variance. That means that each instrument presumably measured different factors for the most fact, which was as it should be.

APPENDIX H

Description	Name	Sig.	g	MVs	Description	Name	Sig.
Order in which S. was run	runorder	0.982			Aviation Safety Attitude scale	asa	0.651
Subject ID#	**idnum**				Anxiety Sensitivity Index, total score	asi	0.143
Takeoff (Y/N)	**takeoff**				Barratt Impulsiveness Scale (BIS-10)--Impulsive Planning scale	imp_plan	0.902
Latency (minutes)	**latcy**				BIS--Motor Impulsivity scale	imp_motr	0.960
Visibility	vis	0.113			BIS-Cognitive Impulsivity scale	imp_cog	0.886
Ceiling	ceil	0.433			BIS--total score	bis	0.896
Incentive	incent	0.070			Eysenck Impulsivity Scale (EIS)--Impulsiveness scale	impuls	0.705
Experimenter	exptr	0.593			EIS--Venturesomeness scale	ventur	0.088
Yr started flying	yr_flg	0.785			EIS--Empathy scale	empath	0.277
Years flying, total	yrs_flg	0.966	2		Hazardous Events Index	hei	0.560
	age	0.653	1		Multidimensional Personality Questionnaire, Brief Form (MPQ-BF)		
	gender	0.461			MPQ--Wellbeing scale	wellbe	0.870
Type of License	type_lic	0.612			MPQ--Social Potency scale	socpot	0.468
Instrument Rating	inst_rtg	0.193			MPQ--Achievement scale	achieve	0.492
Total flight hours	fh_tot	0.410	2		MPQ--Social Closeness scale	socclose	0.290
Flthrs past year	fh_12m	0.536	2		MPQ--Stress Reaction scale	stress	0.528
Flthrs past 90 days	fh_90d	0.444	1		MPQ--Alienation scale	alienate	0.677
High-performance A/C, tot hrs	hp_tot	0.476	2		MPQ--Aggression scale	aggress	0.248
HP last 12 mo	hp_12m	0.287	2		MPQ--Control scale	control	0.540
HP last 90 days	hp_90d	0.151	2		MPQ--Harm Avoidance scale	harmav	0.614
Simulated instrument hrs total	si_tot	0.440	2		MPQ--Traditionalism scale	tradit	0.657
Sim hr last 12 mo	si_12m	0.239	2		MPQ--Absorption scale	absorpt	0.879
Sim hr last 90 d	si_90d	0.235	2		Rohrmann Risk Orientation Questionnaire--Cautiousness scale	roq_c	0.868
Actual instrument hrs, total	ai_tot	0.467	2		Rohrmann Risk Orientation Questionnaire--Risk Propensity scale	roq_p	0.225
AI last 12 mo	ai_12m	0.776	2		Sensation-Seeking Scale	sss	0.886
AI last 90 d	ai_90d	0.868	2		State-Trait Anxiety Inventory--State	anx_st	0.853
Personal visibility minimum	vis_min	0.386	2	1	State-Trait Anxiety Inventory--Trait	anx_tr	0.736
Personal ceiling minimum	ceil_min	0.955	2	1	Balloon Analogue Risk Task (BART)--test duration	durn_sec	0.565
Do you adjust minima?	min_adj	0.398		3	BART--average pumps	pumpsav	0.335
Flown this route before?	fln_rt	0.427		3	BART--average pumps, low-payoff condition	pmpavglo	0.465
Distance through wx imp?	dthruwx	0.813		3	BART--average pumps, medium-payoff condition	pmpavgm	0.630
$ bonus motivating? (Hi Incent only)	buck_mot	0.071		**32**	BART--average pumps, high-payoff condition	pmpavghi	0.198
Passengers change TO willingness	pass_mot	0.837		4	BART--adjusted average	pmpadjav	0.373
More flt hrs change TO willingness	fhincmot	0.893		5	BART--adjusted average, low-pay condn	padjavlo	0.782
Mission-critical chg. w-ness?	missnmot	0.020		**13**	BART--adjusted average, med-pay condn	padjavme	0.868
Flown Malibu chg w-ness?	mal_sub	0.840		**10**	BART--adjusted average, high-pay condn	padjavhi	0.207
Being a simulator chg w-ness?	sim_mot	0.127		**9**	BART--total payoff (cents)	pay_tot	0.790
...more specifically (re prev Q)	simmotsb	0.138		**14**	BART--total payoff, low-pay cond'n	pay_low	0.979
How significant was the $ to you?	buck_sig	0.164		**9**	BART--total payoff, med-pay cond'n	pay_med	0.630
Would crash embarrass you?	crashsig	0.048		**20**	BART--total payoff, high-pay cond'n	pay_hi	0.304
How much did you consider traffic'	tx_mot	0.919		**49**	BART--total balloon explosions	bang_tot	0.259
Ever had a bad wx experience?	badwx	0.318		5	BART--explosions, low-pay cond'n	bang_low	0.422
					BART--explosions, med-pay cond'n	bang_me	0.325
					BART--explosions, high-pay cond'n	bang_hi	0.305

Predictor significances for the full data set (N=60), showing the reliability (expressed by the Wald *p*-value) of individual-predictor models (plus Constant) in logistic regression analysis with *Takeoff* as the dependent variable. Here each model included just one predictor, plus a constant. Subject ID is an identifier, not a predictor, and Latency is a descriptor, hence these lack *p*-values.

These predictors generally show very low reliability, with the exception of those highlighted in gray. However, of those, we should exclude all but *incent* and *ventur* from further consideration, due to high numbers of missing values (MV) for the other three. Note that the reference category for *lic_type* was "Private" (N=39), so *p* expresses the analysis "Private versus All Other Categories." No individual category had a *p* of < .12 in any case.

APPENDIX I

Description	Name	Sig.	g	MVs	Description	Name	Sig.
Order in which S. was run	runorder	0.675			Aviation Safety Attitude scale	asa	0.645
Subject ID#	**idnum**				Anxiety Sensitivity Index, total score	asi	0.127
Takeoff (Y/N)	**takeoff**				Barratt Impulsiveness Scale (BIS-10)--Impulsive Planning scale	imp_plan	0.615
Latency (minutes)	**latcy**				BIS--Motor Impulsivity scale	imp_motr	0.957
Visibility	vis	0.064			BIS-Cognitive Impulsivity scale	imp_cog	0.398
Ceiling	ceil	0.691			BIS--total score	bis	0.562
Incentive	incent				Eysenck Impulsivity Scale (EIS)--Impulsiveness scale	impuls	0.394
Experimenter	exptr	0.261			EIS--Venturesomeness scale	ventur	0.713
Yr started flying	yr_flg				EIS--Empathy scale	empath	0.881
Years flying, total	yrs_flg	0.470	2		Hazardous Events Index	hei	0.221
	age	0.942	1		Multidimensional Personality Questionnaire, Brief Form (MPQ-BF)		
	gender	0.815			MPQ--Wellbeing scale	wellbe	0.896
Type of License	type_lic	0.999			MPQ--Social Potency scale	socpot	0.269
Instrument Rating	inst_rtg	0.873			MPQ--Achievement scale	achieve	0.574
Total flight hours	fh_tot	0.591	1		MPQ--Social Closeness scale	socclose	0.590
Flthrs past year	fh_12m	0.911	2		MPQ--Stress Reaction scale	stress	0.544
Flthrs past 90 days	fh_90d	0.907			MPQ--Alienation scale	alienate	0.787
High-performance A/C, tot hrs	hp_tot	0.347	2		MPQ--Aggression scale	aggress	0.673
HP last 12 mo	hp_12m	0.713	2		MPQ--Control scale	control	0.930
HP last 90 days	hp_90d	0.328	2		MPQ--Harm Avoidance scale	harmav	0.641
Simulated instrument hrs total	si_tot	0.995			MPQ--Traditionalism scale	tradit	0.203
Sim hr last 12 mo	si_12m	0.588	2		MPQ--Absorption scale	absorpt	0.961
Sim hr last 90 d	si_90d	0.982	2		Rohrmann Risk Orientation Questionnaire--Cautiousness scale	roq_c	0.345
Actual instrument hrs, total	ai_tot	0.482	2		Rohrmann Risk Orientation Questionnaire--Risk Propensity scale	roq_p	0.637
AI last 12 mo	ai_12m	0.753	1		Sensation-Seeking Scale	sss	0.888
AI last 90 d	ai_90d	0.512	1		State-Trait Anxiety Inventory--State	anx_st	0.484
Personal visibility minimum	vis_min	0.523			State-Trait Anxiety Inventory--Trait	anx_tr	0.393
Personal ceiling minimum	ceil_min	0.487	1		Balloon Analogue Risk Task (BART)--test duration	durn_se	0.864
Do you adjust minima?	min_adj	0.244			BART--average pumps	pumpsav	0.341
Flown this route before?	fln_rt	0.265			BART--average pumps, low-payoff condition	pmpavgl	0.552
Distance through wx imp?	dthruwx	0.627			BART--average pumps, medium-payoff condition	pmpavgr	0.462
					BART--average pumps, high-payoff condition	pmpavgh	0.234
Passengers change TO willingness	pass_mot	0.175		1	BART--adjusted average	pmpadja	0.460
More flt hrs change TO willingness	fhincmot	0.204			BART--adjusted average, low-pay condn	padjavlo	0.975
Mission-critical chg. w-ness?	missnmot	0.024		**7**	BART--adjusted average, med-pay condn	padjavm	0.768
Flown Malibu chg w-ness?	mal_sub	0.854		**4**	BART--adjusted average, high-pay condn	padjavhi	0.186
Being a simulator chg w-ness?	sim_mot	0.910		**3**	BART--total payoff (cents)	pay_tot	0.749
...more specifically (re prev Q)	simmotsb	0.408		**7**	BART--total payoff, low-pay cond'n	pay_low	0.836
					BART--total payoff, med-pay cond'n	pay_mec	0.990
Would crash embarrass you?	crashsig	0.337		**12**	BART--total payoff, high-pay cond'n	pay_hi	0.365
How much did you consider traffic	tx_mot	0.422		**26**	BART--total balloon explosions	bang_tot	0.272
Ever had a bad wx experience?	badwx	0.472			BART--explosions, low-pay cond'n	bang_lov	0.458
					BART--explosions, med-pay cond'n	bang_me	0.403
					BART--explosions, high-pay cond'n	bang_hi	0.229

Predictor significances for the Low-Incentive data set (N=30), showing the reliability (Wald *p*-value) of individual-predictor models (plus Constant) in logistic regression analysis with *Takeoff* as the dependent variable. The reference category on *type_lic* is "Private," on *simmotsb* it is "Didn't matter." Keep in mind that the SPSS reference category is the one being weighted "0" in the logistic regression prediction equation.

APPENDIX J

Description	Name	Sig.	g	MVs	Description	Name	Sig.
Order in which S. was run	runorder	0.612			Aviation Safety Attitude scale	asa	0.700
Subject ID#	**idnum**				Anxiety Sensitivity Index, total score	asi	0.910
Takeoff (Y/N)	**takeoff**				Barratt Impulsiveness Scale (BIS-10)--Impulsive Planning scale	imp_plan	0.809
Latency (minutes)	**latcy**				BIS--Motor Impulsivity scale	imp_motr	0.820
Visibility	vis	0.655			BIS-Cognitive Impulsivity scale	imp_cog	0.595
Ceiling	ceil	0.466			BIS--total score	bis	0.829
Incentive	incent	N/A			Eysenck Impulsivity Scale (EIS)--Impulsiveness scale	impuls	0.668
Weather severity	wxsvrty	0.364			EIS--Venturesomeness scale	ventur	0.085
Experimenter	exptr	0.080			EIS--Empathy scale	empath	0.296
Years flying, total	yrs_flg	0.451			Hazardous Events Index	hei	0.976
	age	0.420	1		Multidimensional Personality Questionnaire, Brief Form (MPQ-BF)		
	gender	0.476			MPQ--Wellbeing scale	wellbe	0.980
Type of License	type_lic	0.933			MPQ--Social Potency scale	socpot	0.947
Instrument Rating	inst_rtg	0.069			MPQ--Achievement scale	achieve	0.735
Total flight hours	fh_tot	0.420	2		MPQ--Social Closeness scale	socclose	0.117
Flthrs past year	fh_12m	0.385	1		MPQ--Stress Reaction scale	stress	0.980
Flthrs past 90 days	fh_90d	0.192			MPQ--Alienation scale	alienate	0.304
High-performance A/C, tot hrs	hp_tot	0.333			MPQ--Aggression scale	aggress	0.267
HP last 12 mo	hp_12m	0.090	2		MPQ--Control scale	control	0.622
HP last 90 days	hp_90d	0.164	2		MPQ--Harm Avoidance scale	harmav	0.337
Simulated instrument hrs total	si_tot	0.105			MPQ--Traditionalism scale	tradit	0.079
Sim hr last 12 mo	si_12m	0.036			MPQ--Absorption scale	absorpt	0.823
Sim hr last 90 d	si_90d	0.130	1		Rohrmann Risk Orientation Questionnaire--Cautiousness scale	roq_c	0.240
Actual instrument hrs, total	ai_tot	0.625	1		Rohrmann Risk Orientation Questionnaire--Risk Propensity scale	roq_p	0.325
AI last 12 mo	ai_12m	0.481	1		Sensation-Seeking Scale	sss	0.937
AI last 90 d	ai_90d	0.201	2		State-Trait Anxiety Inventory--State	anx_st	0.161
Personal visibility minimum	vis_min	0.519		1	State-Trait Anxiety Inventory--Trait	anx_tr	0.512
Personal ceiling minimum	ceil_min	0.726	1	1	Balloon Analogue Risk Task (BART)--test duration	durn_sec	0.437
Do you adjust minima?	min_adj	0.999		3	BART--average pumps	pumpsavg	0.703
Flown this route before?	fln_rt	0.485		3	BART--average pumps, low-payoff condition	pmpavglo	0.453
Distance through wx imp?	dthruwx	0.638		3	BART--average pumps, medium-payoff condition	pmpavgme	0.812
$ bonus motivating? (Hi Incent only)	buck_mot	0.071		2	BART--average pumps, high-payoff condition	pmpavghi	0.688
Passengers change TO willingness?	pass_mot	0.323		3	BART--adjusted average	pmpadjav	0.682
More flt hrs change TO willingness?	fhincmot	0.491		5	BART--adjusted average, low-pay condn	padjavlo	0.563
Mission-critical chg. w-ness?	missnmot	0.178		6	BART--adjusted average, med-pay condn	padjavme	0.835
Flown Malibu chg w-ness?	mal_sub	0.728		6	BART--adjusted average, high-pay condn	padjavhi	0.874
Being a simulator chg w-ness?	sim_mot	0.159		6	BART--total payoff (cents)	pay_tot	0.679
...more specifically (re prev Q)	simmotsb	0.139		7	BART--total payoff, low-pay cond'n	pay_low	0.868
How significant was the $ to you?	buck_sig	0.126		6	BART--total payoff, med-pay cond'n	pay_med	0.999
Would crash embarrass you?	crashsig	0.135		8	BART--total payoff, high-pay cond'n	pay_hi	0.995
How much did you consider traffic?	tx_mot	0.999		23	BART--total balloon explosions	bang_tot	0.503
Ever had a bad wx experience?	badwx	0.679		5	BART--explosions, low-pay cond'n	bang_low	0.482
					BART--explosions, med-pay cond'n	bang_med	0.533
					BART--explosions, high-pay cond'n	bang_hi	0.783

High Incentive data, N=30, single variable (plus Constant) models. Reference category for *type_lic* is "Private" (no individual p < .187). Reference category for *simmotsb* is "Positives and negatives cancel" (no individual $p <$ .072).

APPENDIX K

Interactions *vis x ceil x factor below*								
Name	MV	Sig.	Ref		Sig.		Name	Sig.
runorder		0.189					imp_motr	0.381
							imp_cog	0.253
exptr		0.134	B	H	0.042		bis	0.392
yrs_flg		0.604					impuls	0.418
age		0.655					ventur	0.252
gender		0.942	M	F	0.444		empath	0.579
type_lic		0.788	Private				hei	0.466
inst_rtg		0.375	N	Y	0.131		w ellbe	0.482
fh_tot		0.282					socpot	0.512
fh_12m		0.559					achieve	0.606
fh_90d		0.473					socclose	0.938
hp_tot		0.380					stress	0.362
hp_12m		0.230					alienate	0.123
hp_90d		0.717					aggress	0.083
si_tot		0.244					control	0.772
si_12m		0.379					harmav	0.764
si_90d		0.802					tradit	0.054
ai_tot		0.305					absorpt	0.896
ai_12m		0.445					roq_c	0.201
ai_90d		0.177					roq_p	0.345
vis_min	1	0.308					sss	0.474
ceil_min	1	0.398					anx_st	0.179
min_adj	*3*	0.101	N	Y	0.997		anx_tr	0.235
fln_rt	*3*	0.363	N	Y	0.093		durn_sec	0.207
dthruwx	*3*	0.859	N	Y	0.961		pumpsavg	0.419
buck_mot	2	0.052	N	Y	0.137		pmpavglo	0.462
pass_mot	3	0.388					pmpavgme	0.568
fhincmot	5	0.194					pmpavghi	0.281
missnmot	6	0.291					pmpadjav	0.372
mal_sub	6	0.754					padjavlo	0.461
sim_mot	6	0.211					padjavme	0.519
simmotsb	7	0.316	"+/- cancel"				padjavhi	0.314
buck_sig	6	0.523					pay_tot	0.453
crashsig	8	0.227					pay_low	0.427
tx_mot	23	0.653					pay_med	0.568
badwx	5	0.806	N	Y	0.658		pay_hi	0.398
							bang_tot	0.438
asa		0.480					bang_low	0.553
asi		0.288					bang_med	0.571
imp_plan		0.643					bang_hi	0.278

High Incentive group only, N = 30. Reference category for *type_lic* was "Private," and results reflect composite significance for all license types. In no case was $p < .436$ for license type Reference category for *simmotsb* was "Positives and negatives cancelled." We were unable to coerce SPSS into defining the reference category as "Didn't matter." SPSS apparently sorts categoricals into frequency counts and assigns its "First" and "Last" categories according to frequency, rather than to the order in which categories are coded. In other words, recoding makes no difference. And, since its only options for assigning reference are "First" or "Last," it was impossible to equilibrate the analysis of *simmotsb* with its Low Incentive counterpart. In any event, the composite significance of *simmotsb* and all its components were all $> .165$, so the matter is irrelevant

APPENDIX L

Interactions *buck_mot x factor below*												
Name	MV	Sig.	Ref		Sig.		Name	MV	Sig.	Ref		Sig.
runorder	2	0.078	N	Y	0.144		asa	2	0.066	N	Y	0.061
vis	2	0.060	N	Y	0.125		asi	2	0.083	N	Y	0.187
ceil	2	0.064	N	Y	0.112		imp_plan	2	0.055	N	Y	0.053
exptr	2	0.064	NH	YH	0.916		imp_motr	2	0.049	N	Y	0.051
		0.877	NB	YB	0.068		imp_cog	2	0.048	N	Y	0.080
yrs_flg	2	0.531	N	Y	0.253		bis	2	0.041	N	Y	0.050
age	2	0.129	N	Y	0.083		impuls	2	0.020	N	Y	0.101
gender	2	0.469	NM	YM	0.706		ventur	2	0.021	N	Y	0.139
		0.030	NF	YF	0.071		empath	2	0.200	N	Y	0.120
type_lic	2	0.974	NP	YP	0.999		hei	2	0.301	N	Y	0.194
inst_rtg	2	0.811	NN	YN	0.066		wellbe	2	0.104	N	Y	0.135
		0.047	NY	YY	0.877		socpot	2	0.116	N	Y	0.167
fh_tot	2	0.231	N	Y	0.134		achieve	2	0.045	N	Y	0.069
fh_12m	2	0.670	N	Y	0.225		socclose	2	0.321	N	Y	0.061
fh_90d	2	0.904	N	Y	0.253		stress	2	0.181	N	Y	0.227
hp_tot	2	0.897	N	Y	0.152		alienate	2	0.156	N	Y	0.527
hp_12m	2	0.860	N	Y	0.103		aggress	2	0.127	N	Y	0.568
hp_90d	2	0.649	N	Y	0.070		control	2	0.363	N	Y	0.248
si_tot	2	0.917	N	Y	0.113		harmav	2	0.801	N	Y	0.088
si_12m	2	0.200	N	Y	0.244		tradit	2	0.053	N	Y	0.411
si_90d	2	0.230	N	Y	0.728		absorpt	2	0.027	N	Y	0.067
ai_tot	2	0.315	N	Y	0.190		roq_c	2	0.084	N	Y	0.166
ai_12m	2	0.107	N	Y	0.246		roq_p	2	0.032	N	Y	0.058
ai_90d	2	0.156	N	Y	0.700		sss	2	0.064	N	Y	0.080
vis_min	2	0.084	N	Y	0.052		anx_st	2	0.022	N	Y	0.138
ceil_min	2	0.387	N	Y	0.137		anx_tr	2	0.050	N	Y	0.062
min_adj	3	0.007	NN	YN	0.228		durn_sec	2	0.122	N	Y	0.336
		0.999	NY	YY	0.999		pumpsavg	2	0.063	N	Y	0.044
fln_rt	3	0.030	NN	YN	0.999		pmpavglo	2	0.032	N	Y	0.039
		0.707	NY	YY	0.999		pmpavgme	2	0.095	N	Y	0.038
dthruwx	3	0.092	NN	YN	0.999		pmpavghi	2	0.095	N	Y	0.075
		0.684	NY	YY	0.825		pmpadjav	2	0.070	N	Y	0.044
pass_mot	4	0.169	N	Y	0.499		padjavlo	2	0.039	N	Y	0.029
fhincmot	5	0.256	N	Y	1.000		padjavme	2	0.132	N	Y	0.039
missnmot	6	0.477	N	Y	0.999		padjavhi	2	0.122	N	Y	0.083
mal_sub	6	0.945	N	Y	0.503		pay_tot	2	0.059	N	Y	0.043
sim_mot	6	0.251	N	Y	1.000		pay_low	2	0.042	N	Y	0.029
simmotsb	7	0.442	N	Y	0.168		pay_med	2	0.107	N	Y	0.042
buck_sig	6	0.031	N	Y	0.120		pay_hi	2	0.083	N	Y	0.071
crashsig	8	0.124	N	Y	0.387		bang_tot	2	0.080	N	Y	0.054
tx_mot	23	0.999	N	Y	0.999		bang_low	2	0.047	N	Y	0.052
badwx	5	0.840	NN	YN	0.414		bang_med	2	0.120	N	Y	0.082
		0.143	NY	YY	0.212		bang_hi	2	0.153	N	Y	0.077

High Incentive, N = 30. Reference category for *type_lic* is "Private." Reference category for *simmotsb* is "Positives and negatives cancel." The reason for most of the missing values here is that *buck_mot* had two itself, so each analysis therefore automatically had to reflect at least these two.