2005

General Aviation VFR-Into-IMC: Z-Score Filtering of Demographic and Personality Variables, and the Personality Paradox

William Knecht

GENERAL AVIATION VFR-INTO-IMC: *Z*-SCORE FILTERING OF DEMOGRAPHIC AND PERSONALITY VARIABLES, AND THE PERSONALITY PARADOX

William Knecht
Civil Aerospace Medical Institute, FAA
Oklahoma City, OK

Does pilot personality affect risk-taking with weather? Armchair logic says "Yes," while data often say "No." In this work, we apply the technique of *z*-score filtering (slice analysis) to pilot takeoff decisions made in the face of simulated adverse weather seen at taxiway level. Such a filtering technique might prove useful, provided emphasis is kept to maintain experiment-wise reliability. Statistical and methodological problems with personality data are discussed. The results of this particular data set showed a strong effect of weather on takeoffs, as measured by visibility, cloud ceiling, and the interaction of the two. But, despite best efforts, no strong effect of personality could be found in this data set. Theoretical reasons are discussed as to why it may be difficult to show that personality predicts behavior.

Introduction

Visual flight rules-flight into instrument meteorological conditions (VFR-into-IMC) is a serious problem in general aviation (Adams, Koonce, & Hwoschinsky, 2002; Hunter, 2002a,b; O'Hare, 1990; O'Hare & Owen, 2002; O'Hare, Chalmers, & Scuffham, 2003; Wiegmann, Goh, & O'Hare, 2002). The FAA has identified VFR-into-IMC as a leading cause of GA fatalities, and has made it a top priority in its 2004 and 2005 *Flight Plan* (FAA, 2004).

It is natural to wonder if pilots' personality influences their risk for venturing into severe weather. Armchair logic says "Of course it does." However, personality tests have a mixed record for being able to predict behavior. This has been called "The Personality Paradox"—the notion that, somehow, personality *must* exist and *must* affect behavior—yet the connection is usually hard to demonstrate.

In aviation psychology, at least one author asserts that virtually all personality research on pilots can be shown to have at least one fatal flaw (Besco, 1994). Besco cites a host of methodological errors, such as weak validation procedures, lack of replication, experimenter biases, "potential for fakery" of responses, and lack of objective performance criteria. Any of these flaws renders research results suspect.

There are also theoretical reasons why personality tests may not predict behavior. Within the field of personality research, a great "Person-Situation Debate" has raged for years. A good summary of this is given in Epstein & O'Brien (1985). To sum up briefly, every behavior is probably specific to some rather narrow environmental context, or *domain* (Weber, Blais, & Betz, 2002). For example, roads and skies are two different domains. A risky driver may not necessarily be a risky pilot. This means that domain-specific tests normed in a non-aviation domain may not have much application to aviation.

A central theoretical issue here is whether or not there even exist any such things as "domain-free personality traits." Such traits would have to be stable and exert an influence on behavior, no matter in what context that behavior took place.

In the present work, a number of common personality measures were examined, as well as two demographic factors commonly assumed to correlate with risk-taking behavior (pilot age and number of flight hours). The idea was to see if any of their scores, or sub-scores could predict takeoff into adverse weather.

Method

Thirty general aviation (GA) pilots were first given an extensive battery of common personality tests (Table 1). Pilots were next positioned on a taxiway in a flight simulator and were told that their aircraft was not currently certified for instrument flight, so any takeoff would have to be VFR. Three levels of simulated ground visibility (V = 1, 3, 5 statute miles) and two levels of cloud ceiling (C = 1000', 2000') were manipulated as independent variables in a 3x2 between-subjects design. Each pilot saw one V,C combination and then had to decide whether or not to take off and fly in that weather. Logistic regression modeling was then conducted to see if personality test scores could predict actual yes/no takeoff decisions.

Instrument	High score implies	Reference
Aviation Safety Attitude Scale	high history of aviation risk behavior	Hunter, 1995, 2002a, 2002b
Anxiety Sensitivity Index	high scores indicate high anxiety	Peterson & Reiss, 1994
Barratt Impusiveness Scale V10	high impulsivity	Barratt, 1975
Eysenck Impulsivity Scale	high impulsivity	Eysenck & Eysenck, 1964,1985
Hazardous Events Index	high history of aviation risk behavior	Hunter, 2002b
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
State-Trait Anxiety Scale	high anxiety	Spielberger, 1983

Table 1. *List of personality tests examined in this study.*

Results

Predictably, the single most significant groupwise factor in pilots' decisions turned out to be the weather itself. Seventy percent of pilots chose to stay on the ground. Contrast this with an expected rate of 100% takeoffs, had there been unlimited visibility and ceiling ($p << .0001$ by binomial expansion, assuming a highly conservative 28/30 takeoff ratio).

Throughout the regression analysis, despite extensive attempts to predict takeoff through seemingly sensible combinations of demographic and personality factors, no model ever seemed to explain much more outcome variance than did weather all by itself (about 50%).

Was this to say that pilot personality did not matter? Or was it was more likely that each pilot had a unique, individual set of motivations and propensities—a "story," if you will—but that there were so many individuals with so many different stories that it made groupwise analysis difficult?

To try to get at these individual stories, predictor scores were converted to *z*-scores, and then threshold-filtered to try to reveal patterns of predictors whose absolute values were high relative to the group mean.

This kind of slice analysis has potential as an analytical technique, particularly in cases where we wish to tell stories about a relatively small number of individuals. However, we do need to keep in mind the effect that looking at many predictors will have on experiment-wide (familywise) error (Keppel, 1982).

These potentials and issues are best seen through example. In this experiment, two pilots chose to take off into the very worst weather presented (1 mile ground visibility plus 1000' cloud ceiling). What, if anything, set these two pilots apart from the other 28?

Using slice analysis, an initial *z*-threshold value (θ_z) of 3.3 standard deviations was established. This theta value corresponded to the Bonferroni correction necessary to maintain familywise error at $\alpha = .05$ (two-tailed), despite the examination of 28 predictors for two subjects. The corrected α was derived from the desired familywise α divided by the number of examinations planned (.05 / (2*28) = .0009. The *z*-level necessary to achieve that new α was $z_{critical} = \pm 3.3$, which yielded an area of .0009 under both tails combined. This can be cross-checked by expanding the binomial $(\alpha, 1-\alpha)^n$ for n=56 factors and noting that $(1-.0009)^{56} = .95 = 1-\alpha$, which equals the chance of zero Type 1 errors (the chance of finding no statistical "significances" where none truly exist).

As Figure 1 illustrates, $\theta_z = 3.3$ was a very stringent criterion. All that remained after thresholding was a single surviving predictor for a single pilot (the first two variable slots merely represented visibility, and ceiling, which were not thresholded). This surviving predictor was the Hazardous Events Index (HEI) score, which measured pilots' past history of hazardous encounters. So it did make sense that an elevated HEI score could relate to risk-taking in this scenario.

At this point, it made some sense to try relaxing the familywise α to assess how this would trade off in terms of increased information. Relaxing to $\alpha = .10$ gave a $\theta_z = 3.1$. That still left 90% assurance that the overall analysis was reliable, which still translated to a best guess of zero expected overall Type I errors. This produced at least one extra piece of information about S 2031, as Figure 2 shows.

Unfortunately, the surviving predictor was a below-average Rohrmann Risk Orientation Questionnaire, *Risk Propensity* index score (ROQ-P). Having a *low* propensity for risk was inconsistent with this pilot's actual takeoff into the very worst conditions. So that left a logical quandary.

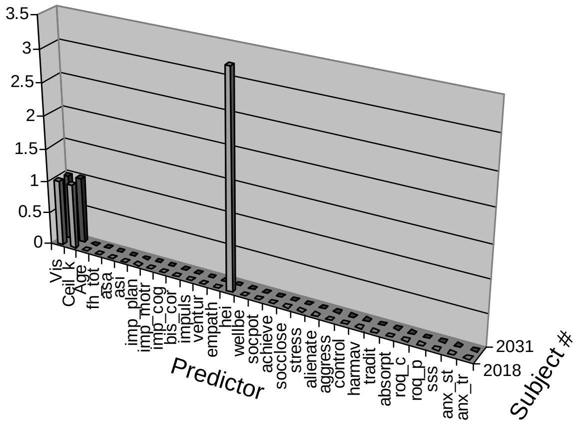

Figure 1. *Predictors thresholded at $\theta_z = 3.3$. This criterion was so rigorous that it failed to show anything other than an elevated Hazardous Events Index (HEI) score for subject 2018.*

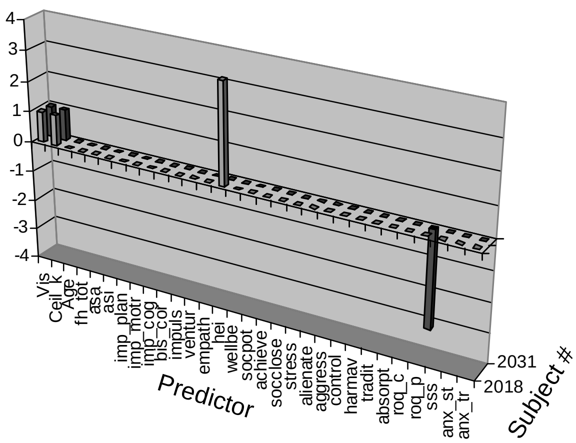

Figure 2. *$\theta_z = 3.11$. Familywise reliability is .90, expected Type I errors still = 0.*

To carry this filtering technique to its conclusion, θ_z was finally lowered all the way to 1.5. This provided only slightly more information, and led familywise reliability to plunge to .0003, with seven expected Type I errors, despite only four predictors surviving threshold.

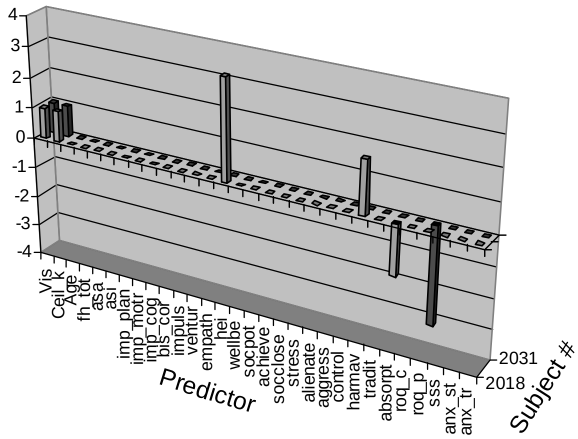

Figure 3. $\theta_z = 1.5$. *Little information is gained, despite a great loss in reliability.*

Discussion

The implication this methodology has for analysis is mixed. On the one hand, it makes it quite easy to imagine a "story" for each pilot—some pattern of predictor scores that might explain why that pilot acted as he or she did in some circumstance. On the other hand, elements of these stories may not be reliable or even make logical sense. In fact, as we can see with this data set, by the time we lower our reliability threshold (θ_z) to a level where we can see emerging patterns, our familywise error rate is in trouble. That means that, under certain circumstances, we could have gotten strong-looking—but counterfeit—patterns simply from random numbers.

So does this mean that pilot personality had nothing to do with pilot behavior? Not necessarily. What seemed more likely was that:

1. Aviation-specific versions of most of these predictors may be needed.
2. Even if the right predictors were tracked, scores may not have differed greatly enough from the mean to statistically distinguish themselves from noise. However, their concomitant traits might still have exerted influence on behavior.
3. Combinations of traits may have acted synergistically to create a "whole greater than the sum of the parts."

Point 1 concerns the notion that risk can be domain- and situation-specific. If so, then we would need aviation-specific personality tests, normed on pilots and specific aviation behaviors (e.g. the HEI).

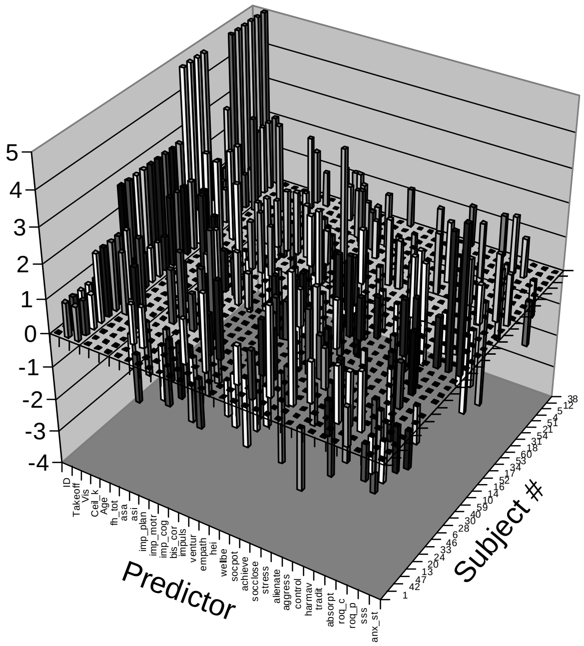

Figure 4. *θ_z = 1. Dense patterns of information emerge. These may have small amplitude, but might still exert true effect on takeoff.*

Figure 4 speaks to Point 2. As this illustrates, by greatly relaxing θ_z we can visualize how each subject may very well have a unique personality profile. But these patterns do seem almost all over the map. The problem is one of reliably demonstrating patterns when most of them lie "submerged" below a statistical threshold elevated by the number of factors being examined.

Points 2 and 3, if true, would make the study of pilot personality very difficult, if not impossible. We could call all this part of the mathematical basis for the Personality Paradox.

First, we have a theoretical situation loosely analogous to Heisenberg's Uncertainty Principle (the impossibility of simultaneously knowing a particle's momentum and position). The act of looking for meaningful patterns—examining many factors simultaneously—decreases the statistical reliability of each score to the point where the information becomes untrustworthy. We seemingly cannot have our cake and eat it too.

Second, and equally bad, if synergy does exist between variables, then the situation worsens because of a possible combinatorial explosion. Equation 1 shows the formula for *n* objects taken *k* at a time:

$$\binom{p}{k} = \frac{p!}{k!(p-k)!} \qquad (1)$$

So, if our personality test has, say, 11 factors, then there are 55 ways we could make pairs, 165 ways for triplets, 330 for quadruplets, and so forth. If, truly, "the action is in the interaction," then, given these kinds of numbers, we run headlong into impossibly strict criteria for limiting familywise error.

In short, we may be statistically caught between a rock and a hard place. The Personality Paradox may be an inevitable mathematical consequence of combinatorics.

Conclusions

It is difficult to dismiss the intuitive notion that "right stuff" personality plays a major role in pilot decision making. A logical next step in pursuing this issue might be to use a "Big Five" OCEAN approach. This would involve testing five commonly accepted factors of Trait Theory: openness to experience, conscientiousness, extroversion, agreeableness, and neuroticism. Popkins (2004) gives an excellent critical review of this approach. Since five is not a very large number, this would go a long way toward reducing combinatorial effects.

Yet we are faced with a burgeoning suspicion that it may be difficult to identify most of the personality factors that putatively affect behavior, even after the fact, let alone before it. There do appear to be so many individuals with so many different "stories" that mathematical arguments arise that indicate groupwise analysis and behavioral prediction on the basis of personality will always be difficult.

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.

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

Besco, R.O. (1994, January). Pilot personality testing and the emperor's new clothes. *Ergonomics in Design*, pp.24-29.

Epstein, S., & O'Brien, E.J. (1985). The person-situation debate in historical and current perspective. *Psychological Bulletin*, 98(3), 513-537.

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.

Federal Aviation Administration, (2004). *Flight Plan*. Retrieved December 29, 2004 from http://www.faa.gov/AboutFAA/Plans.cfm

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

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.

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

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

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

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.

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.

Popkins, N.C. (2004). *The Five-Factor Model: Emergence of a taxonomic model for personality psychology*. Retrieved January 18, 2005 from http://www.personalityresearch.org/papers/popkins.html. Unpublished manuscript, Northwestern University.

Rohrmann, B. (2002). *Risk attitude scales: Concepts and questionnaires. Project report.* Retrieved January 3, 2005 from http://www.rohrmannresearch.net/rohrmann-racreport.pdf. Unpublished manuscript, University of Melbourne, Australia.

Spielberger, C. (1983) State-Trait Anxiety Inventory (Form Y). Palo Alto, California: Mind Garden.

Timberlake, W. (2002). Niche-related learning in laboratory paradigms: the case of maze behavior in Norway rats. *Behavioural Brain Research*, 134, 355-374.

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, 263-290.

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.

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