Automated Trend Analysis for Navy-Carrier Landing Attempts

Neil C. Rowe
U.S. Naval Postgraduate School
Monterey, California, USA
ncrowe@nps.edu

ABSTRACT

A replacement system IPARTS is being built for the current U.S. Navy APARTS handheld data-entry device that records evaluations of landings of pilots on aircraft carriers. Navy aircraft are difficult to land and costly to repair, and extensive training and performance monitoring is important. Part of this task includes summarizing older data on landing attempts for comparison of pilot performances. We built tools for analyzing trends exhibited by pilots, pilot groups, aircraft, and evaluators in regard to grades, landing details, and verbal comments. Results are shown on a sample of 85,571 passes representing about 20% of the current Navy records, a significantly larger study than has ever been conducted. These results enabled building several kinds of predictive models of pilot performance which help identify particular pilot problems, and this should help in designing training programs. Fairness of grading of pilots was also assessed by comparisons between military units, aircraft, and graders. The most novel part of the research was understanding and computing statistics on the comments, which are in a telegraphic format using a unique language; a 2433-rule standardization routine and a parser were built to interpret them. Comments were essential in understanding the context of grades. The comment counts were also especially helpful in designing a user interface for a replacement grading device we designed and tested. This work should provide new insights into the performance of military pilots.

This paper appeared in the 2012 Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL, December 2012.

ABOUT THE AUTHOR

Neil C. Rowe is Professor of Computer Science at the U.S. Naval Postgraduate School where he has been since 1983. He has a Ph.D. in Computer Science from Stanford University (1983) and three other degrees from the Massachusetts Institute of Technology. His main research interests are the modeling of deception, information security, surveillance systems, image processing, and data mining. He is the author of 160 technical papers and a book.

http://faculty.nps.edu/ncrowe/rowe_itsec12_paper12247.htm

Automated Trend Analysis for Navy-Carrier Landing Attempts

Neil C. Rowe
U.S. Naval Postgraduate School
Monterey, California, USA
ncrowe@nps.edu

INTRODUCTION

Landing an aircraft is a difficult skill to acquire (Love, 1995), and is difficult to automate (Durand and Wasicko, 1967; Prickett and Parkes, 2001). Landing on an aircraft-carrier deck is especially challenging due to small size of the landing surface and the motion of the deck (Bennett and Schwirzke, 1992). Thus extensive training and practice is essential for carrier pilots. This is monitored by Landing Signal Officers ("LSO"s) (Figure 1) who watch every landing attempt, on carriers and at training bases, and who assign grades, data, and comments to landings. The current system APARTS (Automated Performance Assessment and Remedial Training System) (Bricston, 1981) uses a handheld device to enter this data. However, APARTS is old technology and needs replacement.

Figure 1: Aircraft and LSOs.

Our IPARTS (Improved Performance Readiness and Training System) project designed and built a replacement device and provided associated software. Tasking also included adding more modern data analysis capabilities and running them on both old ("legacy") and new data. This paper focuses on this analysis, and especially its most critical aspect, the evaluation of trends in pilot performance with the goal of improving training of pilots. (Salas, Milham, and Bowers, 2003) points out how military organizations are often overoptimistic about their training programs because of insufficiently careful evaluation, and military aviation tends to feature close-knit groups rarely subject to outside evaluation. Nonsubjective information about pilot performance such as physiological measurements also helps in the early stages of training (Schnell, Keller, and Poolman, 2008), but once pilots have experience, human judgment is necessary for assessment.

DESIGN

Information collected by the U.S. Navy for each landing attempt ("pass") by a pilot includes time, aircraft identification number, aircraft type, squadron or air wing, pilot name, type of landing attempt (training or operational), "recovery" (the name for the group of landing attempts), grader, grade, result of the attempt, and comments about it. Passes can be training runs on land, use of automatic landing systems, or even simulation runs. After discussions with several LSOs and examination of the few available written documents such as (U.S. Navy, 2001), the most important aspects of pilot performance were concluded to be: (1) average grade, grade during the day, grade during the night, and grade in the last 50 passes; (2) average boarding (landing) rate, boarding rate during the day, boarding rate at night, and boarding rate in the last 50 passes; and (3) counts of verbal comments from LSOs that were atypically common for each pilot. Their calculation was implemented in a Java program. This analysis used approximately 20% of the data for the last few years in the entire U.S. Navy. Obtaining this data was difficult as no centralized repository was available and we had to request it from each air unit separately.

Grades are integers 0-5 (5 is very good and 0 is very bad) with some passes ungraded due to special circumstances (such as a waveoff because of a "foul" (uncleared) deck or turning ship, or test passes). Boarding rates (the fraction of attempts in which the pilot landed) are computed on groups of passes, and rates of 0.9 were typical. Their computation following U.S. Navy policy is tricky because there are several rates computed with different numerators and denominators. For instance, the "combat boarding rate" must exclude passes where pilots where doing "touch and go"s, coming down for a landing but not actually landing, except when they did something dangerous near the touchdown.

Null values for the landing result occurred in some of the older data; the grade field was used to infer values when possible. Blank values, on the other hand, were interpreted as no-count failures to land. Determining the last 50 passes for each pilot used the recovery date, recovery time, and sequence number, since landing times were not generally recorded.

The LSO comments posed the greatest challenge for summarization. Most of these come from the LSO that watches the incoming flight path, with a smaller number from other LSOs. The comments follow a shorthand language with its own grammar (Table 1) intended to be quick to write. For example, "(LO)SLOIC-AR" means that an aircraft was slow and a little low, both when approaching the carrier and just at the edge of the carrier deck; the parentheses mean the lowness was only a little, and the hyphen means the comment applies to the period between being "in close" and "at the ramp". Other common locations are "X" (at the start), "IM" (in the middle), and "IW" (in the wires). Underscores are used for added emphasis, and periods are used to separate code letters that could otherwise be confused; special symbols represent ascending and descending relative to the ideal slope of the aircraft. Unhelpful characters like tabs, commas, semicolons, and double spaces are eliminated from comments before applying these rules, as well as duplicate comments added by different LSOs.

Table 1: Grammar for LSO comments.

Grammar rule	Semantic restrictions
e(code1 <space> code2) ? e(code1) <space> e(code2)	Append resulting lists of codes
e(code1-code2 loc) ? e(code1 loc) <space> e(code2 loc)	Loc is a location; hyphen means "to"
e((code)) ? e(code)	Mark deemphasis
e(_code_) ? e(code)	Mark emphasis
e([code]) ? e(code)	Mark missed communication
e(code times) ? e(code)	If times = x2, x3, etc., mark number of times
e(code1.code2 loc) ? e(code1 loc) <space> e(code2 loc)	Interpret dot as "on", if code1 has no loc
e(code1.code2) ? e(code1) <space> e(code2)	Interpret do as "on"
e(code1 code2) ? e(code1) <space> e(code2)	If code1 and code2 are known abbreviations

651 terminal symbols were defined for the grammar like "AR" ("at the ramp"), "LO" ('low"), and "ENG+PROBLEM" ("engine problem"), most of which are specified in (U.S. Navy, 2001). Other symbols are checked as misspellings. 37 misspellings, like "SBY" for "standby" rather than "STBY", were common enough to correct immediately; several hundred others are corrected after checking their context (neighboring words). In 248 cases new symbols were introduced for frequently seen phrases when no standard abbreviation was being used, like "A/C+IN+LA" for the many ways to say that a waveoff was due to an aircraft was blocking the deck.

Correction of misspellings and mispunctuations, standardization of terms, and substitution of codes for English words were done by 2433 transformation rules applied before parsing, which matched 33,155 instances in the test data. An example misspelling is "AAR" for the harder-to-type "AAAR". Example variant abbreviations are "EWI", "EZW/IT", and "EASYWITHIT" for "EWIT", and "PWR IM" for "PIM". An example terminology difference between air wings is "LUCKYBUCK", "BONGO", "GOLDSTAR", and "$$$$" for the free-pass authorization.

Transformations are tried in a fixed order, so their placement in the list was planned carefully. For instance, "CALLS IM AND IC" gets transformed first to "CALLS IM-IC", then "CALL IM-IC", then "CALL.IM-IC"; and "NEP/COIM LO IN CLOSE HALF FLAPS THRUST CAUTION" gets transformed ultimately to "NEP.COIM LOIC HALF+FLAPS ENG+PROBLEM". As an example of the rules, here are some for replacing expressions about three-point landings with the Navy standard "3pts".

```
three points -> 3-point
three-points -> 3-point
three point -> 3-point
three-point -> 3-point
three pts -> 3-point
three-pts -> 3-point
3 points -> 3-point
3-points -> 3-point
3pts -> 3-point
3pt -> 3-point
tchdn -> touchdown
landed -> land
landing -> land
lndg -> land
3-point touchdown -> 3pts
3-point land -> 3pts
touchdown 3-point -> 3pts
land 3-point -> 3pts
3-point -> 3pts
```

Overall counts are computed on each comment for each pilot, as well as counts on day passes, counts on night passes, counts on low-graded passes, and counts of comments that were atypically frequent for a pilot. Atypically frequent was defined as occurring more than K standard deviations above the norm of a Poisson distribution. In our experience, K=1 worked well to give useful observations. The Poisson distribution is a good model because in our data most comments occurred less than 0.1% of the time, and only 10 occurred more than 10% of the time. In computing counts, 50% less weight was given to deemphasized (parenthesized) comments, and 50% more weight to emphasized (underscored) comments, following discussions with LSOs. So "(LO)SLOIC" generated full weight for SLOIC ("slow in close") and half for LO ("low").

RESULTS

85,571 passes were analyzed in the legacy data. The average grade was 3.43 with a standard deviation of 0.66 on 77,833 passes qualifying for grading, and the average boarding rate was 0.939 with a standard deviation of 0.244 on 80,405 passes qualifying for notation of boarding. Total processing took about 0.002 seconds per pass on a five-year-old 32-bit Windows machine.

The lack of complete data meant that not all passes for the mentioned pilots were included in the data. Nonetheless, it was encouraging how much could be concluded from what was available.

Comment Frequencies

Overall counts on comments on height and power are plotted in Figures 2 and 3. Comments on aircraft attitude had a similar occurrence to those on power, and comments on aircraft lineup decreased uniformly with approach to the carrier.

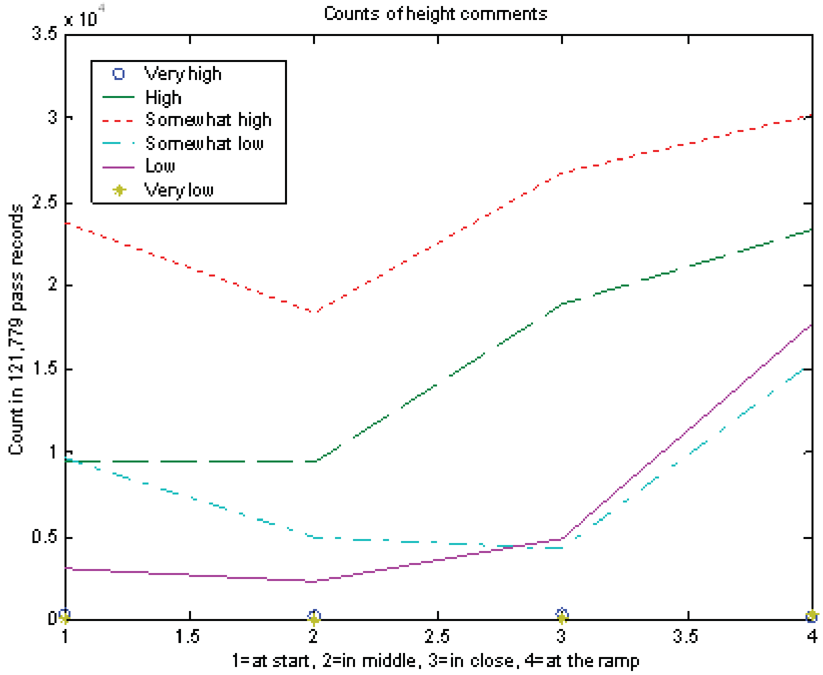

Figure 2: Counts on height comments.

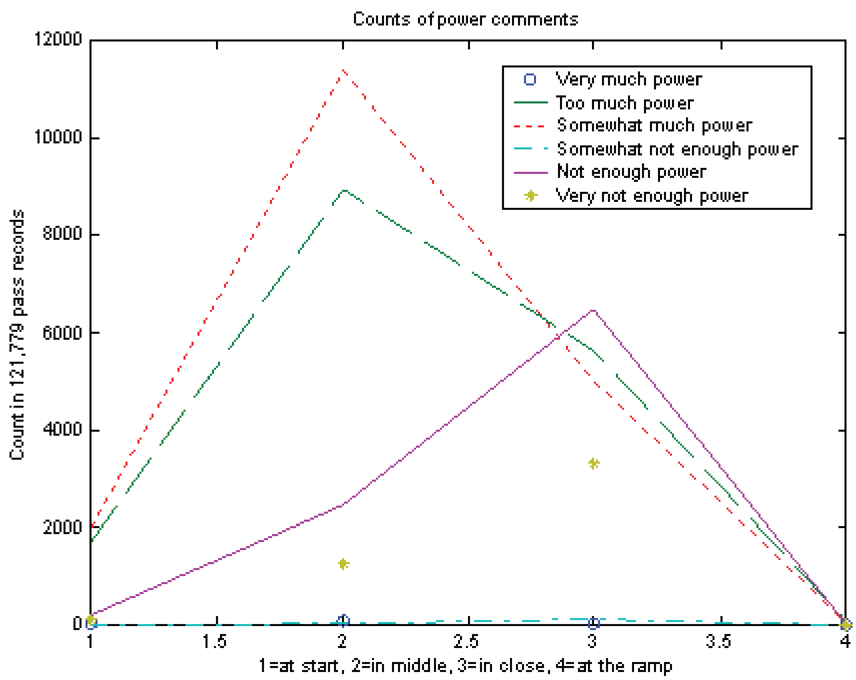

Figure 3: Counts on power comments.

The nonstandardized comments mostly refer to details of the landing, but there are also comments on calls, reasons for a foul deck, comments at earlier parts of the approach, and announcements of upgrades. The most common are listed in Table 2. The "luckybuck" free-pass upgrade was used in 2.3% of the passes. Other upgrades are given for specific circumstances, some apparently arbitrary. These occurred in only 0.2% of the passes, but should be checked to see if they are justifiable.

Table 2: The most common nonstandardized comments.

Problem	Count	Problem	Count
Wind	531	Ship in turn	406
Aircraft in landing area	337	No heads up display	209
No hook	188	Gear up	159
No angle of attack indicator	134	Deck not ready	82
Engine malfunction	71	No radio	53
People in landing area	52	Debris in landing area	46

Several codes listed in (U.S. Navy, 2001) were never used in comments and should probably be retired. Additional codes were observed that probably should standardized, like LTR ("left to right"), NELR ("not enough left rudder"), TMLR ("too much left rudder"), CLARA ("far from glideslope"), the 90 location, the 45 location, something for when an aircraft is in the landing area, something for no heads-up display, TOB ("talking on the ball"), something for gear up, something for no angle-of-attack indicator, NELSO ("not enough LSO"), TMLSO ("too much LSO"), and OT ("out of turn").

Other useful statistics concern which retarding wires the aircraft caught or missed on landing, as these may suggest trends in aircraft operations that require attention. Waveoffs for an unlandable deck occurred 6% of the time, pilot-judgment failures to land occurred another 6% of the time, deliberate practice of touching the ground without landing 4% of the time, and other kinds of waveoffs 1% of the time. All these are important in evaluating the effectiveness of carrier air operations.

Pilot Performance

Table 3 shows example pilot summary data that is prepared to aid the LSOs in debriefing the pilot. The numbers in brackets are the counts on the comments. This pilot had lower grades than the average, had a lower boarding rate than the average, was better at night, and tended to be too high coming in.

Table 3: Example pilot summary data.

Pilot number: 1225	Number of passes: 47
Average grade: 3.170	Average day grade: 3.089
Average night grade:3.312	Boarding rate: 0.893
Day boarding rate: 0.866	Night boarding rate: 0.941
Comments on start of descent: High[7] Too much power [7] A little high [5] A little overshot[4]	Comments on middle of descent: High [8] A little too much power [6] Too much power [6] A little high [4] A little ascending [4] Coming down [4] High coming down [4]
Comments near to carrier: High [9] A little high [7] Descending [4] A little high coming down [4] High coming down [4]	Comments just reaching carrier: A little high coming down [8] High coming down [7] High [5] A little high [5] Low and flat [5]
Nonlocalized comments: Long in groove [6] A little long in groove [6] Very long in groove [5]	Comments atypically frequent for this pilot: Long in groove [16] High at the start [13] High in the middle [13] High coming down at the ramp [11] Too much power at the start [8] High coming down in close [6] Overshoot settling [6] Coming back in the middle [6] Descending in close [5] High coming down in the middle [5] Coming down in the middle [4] Ascending at the start [4] Flat in the wires [4]

For debriefing of pilot, it is also helpful to plot comment severity versus location separately for height, power, and attitude. Below is an example for height where the pilot has a tendency to be high (H) even late in the pass. X is at the start, IM is in the middle, IC is in close, and AR is at the ramp.

```
Glideslope comments for 153 passes of pilot ******* (#844)
            X     IM    IC    AR
_H_         1     1     0     0
H           13    10    22    40
(H)         39    20    40    45
OK          89    114   82    32
(LO)        8     6     4     14
LO          3     2     5     22
_LO_        0     0     0     0
```

Plotting average grade versus average boarding rate for pilots was disappointing in enabling us to distinguish pilots. Other than a few pilots (15 out of 432 with 50 passes or more) who had particular problems with their boarding rates, and 6 especially talented pilots, the rest of the data formed a nice symmetric Gaussian cluster that was not very illuminating. But looking at the data a different way was more helpful. An important issue for pilot training, as indeed for any expensive one, is whether the amount of time allocated is sufficient. Figures 4 and 5 plot average grades and average boarding rate with the pass number in our data for the pilot. The graphs show averages of each group of 20 passes. The number of passes falling into each bin decreased very close to monotonically on a logarithmic scale from 11,132 for passes 0 through 19 to 51 for passes 400 through 420. Performance appears to improve continuously through 400 passes, but note that lower-scoring pilots are being removed as the number of passes increases, and this provides part of the effect.

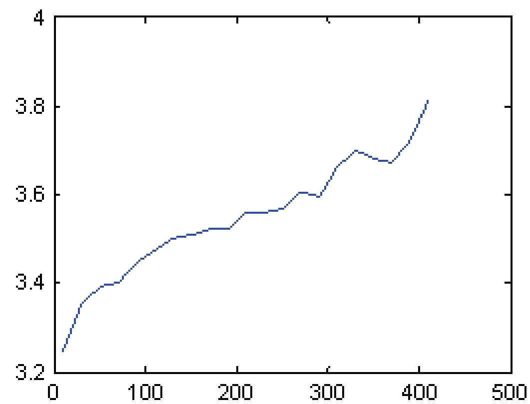

Figure 4: Pilot grade versus number of landing attempts ("passes").

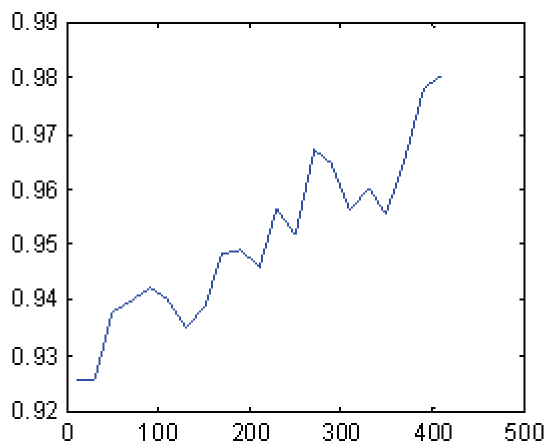

Figure 5: Average boarding rate versus number of passes.

These are classic examples of learning curves. Following the discussion of (Fogliatto and Anzanello, 2011), we concluded that the best model of the grade trend would be a hyperbolic curve of the form $K(x_i+p)/(x_i+p+r)$ where K represents the maximum for human performance, p is the previous experience of the pilot, and r is the inverse of the learning rate of the pilot. That is because experience of a pilot adds to factors in both the numerator and denominator of a performance measure since sometimes feedback is helpful and sometimes not. This formula has three parameters that can be fit using nonlinear least-squares methods. The components of the gradient for steepest-descent optimization of the fit error are:

$$\sum_{i=1}^{N} e_i (x_i + p) / (x_i + p + r),$$

$$\sum_{i=1}^{N} e_i K r / (x_i + p + r)^2,$$

$$\sum_{i=1}^{N} e_i K (x_i + p) / (x_i + p + r)^2$$

where $e_i = ((x_i + p) / (x_i + p + r)) - y_i$. This gradient was used to optimize over the passes from pilots with at least 50 passes, and it estimated overall values of K=3.68 (two thirds of the way from "somewhat OK" to "OK"), p=13.25, and r=4.25.

The three parameters can also be estimated separately for each pilot, using the fit for all pilots as the starting point for optimization. Figure 6 plots the learning rate parameter "r" (vertical) versus the experience effect parameter "p" (horizontal) for the 434 pilots with at least 50 passes. It can be seen that some pilots are clearly anomalous in their response to training and thus may be having problems. Figure 7 similarly plots "r" against the inferred final average grade "K" for the pilot, and indicates a different set of pilots with problems. Here the maximum grade of 5.0 was used as the upper limit on “K”. For best success with this method, however, complete data on pilots is needed since missing passes may represent more experience than the number of observed passes indicates.

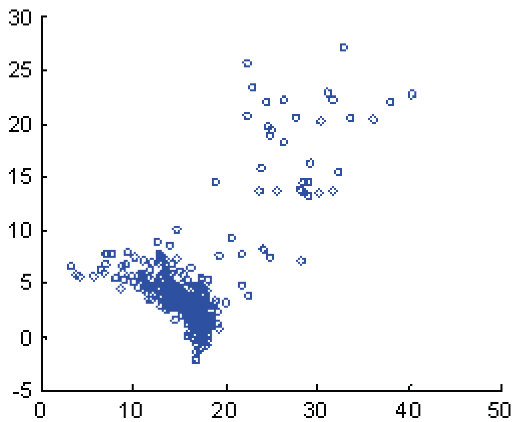

Figure 6: Learning rate parameter r (vertical) versus experience effect parameter p (horizontal) as fit to data for each pilot.

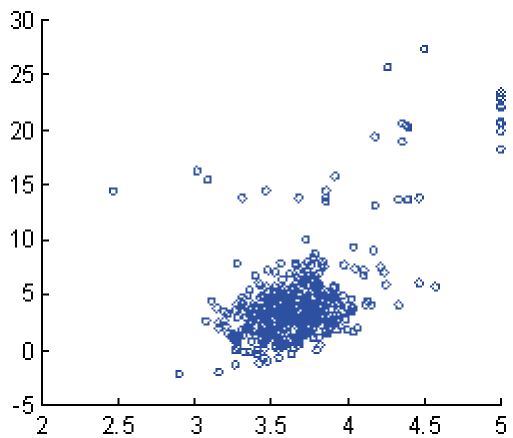

Figure 7: Learning rate parameter r (vertical) versus maximum inferred performance K for the pilot (horizontal).

Pilot Performance Versus Time Gap between Passes

Many researchers studying human training have noted effects of the time gap between successive training experiences (Schendel and Hagman, 1991; Ebbatson et al, 2012). So the average change in pilot grade was calculated as a function of the natural logarithm of the time gap in seconds recorded between successive passes (Figure 8), where the bottom curve is the average change in grade and the top curve is 0.1 times the natural logarithm of the number of passes having the same gap rounded to the nearest integer value (the right peak represents 20,076 passes). The results show a clear decline in performance with time gap as is typical with motor skills. It also suggests that time gaps of 37.8 days or more (value 15 on the horizontal axis) should be avoided as there is an average of at least 0.3 decrease in grade after such gaps.

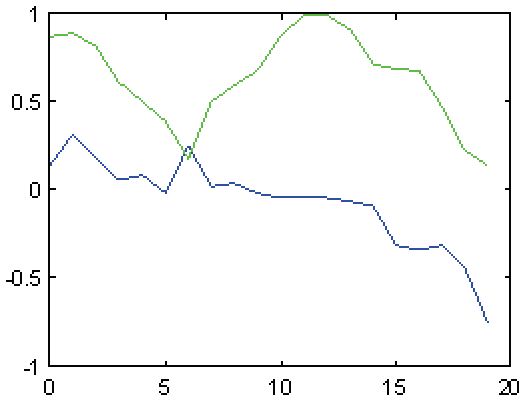

Figure 8: Average change in pilot grade between successive passes (bottom) and 0.1 times the logarithm of number of passes (top), versus logarithm of time gap in seconds (horizontal axis).

Predicting Future Pilot Performance

An issue important to the Navy is predicting of pilot performance from their early passes, since this can be used to more quickly decide which pilots are not going to qualify for retention and potentially save money. Figure 9 plots pilot grade average on the first 40 passes (horizontal) versus their final grade average. They are correlated with linear fit finalgrade = 1.39 + (0.622*initialgrade). But the dispersion is significant and especially for low grades. Thus it appears unfair to exclude pilots based on their grades on their early passes alone. The correlation between average grade on the first 40 passes and the total number of passes flown was slightly negative and unhelpful, as it appears that weaker pilots are allowed a few more passes for additional training.

LSO comments on a pass also could predict future pilot performance. To analyze this, for each atomic comment (after parsing and applying transformation rules), the average subsequent grade of the pilot and the number of subsequent passes that they flew were computed. To determine which comments had a statistically significant effect, these two measures were normalized with regard to the mean and standard deviation over the population according to sampling theory using $(s-3.43)*(\sqrt{N}/0.66)$ where s is either the average grade or average number of passes of a pilot having the comment, and N is the number of times the comment occurred. Only comments whose effect was more than one standard deviation away from the expected value on either the average pilot grade or the average number of passes were considered, to rule out weak correlations. Table 4 shows data for some comments that had significant effects. These clues and their strengths are consistent with LSO experience. For instance, being high before the 180 degree turn is a serious negative clue because it suggests careless flying.

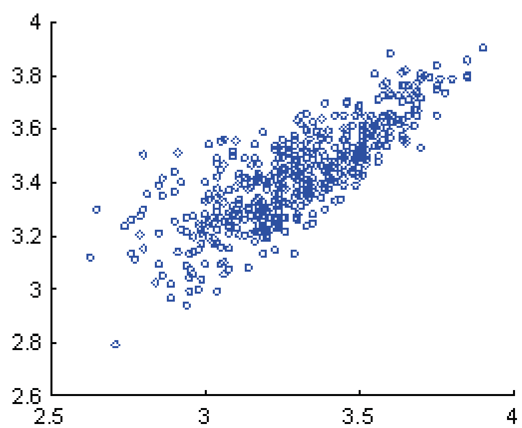

Figure 9: Pilot average grade on first 40 passes (horizontal) versus final average grade.

Table 4: Example comments having a statistically significant effect on grade or number of passes.

Comment	Count	Effect on average grade	Effect on number of passes
High	25480	-0.10	+43
Too much power	10688	-0.10	+44
Very high	490	-0.22	+60
Stopped rate of descent	703	-0.18	+51

High before 180 degree turn	9	-1.21	-64
Nose down	2836	-0.07	+41
Late	4	-1.72	-98
Chased centerline	28	-0.53	-7
Overcorrection on too much power	23	-0.62	+14
Much power in the wires	10	-0.45	-3
Not enough rudder	66	-0.15	+31
Tanker drill	6	+0.06	+77
Too much showing off	4	-0.67	-27
Showing off	163	+0.09	+0
Nose up a little	4217	+0.10	+11

An important question is how good a predictor the grade average on the first few passes is compared to the comments on the first few passes. An estimated grade was computed based on the comments on the first 40 passes by adding the associated effect numbers for each comment that occurred and was statistically significant, multiplying by a weighting constant, and adding to 3.43, the average grade over all pilots. Best fit was found with a weighting constant of 0.2. This estimate had an average absolute error of 0.143 in estimating the final grade average of the pilot versus an average absolute error of 0.146 for an estimate based on the average grade of the first 40 passes. Thus both are reasonable estimates. A natural next question is whether a weighted average of the two could be an even better estimate. Best results were obtained with a weighting of 0.45 on each of the estimates, plus a weight of 0.1 on 3.43 as a kind of enforced regression to the mean. This weighted average had an average absolute error of 0.108, a significant improvement. That shows that both a pilot's grades and comments are necessary to make a good prediction of how well they will do in their flying career.

Also interesting are the other extreme of comments that seem to have no effect on the average pilot grade, as these may be redundant and LSOs have much to write. 28 comments occurring at least 100 times had less than one standard deviation effect on both the overall pilot grade average and the total number of passes, so these would seem good candidates to eliminate. Examples are "a little slow" and "deck down a little", which are too mild to mean much for the pilot's future.

Squadron, LSO, and Aircraft Performance

Statistical summaries are also prepared by our software for each unit (squadron or air wing), each controlling LSO, and each type of aircraft. Other output files produced are a listing of comment counts for all pilots, a "night currency" summary of the latest night passes for each pilot, a list of pilot names found (to check different names for the same pilot), and a list of symbols in the comments that could not be interpreted and thus may require additions to the list of transformation rules or the list of code words.

Figure 10 plots the average grades of units against their average boarding rates, with area of the circle proportional to the number of passes for that unit (the largest circle represents 6950 passes). This includes 15 units and 3 more general categories. No particular correlation is obvious between grade and boarding rate, which suggests they are relatively independent. The data from the two squadrons at (3.2, 0.89) and (3.30.90) suggests some attention. Low boarding rates per se are not a concern because some units landed in more difficult conditions than others, and low grades may be due to having many new pilots, but having both low is cause for concern.

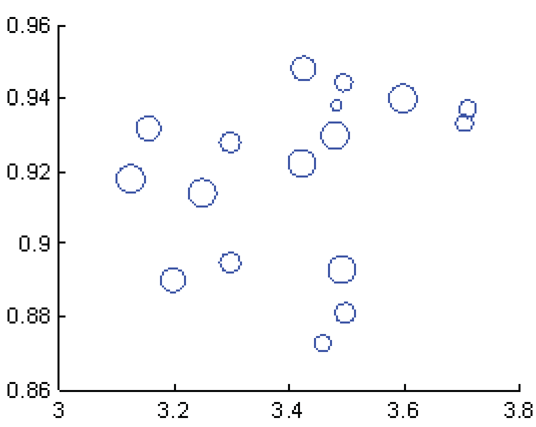

Figure 10: Average grade (horizontal) versus boarding rate for each squadron and air wing examined.

Figure 11 plots the normalized grade average (horizontal) for each of 325 controlling LSOs who judged 20 or more passes, plotted against the square root of the number of passes (vertical). The normalization was again the standard one for samples of a Gaussian population with mean 3.43 and standard deviation 0.66, or $(g-3.43)*(\sqrt{N}/0.66)$

where g is the average grade for the LSO and N is the number of passes they graded. Large positive values indicate LSOs that are too lenient in grading, and large negative values indicate LSOs that are too strict. The extreme values here, beyond three standard deviations from the mean, are well beyond chance. It appears important that the Navy take steps to ensure more uniformity of standards of grading.

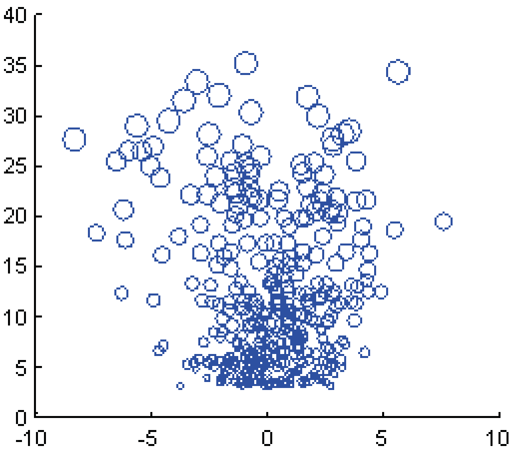

Figure 11: Normalized deviation from the overall average of average LSO scores (horizontal) versus square root of number of grades given by an LSO (vertical).

Relative performance of different aircraft was assessed by comparing the grades of pilots in those aircraft. The S-3B had the lowest average grade of 3.258 and the F/A-18E had the highest of 3.469. The differences between aircraft were not significant enough to indicate any trends. Nonetheless, these statistics should be calculated routinely to get advance warning of aircraft problems.

THE NEW IPARTS SYSTEM

This data analysis has been useful in designing a new handheld device to be used by LSOs in recording data as aircraft make landing attempts (Figure 12). A major challenge is providing the wide range of buttons for comment symbols, since minimizing text entry is highly desirable to reduce the many typing errors seen in the legacy data. So we designed three subscreens of the most common comment symbols. These covered all symbols in the 560,359 atomic symbol occurrences in our legacy data which occurred 200 times or more. They were grouped together intelligently within each subscreen to aid the user in finding them. LSOs seemed happy with the interface in tests.

Figure 12: Example screen view on the new (IPARTS) handheld device prototype.

We tested our device and interface in a Limited Operational Experiment both on land and on a carrier, and obtained 4563 additional pass records. Statistics on grades, boarding rates, and the most common comments were similar to those for the legacy data. However, the less-common comments were far fewer, from which we conclude that the interface did not support their entry very well.

Placement of buttons on the screens can be improved by formulating it as an optimization problem. After allocating required buttons, there was room for 29 options on the first subscreen, 40 on the second, and 40 on the third, to be chosen from 308 possibilities. (Possibilities not allocated buttons can be entered using a keyboard, but it is inconvenient.) A "greedy" ("hill-climbing") algorithm was implemented to test all interchanges of buttons between menus and between buttons and the stock of unused symbols. It successively chooses the best interchange until the placement could not be improved. To evaluate changes, it used statistics on successive atomic comment sequences in the legacy data, and gave a weight of 2 for consecutive comments on the same menu, a weight of 1 for consecutive comments on different menus, and a weight of 0 for consecutive comments involving interchanges with comments not currently on a menu. For instance, for the LSO pass comments of "HCDAR EWIT", atomic comments "H", "CD", “AR”, and "EWIT" were extracted; the number of times "H" was followed by "CD", the number of times "CD" was followed by “AR”, and the number of times “AR” was followed by "EWIT" in our legacy data were added, each multiplied by the appropriate weight based on their assigned menus. Using this, a locally optimal button placement was found in 34 steps of interchanges, and was calculated to save 306,002 units of effort on the legacy data compared to the original intuitively-designed layout. With 560,359 atomic comments total in the legacy data, these savings amount to one button press saved for every 1.8 atomic comments, so they are significant. While this placement does not necessarily group similar buttons together, it does optimize recording speed. This placement will be subject to future tests with LSOs.

Data recorded from the handheld devices during passes is then downloaded onto a repository laptop computer provided each ship and training base. The interface on the laptop provides statistical routines described in this paper so that users can run quick assessments of each pilot and unit. Repository data is periodically downloaded to a central database that provides access to all Navy data by a wireless connection.

CONCLUSIONS

Learning to land a military aircraft on a carrier is a difficult skill. Considering how expensive the aircraft are and how many hours of training are required for them, it is important to monitor pilot performance carefully. The methods developed here make it less subjective to recognize underperforming pilots, unfair grading of pilots, and unreliable aircraft, as well as the reasons for these problems through the statistics on comments. More complete data than 20% of the Navy is needed to make better predictions, though our study used significantly more data than any previous studies. This work should help improve the preparedness of naval aviators.

ACKNOWLEDGEMENTS

This work was sponsored by the Office of Naval Research under the Technology Solutions program. The views expressed are those of the author and do not represent those of the U.S. government. Michael G. Ross and Arijit Das helped obtain the data, and Mathias Kolsch, Michael McCauley, Thomas Batcha, Eric Lowney, and Greg Belli provided additional support.

REFERENCES

Bennett, C., & Schwirzke, M. (1992). Analysis of Accidents during Instrument Approaches. Aviation, Space, and Environmental Medicine, Vol. 63, pp. 253-261.

Bricston, C. (1981). Automated Performance Measurement for Naval Aviation: APARTS, A Landing Signal Officer Training Aid, pp. 445-448. Retrieved January 12, 2012 from www.faa.gov/ library/online_libraries/aerospace_medicine/sd/ media/brictson.pdf.

Durand, T., and Wasicko, R. (1967). Factors Influencing Glide Path Control in Carrier Landing, *Journal of Aircraft*, Vol. 4, No. 2, pp. 146-158.

Ebbatson, M., Harris, D., Huddleston, J., and Sears, R. (2012). Manual Flying Skill Decay. In De Voogt, A., and D’Olivera, T., (eds.), *Mechanisms in the Chain of Safety: Research and Operational Experiences in Aviation Psychology*, Farham, UK: Ashgate, pp. 67-80.

Fogliatto, M., and Anzanello, M. (2011). Learning Curves: The State of the Art and Research Directions. In Jaben, M. (ed.), *Learning Curves: Theory, Models, and Applications*, Boca Raton, FL: CRC Press, pp. 3-21.

Love, M. C. (1995). *Better Takeoffs and Landings*. Columbus, OH: Tab Books / McGraw-Hill.

Prickett, A., and Parkes, C. (2001). Flight Testing of the F/A-18E/F Automatic Carrier Landing System. Proc. IEEE Aerospace Conference, pp. 5:2593-2612.

Salas, E., Milham, L., and Bowers, C. (2003). Training Evaluation in the Military: Misconceptions, Opportunities, and Challenges, *Military Psychology*, Vol. 15, No. 1, pp. 3-16.

Schendel, J., and Hagman, J. (1991). Long-Term Retention of Motor Skills. In *Training for Performance: Principles of Applied Human Learning*, Chichester, UK: Wiley, pp. 53-92.

Schnell, T., Keller, M., and Poolman, P., 2008. Quality of Training Effectiveness Assessment (QTEA): A Neurophysiologically Based Method to Enhance Flight Training. Proc. 27th Digital Avionics Systems Conference, October, pp. 4.D.6-1-4.D.6-13.

U.S. Navy (2001). *NATOPS Landing Signal Officer Manual*, NAVAIR Publication 00-80T-104.