Final Project Report for FAA Grant 00-G-014

Training and Assessment of Aircrew Skills:

lethods to Achieve Reliable and Valid Performance Data

Timothy E Goldsmith

University of New Mexico

Project Summary

The project aimed to develop and validate procedures for training and assessing the knowledge and skills of airline pilots. The focus was on training and calibrating evaluators of aircrews and on the traditionally difficult to train and assess area of crew resource management skills. A central aim of the project was to aid airlines and other entities in collecting and using pilot performance data to make valid training and assessment decisions. The project employed several methods to achieve its goals including statistical modeling of flight parameter data, analysis of existing pilot performance databases, analysis of incident and accident reports, and controlled laboratory studies. Some of the major developments and findings of the project included (a) a software package to train and calibrate evaluators' judgments of aircrew performance, (b) statistical models of evaluators' performance judgments based on real-time, physical flight parameters, (c) evidence that crew resource management skills are highly context dependent, (d) quantification of the decline in aircrew skills over a 12 month retention interval, and (e) evidence that degree of expectancy affects performance on unlikely to occur emergency maneuvers.

Training and Calibrating Instructor/Evaluators

A major goal of the *Advanced Qualification Program* (AQP) is to ensure that carriers have a *quality assurance* program which ensures that aircrew members have the highest possible level of proficiency on all technical and management skills relevant to the safe and efficient operation of the aircraft. The implementation of a quality assurance program carries with it certain requirements, beginning with a database that defines an explicit set of qualification standards that are based on job task listings. These qualification standards drive the content of the curriculum, which in turn drive an assessment process that explicitly evaluates pilots on these performance standards. These assessment data provide feedback regarding the content and delivery of the curriculum. This feedback in turn allows for continuous improvements in curriculum design, as well as improving the allocation of training efforts to those knowledge and skills that are weakest. When functioning properly this system will ensure that all aircrew members attain and maintain a pre-specified standard of proficiency. Thus, it can be seen that quality assurance requires *quality assessment*.

A quality assurance program can only be as good as its weakest link. The qualification standards must be based on a careful analysis of job task listings. The curriculum and instruction must be designed to train to the qualification standards. And finally, the focus of this research proposal, the assessment tools must provide a *reliable* and *valid* evaluation of aircrew performance. It is of the utmost importance to realize that under AQP we are not simply assessing individuals, we are assessing the viability of the curriculum, the instructors, and the evaluators. From this perspective, the primary function of assessment is to improve training and thereby provide highly qualified aircrews.

Instructor/Evaluator Training and Calibration Tool (IETC). During the grant period continued to develop and refine the instructor/evaluator training and calibration (IETC) software tools. These tools are designed to facilitate the training and calibration of I/E's in their evaluations of crew performance during flight simulations. The work is currently directed toward evaluations of crews performing LOEs, but it has natural extensions to other types of evaluation including technical maneuvers. Indeed, the LOE appears to us to be the most difficult type of flying performance to grade, and so if we can succeed in this arena we feel confident that we can also help train and calibrate evaluators on other grading tasks.

There are two basic types of training and calibrating sessions, one involving a group of I/Es who come together at one time and the other involving a single I/E. Group sessions occur when several I/Es view a single video of a crew flying various scenarios from an LOE (either actual or staged) and then grade the crew using the gradesheet that they would normally use for grading an LOE. Various types of feedback are then given to the I/Es to show them how their grades compare to grades given by others in the group and also to the grades that they should have given, which have been defined by the qualification standards and grading scale criteria. The IETC software tool aids this process by creating an electronic version of the gradesheet, providing a tool for easily entering grades on this electronic gradesheet for multiple I/Es, providing tables in an Access database to house the grade data, and generating various types of statistical reports to be given to the I/Es as feedback.

An individual session using IETC consists of a single I/E performing a similar exercise in training and calibrating but now everything in the session occurs through a personal computer. The I/E selects a training session that would correspond to a particular video and gradesheet (or perhaps his supervisor would select one for him). An electronic gradesheet would then appear on the screen with the requisite tools for him to enter and change grades and move through the items on the gradesheet. Controls are provided for him to play the video on the computer screen, and then once grading is complete, he is provided with immediate feedback on his grading. In addition to the standard feedback given during a group session, the I/E user of the individual software is allowed to review segments of the video where his grading might have differed from the desired grade, review an online version of the qualification standards, again with the aim of understanding why a particular task item should have been graded in a certain way, and also to review an online description of the grading scale criteria. The IETC software tool provides tables to house I/Es grading performance over time.

During the grant period we modified the IETC software with the goal of creating a version that would be stable and usable by the airline community. Much of this work involved standard improvements to the code such as generalizing data structures and variables, ensuring the code would work on various platforms and configurations, and changing from the existing avi video files to the more efficient mpeg format. A more substantial change was to create software that would allow a user at a carrier (say an I/E supervisor or computer systems person) to easily create an electronic gradesheet and tie it into the appropriate tables in the Access database. This was accomplished through MS Word Forms and Visual Basic code. Version 1.0 of the group IETC tool has been completed and sent to a large group of US carriers. In addition the software and user manual can be downloaded from http://faa.unm.edu. The IETC User's Manual is included as a pdf attachment.

As mentioned above, the IETC tool is currently aimed at training and calibrating the assessment of an LOE. An obvious extension of the tool is to use it to train and calibrate maneuver validation grading. An important difference between an LOE and a maneuver validation is the importance of having available much more detailed technical information for grading a specific maneuver. The qualification standards of maneuvers are written with highly specific statements about physical characteristics of the aircraft (e.g., +/- 10 deg heading). Currently this level of information would not be available with a video camera in the simulator. However, it is possible to download a set of physical flight parameters from a simulator and then replay the flight on a desktop computer. Sets of flight parameters could be obtained for various levels of performance on critical maneuvers and then replayed for training and calibrating I/Es to grade the maneuvers. We propose to investigate integrating the animation of a flight with the IETC tool in order to extend the IETC tool to maneuver validation.

Evaluation of Calibration Software. The goal of the IETC software is to improve I/Es' LOE grading skills, which in turn will improve the quality of the performance data that is entered into the PPDB. Clearly, AQP can only be as good as the tools and the data that are used to assess proficiency. With the development of this software the final and most important step is to assess the effectiveness of IETC. To our knowledge there is little information regarding the permanence or the transfer of calibration training to new event sets and LOEs. Obviously, there is little point in calibrating I/Es if the beneficial effects are short lived and fail to transfer to event sets other than the ones trained.

In our view there are essentially two sources of data by which we can evaluate the effects of calibration training on I/E performance. One is to track I/Es' performance across repeated calibration sessions; the other is to observe the effects of calibration on actual LOE grading performance. There are certain obvious problems in attempting to systematically track changes in calibration data. Several factors can be expected to change across time (e.g., the population of I/Es, the population and/or sample of pilots, the contents of the calibration video, or the possibility of I/Es remembering the video if the same one is used, additional training experience, etc.). With a sufficiently large sample of I/Es we may be able to safely assume that the effect of some of these potential confounds would be randomized out. However, the sample size of I/Es for a given fleet within a carrier, for which we have calibration data, is usually not very large. Regardless of the various problems, these kinds of data need to be systematically collected and analyzed. Ideally, we would document the effects of both group and individual calibration on the statistics that are computed with the IETC software (RRR, IRR, and MAD).

Turning to the analysis of LOE grading data there is somewhat of a good-news-bad-news story. On the positive side we would be looking at the effects of calibration on real LOE grading. Also, even with a relatively small fleet the sample size would be expected to accumulate quite rapidly over time. The problem with this measure is that we cannot estimate RRR or IRR because the I/Es are not grading the same performance. Thus, all that remains is to examine the distribution of LOE grades across I/Es. This information will allow us to identify those instructors who are grading too high or low, or those for whom there is virtually no variance in the grades they assign. Certainly, if calibration is having a positive influence on I/E grading we should expect to observe a measurable change in these variables. One might assume then, that if the calibration training produced desirable changes in the distributional characteristics of an I/E's grades, it would also improve his ability to discriminate different levels of performance and assign to them the appropriate grade.

Instructor/Evaluator Reliability Check (IERC). In addition to the IETC software we have created a simple software package that airlines can use to analyze their instructor/evaluator calibration data. The software performs statistical analyses on data stored in an Excel spreadsheet and outputs the results (tables and charts) on a separate Excel worksheet. The purpose of the software was to provide a very simply format for analyzing training and calibration data that an airline could use for existing data. The software package is available for download at http://faa.unm.edu. The user's guide is provided as a separate pdf file.

Performance Differences on Rejected Takeoffs as a Function of Expectancy

Unexpected emergency situations in the aviation realm (e.g., rejected takeoffs) demand an immediate response from the pilot in order to avoid severe consequences. Commercial pilots receive extensive training on emergency maneuvers in simulators; however, on rare occasions pilots experience an unexpected event and perform poorly. We completed a series of studies aimed at investigating the effects of expectancy on performance for rejected takeoffs. We found that undergraduate students had a significant degradation in performance for unexpected rejected takeoffs. These results have implications for pilots who experience unexpected events on the line.

It is generally assumed that flying skills trained and assessed in the simulator will transfer to the line, and in general there is good reason to assume this. However, there is a class of maneuvers for which such transfer can be questioned and for which there is little or no empirical data to support a transfer assumption. These are maneuvers that demand an immediate response to unexpected events (e.g., rejected takeoff; RTO). The assessment of these types of abnormal events occurs exclusively in the simulator where a crew has a high expectancy of encountering an emergency maneuver. This stands in stark contrast to the line where a pilot may never encounter a RTO. Hence, we end up assuming that performance on events that occur under conditions of high expectancy will transfer to conditions where the expectation of these events is extremely low. We know of no empirical support for this assumption.

Unexpected events in the aviation realm (e.g., RTOs) require an immediate response from the pilot in order to avoid serious consequences. Although commercial pilots receive extensive training and evaluation on emergency maneuvers in simulators, they may never experience an emergency maneuver during actual flying. Hence, their expectation of an emergency maneuver is very low. The question, then, is if they do experience an emergency situation, how well do they deal with it?

The question is motivated by an extensive body of research that shows that people are generally shower to respond to an unexpected event that to an expected event. For example, Kirby (1976) looked at expectancy by presenting his participants with two different stimuli (consisting of two lights). The participants were informed to respond to the presented light by pressing a corresponding key. Kirby presented his participants with a long series of the same light and then on the test trial presented them with either the same light again or the alternate light and noted their reaction time. He found that his participants were much slower to respond to the different light than they were to respond to the same light.

Posner and his colleagues (1978/1980) found that participants were slower to detect a stimulus that occurred in an unexpected position. Participants were asked to respond to a stimulus on a screen by pressing a button. A cue indicating which part of the screen the stimulus would appear was presented before the trial began. Posner and his colleagues found that the participants were slower to respond to the stimulus when the cue was incorrect (i.e., when the stimulus did not appear where the cue indicated it would). That is, when the stimulus occurred in an unexpected position, the participants were slower to respond.

Mattes et al. (1997) found that participants were slower to react to stimuli that had a low probability of appearing. In their experiment, the researchers told the participants the probability of the stimulus appearing at the beginning of each trail. The participants' task was to respond to the stimulus (if it appeared) as quickly as possible by pressing a key. The researchers found that when the participants were told that the stimulus had a low probability of appearing, the participants were slower to respond when it actually did appear. Finally, Wolfe et al (2005) found that participants failed to detect key targets in a visual task when the targets appeared infrequently but not otherwise. Thus, people's performance degrades for unexpected events.

The expectancy effect for pilots can be thought of as a decrement in performance due to low expectancy of an event. Perhaps one reason why this area of research is so crucial to study is because the consequences of an error are very severe; there have been instances in which lives have been lost and airplanes have been damaged as a result of pilots not responding properly to an unexpected event.

The paucity of empirical data on this potentially serious safety problem is related to the difficulty of manipulating pilot's expectancy for abnormal events in the simulator and the rare occurrence of aborted takeoffs on the line. Given these real-world constraints, we have initiated studies aimed at investigating the effects of expectancy on performance for unexpected events in a laboratory situation with undergraduate students.

In a series of studies, we investigated the effects of expectancy on performance for an unexpected event (i.e., RTO) in a simulated flying task with undergraduate college students. One specific type of RTO occurs when an engine fails prior to rotation speed. The pilot must abort the takeoff and quickly stop the plane on the runway. This is the type of RTO that we studied in our study. Participants were first trained to perform both normal and rejected takeoffs on a PC-based flight simulator. We then tested their performance on a RTO during a subsequent test session where we provided a long series of normal takeoffs (NTOs) followed by two RTOs.

We manipulated expectancy in two ways: by the relative frequency of a RTO in the test session and by the instructions that the participants received in the test session. All of our participants received a long series of NTOs followed by two RTOs in the test session. We expected our participants to have a relatively low expectancy for the two RTOs after the long series of NTOs. In addition, some of our participants received explicit warnings in their test session instructions before the RTO trials, while others received no warning or a misleading warning in their instructions. We expected performance for the latter group of participants to degrade compared to those who received a warning before the RTO trial. The participants who received an explicit warning about an impending RTO were expected to have a higher expectancy for the RTO than those who were not warned or were mislead about the impending RTO trial.

Method. Two hundred and sixty four undergraduate students from the University of New Mexico served as participants. The students received course credit for their participation. Over six percent of the students reported having some experience with Flight Simulator, taking flying lessons or being a pilot.

Microsoft's Flight Simulator 2004 displayed on a personal computer was used to create a well controlled laboratory task to study performance on RTOs. The aircraft was a Boeing 737. The task achieved at least some level of realism. Flight Simulator was supplemented with (a) Adventure Basic Language Software Development Kit (ABL) to control and monitor the flights, (b) Flight Recorder for Microsoft Flight Simulator (FLTREC) to record the flight parameter data, and (c) Perl to run and coordinate the various simulation components. Participants used a yoke (CH Products USB LE Flight Simulator Yoke FSY208LE 200-608) and rudder pedals (CH Products PRO Pedals USB PP99USB 300-111) to control the aircraft. Instructions were played via Windows Media Player and were recorded (via Audacity) in a male voice. There were four different sets of instructions; each set described different task instructions for various phases of the experiment.

The participants were tested individually in a small room. The first set of instructions was displayed (via Media Player) informing the participant about the study and asking him to sign an informed consent. After signing the informed consent, the second set of instructions described the training session. These instructions described and explained the controls of the aircraft in flight simulator (i.e., the yoke, throttle, rudder pedals, air speed indicator, attitude indicator and altimeter). The experimenter pointed out the controls as they were being mentioned in the instructions. Next, the participant was informed that he would perform a series of NTOs. A NTO consisted of taking off and then turning the aircraft.

Each trial started with the plane lined up on the runway, engines running and ready for takeoff. The altitude of the runway was 440 feet and the heading was 160 degrees. Once the participant released the break and engaged the throttle, a text message displayed the takeoff speed (either 145 or 155 knots; randomly determined), the turn altitude (always 1000 feet) and the heading (140 degrees [left turn] or 180 degrees [right turn]; randomly determined). The participant was asked to stay as close to the centerline as possible when traveling down the runway. Once the participant took off, he received a message indicating whether or not he took off at the correct speed. After the specified heading was reached, the program stopped and reset itself for another trial. The participant performed 10 NTOs.

Once the participant completed the NTOs, a third set of instructions described how to perform a RTO. A RTO trial began the same as a NTO but before the plane reached the designated rotation speed, one of the engines failed. The participant was informed to close down the throttle as quickly as possible and then to steer and brake the plane to keep it as close to the centerline as possible. Once the plane came to a complete stop, the program stopped and reset itself for another trial. The participant performed 15 RTOs. On average, the training session took forty five minutes to complete. After the participant completed these training trials, he was told he could take a break if he desired (most people declined).

After the training session was completed, the second test session began. The participants were divided into five groups which all differed in their instructions for this second session.

One way in which we manipulated expectancy was by the instructions that our participants received during the test session. There were five different groups that all differed in their test session instructions. Group A was instructed at the beginning of the session to "focus on the initial training you received in session one." These instructions were intended to be vague and not informative of whether or not the participant would expect to receive any RTOs in this session. Thus, we expected Group A to have the lowest expectancy for a RTO and consequently to have the worst test RTO performance (compared to baseline RTO performance) of all groups.

Group B was informed at the beginning of the session that "in session two you will perform another set of takeoffs but you will not know when a RTO might occur". This set of instructions was still rather vague about expecting a RTO, but the participants in this group were told that a RTO would occur at some point in the test session, unlike the participants in Group A. Thus, we expected Group B's performance to be better (i.e., Group B would not show as large of an expectancy effect) than Group A's performance. Group C was given the same instructions as Group B but the message "this flight will be a rejected takeoff – the engine will fail on the runway sometime before V1" was also displayed onscreen before each RTO (but not the NTOs). Since Group C was given an explicit warning before the RTO trial that a RTO would occur, they had a high expectancy for the RTO to occur. Thus, we did not expect to find an expectancy effect; we expected Group C to perform the best out of all of the groups.

Group D was provided the same instructions as Group B but was also told to "be prepared for a RTO" before every trial (i.e., these participants received the warning regardless of whether the subsequent trial consisted of a RTO or not). The reasoning behind these instructions was to see if the warning loses its effectiveness over trials. Since the participants received a long series of NTOs before they received the RTOs, it was expected that the warning would lose effectiveness by the time the RTO trials occurred. Thus, we expected Group D's RTO performance to degrade. Finally, Group E was given the same instructions as Group B but was also given various warnings suggesting the possibility of an impending RTO before each trial. The warnings for Group E varied in semantic content but all involved situations that might potentially result in a RTO (e.g., birds on the runway, fuel contamination, extreme outside temperatures). These warnings were presented regardless of whether the subsequent trial was a RTO or not. The rationale behind this set of warnings was to assess whether the misleading warnings might maintain their effectiveness if they involved varied semantic content. We hypothesized that the warnings would not lose their effect and thus the participants' performance would not degrade.

All of the participants performed 15 trials in the test session; 13 NTOs followed by 2 RTOs.

Results. Several measures of participants' performance on the RTOs were obtained (e.g., stopping time, stopping distance), but response time to shut down the throttle after engine failure and maximum distance off center line were used for statistical analysis. Extreme throttle and off-center values were scaled such that throttle values above 6 seconds were scaled to 6 seconds and off-center values over 65 feet were scaled to 65 feet. The last 5 RTOs during training were averaged for each participant to establish baseline performance measures. These baseline measures were compared to the participant's performance on the first RTO of the test session. The participants' performance for these measures is shown in Tables 1 and 2.

A paired t-test for Group A revealed a significant throttle effect (t(61) = 4.44, p<.001) and a significant off-center effect (t(61) = 3.05, p<.001) indicating that participants were slower to shut down the throttle and deviated more from center line during the unexpected RTO in the test session than their baseline performance. Analysis for Group B revealed a significant throttle effect (t(42) = 4.19, p<.001) but off-center was not significant (t(42) = 0.16, p>.05). Analysis for Group C revealed an insignificant throttle (t(43) = 0.92, p>.05) but a significant maximum off-center effect (t(43) = -3.13, p<.001). Interestingly, performance for Group C revealed a significant effect in the opposite direction (i.e., the participants actually performed significantly better on their test trial than on their baseline trials). Analysis for Group D revealed a significant throttle effect (t(54) = 3.26, p<.002) but off-center was not significant (t(54) = 0.87, p>.05). Analysis for Group E revealed a significant throttle effect (t(63) = 3.97, p<.001), but off-center was not significant (t(63) = -.604, p>.05).

In order to assess whether the participants' performance on the test RTO was a result of skill decay (i.e., loss of trained skills after a period of nonuse) rather than expectancy, we compared the participants' second test RTO performance with their performance on the first RTO and baseline throttle response. The differences between the first and second test RTOs show a surprising pattern of results (see **Table 3**).

After receiving the first RTO in the test session, the performance of Group A on the subsequent RTO was not different from their baseline RTO performance (t(61) = -0.28, p>.05) but was faster than the first RTO (t(61) = 5.08, p<.001). The degradation in performance on the first test RTO was seemingly due to the expectancy effect and not due to skill decay.

However, Groups B, C, D, and E did not show any overall significant differences between throttle response on the first and second test RTOs, although it would not be expected for Group C, where the mean difference between baseline, test, and post-test did not exceed 0.087 seconds. The other groups were split, with approximately half the participants showing an improvement in throttle response for the second RTO, but the other half not showing any change.

Table 1

Mean response time to shut down throttle after engine failure in seconds

Group	Baseline	Test **RTO**	Cohen's d	P
	Mean (std)			
A	**1.41 (0.75)**	2.26 (1.26)	0.54	**<.001**
B	**1.07 (0.47)**	1.62 (0.75)	0.68	**<.001**
C	**1.09 (0.47)**	1.11 (0.47)	0.06	**ns**
D	***1.13 (0.41)***	*1.46 (0.85)*	*0.60*	**<.002**
E	**0.90 (0.38)**	1.34 (0.91)	0.57	**<.001**

Table 2

Maximum off center distance in feet

Group	Baseline	Test RTO	Cohen’s d	P
	Mean (std)			
A	28.98 (12.59)	37.00 (20.54)	0.46	<.001
B	26.07 (13.25)	25.59 (18.02)	0.03	ns
C	24.47 (11.35)	19.47 (14.72)	0.88	<.001
D	29.55 (14.02)	31.93 (22.17)	0.16	ns
E	21.80 (11.17)	20.75 (17.02)	0.12	ns

Table 3

Mean difference in throttle response between the first and second test RTOs and number of participants showing improved versus degraded performance

Group	Improved		Degraded	
	Mean (std)	N	Mean (std)	N
	1.21 (1.2)	48	-0.50 (0.31)	14
B	0.64 (0.66)	24	-0.42 (0.63)	18
C	0.31 (0.34)	24	-0.55 (0.53)	19
D	0.43 (0.38)	29	-0.75 (1.04)	26
E	0.69 (0.91)	30	-0.55 (0.73)	34

Thus, all groups except for Group C exhibited a significant degradation in their throttle performance (i.e., they were slower to shut down the throttle for the unexpected RTO in the test session). Only Groups A and C exhibited a significant difference in their off-center performance; Group A showed a significant degradation in their test session performance (i.e., they deviated more from centerline on the test RTO trial compared to their baseline performance). Group C showed a significant improvement in their performance (i.e., they deviated less from centerline on their test RTO trial compared to their baseline performance).

Discussion. The results indicated that performance can significantly degrade as a function of expectancy. We hypothesized that Group A would have the worst overall test performance of the groups because these participants had the lowest expectancy of a RTO in the test session (due to the nature of the instructions they received). We found that both the test throttle and off-center test performance was significantly worse than baseline performance which is consistent with our prediction. In fact, Group A was the only group that had degradation in their off-center test RTO performance. Group B was expected to perform better than Group A and again this prediction was supported; while Group B's throttle performance was significantly worse for their test session than their baseline session, there was no significant difference in their off-center performance (unlike Group A who had a significant degradation in off-center performance). We had expected Group C to have the best overall performance (i.e., to show no significant degradation in their test RTO performance compared to their baseline performance) since these participants had the highest expectancy for the occurrence of a RTO due to the nature of their instructions. Indeed, Group C did not have a significant difference in their test and baseline throttle performance and their test off-center performance was actually significantly better than their baseline performance. Thus, the expectancy effect can be eliminated by an explicit warning immediately prior to a RTO. Group D was expected to perform significantly worse on their test throttle and off-center RTO performance than their baseline RTO performance. Indeed, they did perform worse for throttle performance but there was no difference between their off-center performances. The same was found for Group E, negating our hypothesis that the misleading warnings would not lose their effect if they varied in semantic content. Thus, repeated misleading warnings of a RTO (whether of varying semantic content or of the same content) do not mitigate the expectancy effect. In addition, the inconsistent and unexpected throttle performance seen in the second test RTO suggests the need for further investigation.

Thus, we found that the group who had the lowest expectancy for a RTO (Group A) performed the worst and the group that had the highest expectancy for a RTO (Group C) performed the best. This was found after just one hour of training and when the unexpected event occurred just minutes after the training session. This has serious implications for pilots who are trained and then go years before experiencing an unexpected event. Pilots who are not expecting a RTO may be slower to respond appropriately when a RTO occurs. In addition, we found that repeated misleading warnings of a RTO occurring were not effective in mitigating the effect. This result could be comparable to the mental checklist that a pilot rehearses before every takeoff. A pilot is supposed to mentally remind him or herself what to do in case of an emergency situation such as a rejected takeoff. However, this rehearsal might lose its effectiveness since a RTO usually does not occur – much like the misleading warnings we gave our participants lost their effect after a long series of NTOs. More research is needed to determine what other factors might mitigate the expectancy effect and what can be done to help pilots respond appropriately to these unexpected events.

Retention of Airline Pilots' Knowledge and Skill

A current issue for airlines is to determine the appropriate recurrent training schedules for flight crews. The present study offers empirical data on the retention of airline pilots' knowledge and skill. We report data on flight crews evaluated at six and 12-months post training. Both normal and emergency flight maneuvers experienced significantly higher decay at the 12-month assessment when compared to the six-month assessment. Maneuvers briefed before the evaluation, allowing the pilot to mentally review the appropriate procedures, showed less decay compared to non-briefed maneuvers for both the six and 12-month assessments. The greater decay suffered by the first look maneuvers suggests engaging in mental rehearsal before on-the-line flying might mitigate the decay. The results, furthermore, suggest the importance of upholding a six-month training schedule.

Under the Federal Aviation Administration's Advanced Qualification Program (AQP), airlines may propose new recurrent training intervals for pilots. These proposed training intervals must be justified by empirical pilot performance data (FAA Advisory Circular 120-54, 1991). The purpose of this study is to inform these planning efforts by offering empirical data on airline pilots' knowledge and skill retention.

Airlines typically retrain pilots on all critical skills and knowledge in a short (e.g., three days) period, but provide little training during the ensuing retention interval. This retraining model implicitly assumes all knowledge and skills decay at the same rate. However, there appears to be little published support for this assumption. Are the decay rates for emergency maneuvers, which are rarely if ever practiced on the line, the same as highly practiced maneuvers? Is a 12-month retention interval too long? Valid and reliable performance data are needed to answer these and related questions.

Few empirical studies address skill decay of airline pilots, but there are several relevant reviews of skill decay. Arthur, Bennett, Stanush, and McNelly (1998) identified variables that exert the greatest effect on skill decay: retention interval, differences in context between learning and recall, degree of over-learning, cognitive vs. physical task characteristics, as well as training and testing methods. Their review suggests decay increases with longer retention intervals, but decay rates may vary drastically between physical tasks, where proficiency lasts longer, compared to cognitive tasks.

Marmie and Healy (1995) investigated recall of a complex task over a period of disuse - participants' accuracy in tank gunner skills while in a TopGun tank simulator. This procedurally complex task involved both physical and cognitive tasks, but Marmie and Healy found high recall rates even up to 22 months after training. They did report a significant decline in percent hits on target between the one-month retention test (97.5% accuracy) and the six-month retention test (95.2% accuracy). Although this decline was statistically significant, both performances were near the ceiling. The high recall was linked to the readily available ability to reinstate procedures used in tank gunner skill training at the time of testing due to the use of the simulator environment.

In a study of piloting skills, Childs and Spears (1986) found cognitive and procedural elements of flying decayed more rapidly than control-oriented skills. Pilots were observed to have difficulty correctly identifying cues and classifying situations, but once a situation was correctly classified, they remembered what to do. Therefore, a challenge exists in preparing pilots to recognize problems as they occur and implement the correct response.

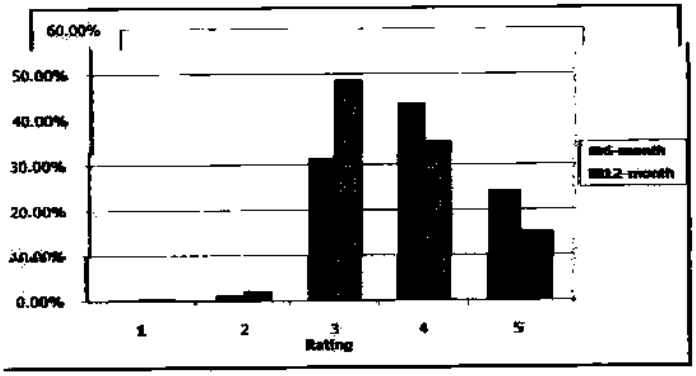
Finally, over-learning - continuing to practice a skill even after proficiency has been acquired - appears to be one of the strongest moderators of skill decay (Driskell, Willis, & Copper, 1992). However, AQP's focus on proficiency-based training allows skills to be trained only until proficiency is achieved. Such a model achieves training efficiency but perhaps at the cost of higher subsequent rates of skill decay. Proficiency-based training needs to be guided by empirical data so the limited training resources are used to achieve maximum proficiency of all critical skills and knowledge throughout the retention interval.

Method. Data were obtained from flight crews evaluated for continuing qualification at either 6-months (N = 274) or 12-months (N = 192) post training. All crewmembers were from the same airline and were trained on large commercial aircraft. Pilots were evaluated in a full motion simulator on 12 emergency maneuvers (e.g., engine out landing crew coordination, rejected landing, visual engine out landing) and 25 normal maneuvers (e.g., approach briefing, approach normal procedures, ground operations normal procedures). Four of the maneuvers (two emergency and two normal) were assessed in a first-look mode and were not briefed to the crew before being evaluated. All maneuvers were rated on a 5-point scale (1 being unsatisfactory and 5 being excellent).

Results. The rating distributions for the 6-month and 12-month assessments were compared to initially judge if any difference existed in the ratings assigned at the two assessments. Figure 1 displays the rating distributions for the two assessments. There was a significant difference in the distributions, Chi Square (4) = 296.09, $p < .001$. Of particular interest is the increase in the assignment of ratings 3 and below (standard, poor and failing performances) in the 12-month evaluation and the subsequent decline in the assignment of ratings 4 and 5 (good and excellent performance). The 6-month assessment shows an opposite pattern with the largest proportion of ratings being a 4.

To specifically test the difference between the proportion of low and high ratings and determine if higher ratings were assigned in the 6-month assessment compared to the 12-month assessment, the ratings were divided into two categories of high (rating 4 and 5) and low (ratings 1, 2, or 3). A chi-square test was performed on these distributions. A significant difference was found between the distributions (Chi Square (1) = 281.47, $p < .001$), indicating a larger proportion of high ratings were assigned at the 6-month assessment.

Figure 1. Rating distribution for 6 and 12-month assessments.



Evaluators generally believe that simulators respond somewhat differently than an aircraft. There was some concern that part of the decay found at the 12-month interval could be due to a lack of simulator familiarity. Thus, part of the decay observed from six to 12-months could be related to skills specific to flying a simulator. To assess this possibility we looked at the relationship between ratings and the order in which the maneuvers were presented. A positive correlation could be interpreted to mean that performance improves as pilots gain experience with the simulator. For the six and 12-month evaluations, the correlations were -.42 ($n = 38$, $p < .01$) and -.46 ($n = 38$, $p < .01$) respectively. These negative correlations offer no support for the hypothesized "simulator effect', instead indicating that performance levels are declining with maneuvers occurring later in the assessment sequence. This observation is more consistent with some type of fatigue factor at work causing the ratings to get worse as the assessment progresses or some type of evaluator bias in becoming less lenient as the evaluation continues.

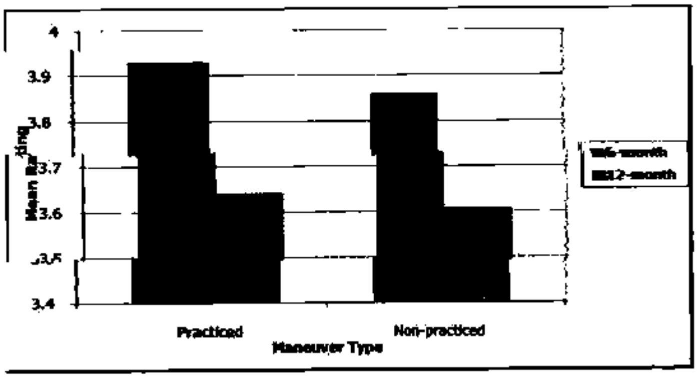
Mentally practicing a task can aid in its performance if the task is to be completed shortly after the practice (Driskell, Copper, & Moran, 1994). To examine the effect of mental rehearsal on the performance of flight skills, the ratings received for maneuvers briefed before assessment were compared to the ratings on first look maneuvers. The mean differences of these two types of maneuvers for the retention intervals are shown in Figure 2. A two-way ANOVA showed no significant interaction between the assessment month and maneuver type, but both main effects were significant. The 6-month assessment had a significantly higher mean rating than the 12-month assessment, $F(1, 70) = 51.17, p < .001$, and, the first look maneuvers received significantly lower ratings than the briefed maneuvers, $F(1, 70) = 27.81, p < .001$.

Figure 2. Changes in ratings assigned for briefed and first-look maneuvers as a function of tention interval.

Aside from mental rehearsal, practicing the skill regularly may also retard the decay of a skill. To effectively test this hypothesis, we compared the performance of normal maneuvers with the performance of emergency maneuvers. Emergency maneuvers were chosen for this comparison as they are practiced only very rarely, if at all, in real flights. As emergency maneuvers are performed much less frequently than normal maneuvers, it is conceivable that skill on these items would degrade faster and would, therefore, need to be refreshed more often.

Figure 3 shows the mean ratings received for normal (practiced) and emergency (non-practiced) maneuvers at the 6 and 12-month assessments. A two-way ANOVA showed no significant interaction between type of maneuver and assessment period. However, there was a significant main effect for the decay of ratings from the 6-month assessment to the 12-month assessment ($F(1, 70) = 55.70, p < .001$) indicating a higher mean rating assigned at the 6-month assessment. There was no significant difference found between the mean rating assigned to the practiced vs. non-practiced maneuvers. Although it was shown previously that there is a need for some type of refresher training or mental rehearsal before engaging in the flight task, there was no evidence for the general belief that a skill not practiced in normal flight (presumably the emergency maneuvers) will decay at a faster rate.

Figure 3. Changes in ratings assigned for practiced and non-practiced (i.e. emergency) maneuvers as a function of retention interval.



Discussion. There is a paucity of research in the area of skill decay as related to flight skill and training interval. The purpose of this study was to evaluate the amount of decay in flight skills that would occur if the continuing qualification training and evaluation were extended to a year instead of six months. The present findings suggest complex skills in operating an aircraft suffer more decay when the retention interval is extended to 12-months as compared to 6-months. This decay occurs in normal (practiced) and emergency (non-practiced) maneuvers.

Driskell et al. (1994) reported that tasks containing cognitive components benefit the most from mental rehearsal. Childs and Spears (1986) suggest alleviating flight skill decay by having pilots practice mental rehearsal in the cognitive and perceptual aspects of flying. Mental rehearsal is particularly important in preparing for those maneuvers the flight crew does not encounter regularly or is unable to practice (i.e. emergency maneuvers). In the present study, pre-briefing pilots about the maneuvers they were to perform benefited their performance, however, they still showed a significant decline moving from a 6 to a 12-month retention interval.

The analysis comparing first-look and briefed maneuvers addresses the issue of the pre-evaluation briefing effectiveness. The more rapid decay of the maneuvers observed and rated in the first-look condition points to the efficacy of a refreshment briefing directly before participating in the flight or maneuver. That a briefing can retard decay suggests that at least part of the decay in first-look maneuvers may be due to a mental rehearsal component and not conventional practice effects. Thus, the decay problem may also be related to difficulty in retrieval instead of an actual skill loss. These findings also suggest the importance of using first-look evaluations to obtain accurate assessments of pilots' performance.

One suggestion for incorporating mental rehearsal into the daily routine is the inclusion of checklists. Pilots and crew are instructed to mentally rehearse emergency procedures before every flight, and Childs and Spears (1986) discuss how the use of cognitive pre-training has shown benefits in improving pilots' situational awareness. However, it is difficult for a crew to systematically review every emergency that may be encountered in the flight. To assist the crew in reviewing procedures and increase the use of mental rehearsal, carriers and crews employ the use of checklists. These checklists not only function to ensure the crew observe all proper procedures in readying the aircraft for flight, but they also help prompt the individual to consider appropriate actions in case of an emergency. Future studies should be done to investigate other forms of rehearsal that can easily be incorporated into the pilot's daily routine.

Assessment of CRM Subskills

The goal of AQP is to maintain pilot proficiency on all critical skills. Thus the implementation of AQP means that carriers must accurately assess pilot skills on a timely schedule. An important advantage of this approach is that it allows the carrier to focus training only on those skills that need additional training and only when the training is needed. The potential benefits, however, are dependent on the assessment tools being capable of diagnosing specific skill deficits. Clearly, if the assessment methods are unable to assess anything more specific than an overall CRM deficit the benefits to the carrier are less than if it is possible to identify a specific CRM subskill (e.g., workload management) that is in need of additional training.

It is widely assumed within the airline industry that pilot performance can be accurately assessed in terms of Technical and CRM types of skills, and that CRM skills can be further differentiated into a number of subskills (e.g., situational awareness, workload management, crew communications, etc.,). While this assumption would appear to have a great deal of face validity (e.g., curricula are designed around this assumption) we were unable to find any empirical evidence that directly supported the assumptions that: 1) there are distinct CRM ;ubskills; and 2) that they can be reliably evaluated using available assessment tools.

The purpose of this project was to assess whether IE grading of CRM discriminates among subskills. In Study I we analyzed an extensive database of LOE OB grades from a major carrier. In Study 2 we analyze a large sample of Line-performance grades provided by Dr. Robert Helmreich's research group at the University of Texas.

General Methodology. Although the Line-performance grades are based on somewhat less specific items than the typical LOB OB items we will simply refer to the graded items as OBs throughout this report. The data were analyzed using the following statistical methods:

Pairwise-correlations. This method simply compares the average correlations between performance on OBs that are either from the same subskill or from different subskills. If the subskill assumption is valid correlations within a subskill should, on average, be significantly higher than correlations between OBs from different subskills.

Regression analysis. Using a regression approach the variance in OB grades is partitioned into three components: 1) a subskill effect; 2) a context effect (i.e., OBs within, versus between, phases of flight; and 3) a general skill effect (i.e., the correlation between OBs from different subskills and different phases of flight.

Cluster analysis. Rather than assuming the existing CRM categories we used various clustering methods to assess what if any categories emerged from the data

Both of these methods are correlational approaches so the subskill effect is measured in terms of the proportion of total variance in grades that is exclusively related to the influence of grading on the basis subskills.

LOE OB Grades The analyses were conducted on 348 crews in continuing qualification with a major US carrier.The data are based on five different LOEs, 12 different event sets, and a total of 72 different OBs The LOE OBs were designed to discriminate among five different CRM subskills (crew coordination, communication, decision making, planning, situational awareness) and one Technical category. Variance in grades was partitioned into the following three effects: 1) General Skill – the average (squared) correlation between grades from a different subskill and from a different phase of flight; 2) Context – the average (squared) correlation between grades from the same phase of flight, but different subskills; and 3) Subskill – the average (squared) correlation between grades from the same subskill, but different phases of flight. The unique variance related to subskill is estimated by the difference between square Subskill correlation and the squared General skill correlation (i.e., the difference between within subskill and between subskill variance).

Primary Findings. Grading was done on a 5-point scale with a grade of 1 being outstanding and a grade of 5 being a fail. Grades were distributed across the 5-point scale as follows: 28% 1, 49% 2, 24% 3, 1% 4, and < 1% 5. From this it can be seen that grading was actually on a 3-point scale and the vast majority of grades were 2. This low variability in LOE grades is a common occurrence and can have a significant effect on the magnitude of Pearson-based correlations. To investigate how this may influence our estimates of the subskill effect we used a Monte Carlo procedure where we approximated the distributional characteristics of the present data and then systematically assigned grades to individuals in a manner that would either maximize or minimize the subskill effect. When grades were assigned randomly for each pilot the average within and between subskill squared-correlations were .168 and .168 respectively. When the grades were assigned to maximize the subskill effect the within and between squared-correlations were .757 and .109 respectively. Thus the subskill effect was .00 as it should be with random assignment of grades to individuals and the maximum subskill effect was .548. When we repeated the above analyses on a data set where the grades were distributed almost equally across all five grades and assigned randomly across individuals the within and between squared-correlations were .00 and .00 respectively. When the grades were assigned to individuals to maximize the subskill effect the within and between squared-correlations were .846 and -.036 (i.e., a maximum subskill effect of .882). In concluding, the distributional characteristics of LOE grades appears to have a profound effect on Pearson Product-moment correlations. In the present context it: 1) substantially inflated the average correlation between unrelated OBs (i.e, the General skill effect); and 2) significantly diminished the upper limit of a subskill effect. More work is planned to investigate whether some nonparameteric correlational measures maybe less affected by the distributions often found with pilot performance ratings.

A Cluster analysis of the CRM OBs only showed evidence of clustering for event set or phase of flight. The same analysis of Technical OBs also showed evidence of a phase of flight cluster, but in addition showed clustering of four technical categories (e.g., flight management system items clustered)

Perhaps the single most important finding came out of the pairwise-correlational and regression analyses, which indicated that less than 1% of the variance in CRM grades is related to a subskill effect. Thus there was no support for the idea that LOE grades are discriminating CRM subskills.

Whereas it initially appeared that certain OBs were better at discriminating CRM subskills, more extensive analysis over three LOEs revealed that there were no OBs that were superior subskill discriminators across all three LOEs (e.g., an OB that is diagnostic of situational awareness in one LOE is not diagnostic of situational awareness in another LOE).

The correlation between OBs within CRM or within Technical was no higher than the correlation between CRM and Technical OBs. This suggests that OBs were not discriminating between CRM and Technical skills. Context, or phase of flight, accounts for more variance (9%) in grades than subskill. . Thus a crew performs more consistenly within a phase of flight (regardless of the subskill) than it does for a specific CRM subskill across phases of flight. Looking at only the grades from IEs that had the highest variance in their grading (presumably the most discriminating IEs) did not increase the effect of subskill. Looking at only the grades from the most difficult phases of flight (e.g., landing) also had no effect on the contribution of subskill to performance.

Line-check Grades. These data were collected by evaluators flying the jump seat with five major US carriers over a four-year period. This resulted in a total of 3,241 flights that met the criteria for inclusion in the study. The evaluators were both pilots from the carriers participating in the study and University of Texas personnel who were trained to conduct the Line-check grading. Grading was done on 10 CRM types of measures and 3 Technical types of measures. The 10 CRM measures were assumed to measure two different types of skills (Team and Task). There were five Team measures (Leadership, Team Environment, Decisions Stated, Effective Inquiry, and Assertion) and five Task measures (Task Distribution, Stay Ahead of Curve, Vigilance, Briefing Content, and Tasks Prioritized). The three technical measures were Sterile Cockpit, Altitude, and Checklist Management.

Primary Results. The distribution of grades in these data are even more constrained than what we reported for the LOE data in Study I. With a grade of 4 being outstanding the percentage of grades were 12%, 77%, 9.5%, and 1.5% for the grades 4, 3, 2, and 1 respectively. We did not conduct a Monte Carlo analysis with these constraints, however it can be assumed that the trends revealed with the LOE data are likely more pronounced with the Line-data.

A cluster analysis revealed evidence for the Team, Task, and Technical categories when the analysis was conducted across phases of flight, however when the cluster analysis was done on each separate phase of flight there was only evidence for a CRM and a Technical category.

Pairwise-correlations and Regression analysis indicated that approximately 4.5% of the variance in grades could be attributed to the influence of subskill. This effect was statistically significant ($p < .001$) and continued to appear across the different carriers and different IE graders (e.g., UT graders versus pilot graders).

The mean correlation between items within the same subskill but across phases of flight was substantially higher (.80 for Team measures and .75 for Task factors) than the mean correlation between different subskills rather they were within a factor (.38 for Team and .50 for Task measures) or between factors (.36). The higher correlation within a subskill across phases of flights may be the first strong evidence of a subskill effect that we have found. However, it is possible that these correlations are inflated because of the physical layout of the gradesheet. Grades for the same subskill are entered across columns in a single row. We have no direct evidence that physical arrangement of the data entry matrix influenced IE grading, but in Study I where the data entry was different the correlations between similar OBs was not higher than randomly paired OBs.

The validity of the distinction between Team and Task factors was further assessed by determining whether one of them was consistently a better predictor of the three Technical measures. This did not appear to be the case. Task accounted for more unique variance for Sterile Cockpit and Checkist Management, but Task accounted for more unique variance for Altitude. However, a more detailed analysis suggested that the predictiveness of the Technical measures was better characterized in terms of the specific measures within the Team and Task factors. Some measures (e.g., Leadership, Decisions Stated, Vigilance) tended to be relative good predictors for the Technical measures, others were never predictive (e.g., Effective Inquiry, Assertion, Briefing Content) and one measure (Briefing Content) was consistently negatively correlated with the three Technical measures.

Summary and Implications. Over two large databases that varied on several dimensions (e.g., Simulator- versus Line-data, taxonomies categorizing CRM subskills into 5- versus 2-subskills, and grading by highly experienced pilot IEs versus non-pilot student IEs) the results consistently showed that less than 5% of the variance in OB grades was related to a subskill effect. It is true that the subskill effect was likely underestimated because of the constrained distribution of the grades in both studies. However, until IEs distribute their grades more equally across the grade scale the present estimates of the subskill effect must be considered valid. If this is the current state of affairs, there would seem to be little value in continuing to CRM data in the manner it was done in the present two studies.

The present findings leave us with two possible implications: 1) that the fundamental assumption underlying CRM subskills (i.e., that they can be diagnosed and selectively trained) is false; or 2) the subskill assumption is valid, but there are various methodological problems (e.g., problems with the methods or tools, or problems with the IEs use of these tools) that prevent their assessment.

Modeling I/E Grades from Flight Parameters

The primary goal of this project was to build statistical models based solely on simulator flight parameters that were capable of predicting I/Es grading of critical maneuvers during maneuvers validation. We began with flight parameter data files consisting of a matrix, where the columns are different parameters and the rows are successive readings (mostly at 1-second intervals). To use these data to predict grades we first selected a subset of parameters that have some possibility of being related to behavioral grades. This was done on the basis of qualification standards and SME input. We then performed various transformations of the raw flight parameter data that were designed to capture the desired measures. Typically this involved selecting particular temporal segments of data from specific flight parameters and then using some statistic to covert the vector of temporal data into a single value. A simple example would be to compute the mean air speed for a segment of time surrounding 'touch-down' to estimate landing speed.

We collected data and developed transforms for four maneuvers (Rejected Take-offs, V1Cuts, Normal Take-offs, and Normal Landings) from an MD80 simulator.

Model Validation. We used Classification Tree Analysis to identify the best combination of flight parameter-based predictors of behavior grades. Because of the complexity of these models (e.g., there can often be six or more predictor variables in the models) and the fact that there are a disproportionate small number of low grades (e.g., 20%), we need fairly large sample sizes (e.g., 200 at a minimum) before the models can be expected to stabilize and generalize to new data samples. When the sample was sufficiently large for a maneuver we built a model on a random sample of half of the observations and then tested the model on the remaining sample.

Summary Report of Models. When we have completed the collection and analysis of simulator flight data, a report discussing the success of the models for each of the four maneuvers will be written. Success will be evaluated in terms of predictiveness (i.e., the proportion of crew grades that are correctly predicted), generalization of model to new data samples, and face validity of the predictors as based on SMEs interpretation of relevant qualification standards.

Assessing Importance of Audio-Visual Cockpit Information. In all of our modeling efforts we attempted to predict the grading performance of I/Es who had access not only to the flight parameter information, but to what we earlier referred to as 'crew information' (i.e., all of the audio-visual information pertaining to the crew that occurred during the performance of the maneuver). Thus, to the extent that verbal and nonverbal communications, as well as arm/hand movements of the crew affect behavioral grades the predictiveness of our models would be expected to suffer. The results of our modeling efforts is included as a pdf attachment.

Generalization of the Model to FOQA Data. The models developed from the above analysis have two potentially important applications. First, they can serve as a basis for training I/Es. The models reveal what experienced and reliable I/Es are actually using as a basis for grading maneuvers. This information can now be easily communicated and illustrated to other I/Es.

The second application is to provide an objective basis for evaluating how well crews are flying critical maneuvers on the line. However, for the FOQA data to be used in this regard it must first be appropriately analyzed. It is here where the models that we developed from the simulator data can be used to analyze the FOQA data. Applying the models to the FOQA data will result in a behavioral grade. By analyzing samples of FOQA data we can determine whether crews are flying maneuvers at the same level they are in the simulator.

From our previous work with flight simulator data, we determined that several of the qualification standards for grading a landing were directly or indirectly related to the flight parameter transforms that served as the basis for predicting the landing grades given by an instructor/evaluator (IE) in the simulator. Further, there was sufficient variability in the values of these transforms to impact grades. Nonetheless, we found the same grade given to flights for which there seemed to be considerable variability in the flight parameters.

To examine this issue further we identified 29 landings (out of approximately 700 landings) that had touchdowns which exceeded the touchdown zone allowed under qualification standards. Of these 29 landings, 3 had low grades (1 or 2) with the remaining graded as 3 or 4.

This finding raises some important questions regarding IE grading. One interpretation is that the IEs recognized the exceedences on touchdown, but because of various contextual factors (e.g., weather conditions, approach limitations at the specific airport, engine out, etc.) they made reasoned allowances (i.e., given the conditions of the landing the crew performed as well as could be expected). A less favorable interpretation is that IEs are not placing sufficient weight on qualification standards in their grading.

A Vector Representation Approach to Analyzing Aviation Safety Reports

Aviation safety programs such as the National Transportation Safety Board (NTSB) and voluntary safety reporting programs such as the Aviation Safety Action Program (ASAP) and the Aviation Safety Reporting System (ASRS) contain large and continuing expanding sets of accident/incident reports. Many of these reports contain a narrative description of the event written from a first-person perspective. There is a growing recognition that these narrative descriptions contain a wealth of safety-related information. However the sheer volume of reports presents a formidable challenge in efforts to classify, summarize, or extract meaningful information from them. At the time of this writing the total ASRS report count was listed as 167,414. In this report, we describe an effort to develop a method for automatic text analysis that could be used to extract meaningful information from narrative safety reports. We investigate a vector representation approach analyzing a large corpus of text, with ASRS narrative reports serving as the target application.

Despite the processing of incoming reports by analysts at NASA and the National Aviation Safety Data Analysis Center (NASDAC), the primary means of extracting document collections is via keyword searches using the ASRS Database Query Tool, even though there are also several proprietary data mining tools available for licensing, such as those listed on the FAA Human Factors Workbench Data Mining Tools web page http://www.hf.faa.gov/WorkbenchTools.

ASRS narratives contain meaningful descriptions of flying events written, for the most part, by experts in the aviation world. Our goal is to develop a statistical method for aiding a safety analyst in extracting information from ASRS reports. Some specific applications include discovering the main topic of a set of narratives, providing quantitative measures of the similarity between documents, or between terms and documents, and identifying emerging trends over in time or other changes in context.

Text Analysis Methods. Methods for performing document search have been around for many years, but these early methods relied on simple keyword matches. Literal keyword matches often fail because multiple concepts can be indexed by a single term (polysemy) or a number of terms can be indexed by a single concept (synonymy). Relevant documents are missed because the documents contain synonyms of the key word or irrelevant documents are generated because the documents contain the key words but used in a different sense. Recent approaches discussed in this report have attempted to apply more sophisticated search methods.

Quorum & Perilog. McGreevy has developed text processing tools aimed specifically at analyzing the content of ASRS reports. His Quorum model (McGreevy, 1996; 1997; 2001) performs keyword searches, phrase searches, phrase generations, and phrase discoveries. ASRS incident narratives are searched for key word phrases that contain one or more user-specified keywords in from a selected context. Phrase searches retrieve narratives that contain one or more user-specified phrases and then ranks the narratives based on their relevance to the phrases. Phrase generations produce a list of phrases from the ASRS database that contain a user-specified word or phrase, and phrase discoveries find phrases that are related to topics of interest.

Quorum uses word-pair models derived from prominently associated word pairs contained in the narratives. Words which are frequently found close together are considered to have a greater degree of association than those found further apart, and each pair is assigned a number reflecting their degree of relatedness. The most prominent word pairs are considered to represent prominent concerns in the reports. Quorum ranks the relevance of ASRS narratives by using a proximity-weighted co-occurrence metric to discover and rank prominent textual relations in the narratives.

A more recent version of the software, called Perilog (McGreevy, 2004), allows for a context sensitive text analysis. Users can search for relevant phrases using a keyword-in-context search method. Phrases that are contextually and situationally associated with a keyword can be found using the phrase review feature. Perilog's phrase generation tool allows a user to find specific concepts involving a prominent word. The output of phrase generation is a list of phrases that are likely to occur in the text, with the more likely phrases appearing toward the top of the list.

Vector Representation Method. A quite different approach to text analysis is to represent each document and each term as a vector of real values. This approach begins by constructing a term by document frequency matrix where rows of the matrix correspond to unique letter strings (words) and columns of the matrix correspond to distinct documents. The cells of the matrix contain how often each term occurs within each document. Hence, each term is represented as a vector of frequency values across documents, and each document is represented as a vector of frequency values across terms. Large collections of documents can result in matrices on the order of tens of thousands of rows and columns.

The columns of the term by document frequency matrix are typically a single document, such as an encyclopedia article, however, the columns can correspond to any unit of text including sentences, paragraphs, chapters, etc. The decision about what constitutes the appropriate scope for defining a unit of text is potentially important because all of the words within a unit of text are viewed simply as a bag of words; the order in which the words occur within the unit is lost.

It is these vectors in space that purportedly capture the semantics of words and documents. Once represented as vectors, words, phrases, and even whole documents can be compared to one another by computing their vector similarity. These similarity values between text segments offer the basis for giving meaningful answers to text analysis questions.

Preprocessing of Text. There are standard methods for processing term by document frequency values before computing similarities. Typically very frequent and infrequent terms are pruned because they carry little information. Standard lists of stop words (e.g., the, and, etc.) are automatically eliminated. Inflected words (e.g., noun-plurals, verb-tenses) are replaced by their root form in order to condense terms with the same semantic content.

In most realistic applications, the term by document matrix contains tens of thousands of rows and columns, with the number of words typically being much larger than the number of documents. But because only a small number of unique words will occur in any given document most of the cells in the matrix contain zeros. As a result the resulting matrix is very sparse.

Weighting Schemes. It is common to transform the raw frequency values in a term by document matrix by some weighting function and/or a normalization. These transformations take into account the frequency of appearance of words within and across documents. There are two basic types of weighting schemes, local and global weighting. Local weighting reflects the occurrence of a term within a particular document and global weighting reflects a term's occurrence across all of the documents. A further transformation can occur by normalizing a term vector by the length of that vector.

Popular weighting schemes are variations of the term frequency, inverse document frequency model. Term frequency refers to the number of times a term occurs within in a document, and the inverse document frequency is the inverse of the number of documents a word occurs in.

Latent Semantic Analysis. When dealing with large collections of documents the resulting term document frequency matrix, can reach dimensions in the tens or hundreds of thousands, much of which is irrelevant information. A method for further processing the transformed term by document matrix that has received considerable attention in recent years is latent semantic analysis (LSA; Deerwester et al., 1990; Ladauer & Dumais, 1997; Landauer, Foltz, & Laham, 1988) and its variant latent semantic indexing (LSI; Foltz & Dumais, 1992). LSA is a statistical method for analyzing written text. The method was originally developed to perform document search, but has now been extended to include other text analyses such as grading student essays, measuring the coherence of text, and even serving as a model of human semantics. The principal claim of LSA is that it can extract meaningful information from text without human intervention.

At the heart of LSA is data reduction method known as singular value decomposition (SVD). The claim of LSA is that SVD retains much of the lexical and semantic information in the original matrix while removing irrelevant noise. The resulting matrix has a greatly reduced dimensionality compared to the original matrix. Similarities between terms and documents can then be computed with the reduced matrix. These similarities purportedly capture meaningful semantics of the domain.

Early applications of LSI typically involved document retrieval, where document vectors in the reduced term document frequency matrix were selected by their similarity to the queries submitted to the system. This resulted in a list of documents ranked according to their similarity to the query.

The disadvantage of the method involved the computational burden of maintaining and updating the matrix with new documents as well as computing similarities for queries not previously submitted. Consequently it was more the object of theoretical research than a practical application. However, the value of the SVD as a general method of feature extraction and dimension reduction was increasingly being appreciated, particularly in situations where the vector space model was being investigated with a variety of classification methods used in data mining and machine learning (Husbands, Simon & Ding, 2001; Tang, Shepherd, Milios & Heywood, 2005). It was seen a potential way to deal with the so called curse of dimensionality (Aggarwal, Hinneburg & Keim, 2001), which plagued virtually any method dealing with distance-based classification methods as they were applied to the vector space model.

The most typical example of feature extraction is the creation of linear combinations of features, where the features are terms in the term document matrix. Principal components analysis is a similar method also based on SVD. Another way to attenuate the problems encountered with high dimensionality is through feature selection, which is implicitly incorporated into LSA in the construction of the term lists (or possibly post-reduction depending on how the solution is applied), but there are no limitations on the manner in which (or how aggressively) that process is applied.

Mathematics of SVD. SVD is a mathematical technique for decomposing an arbitrary matrix into the product of three other matrices. It is a type of eigenvector decomposition. Beginning with an input matrix X consisting of a term by document frequency matrix, X is decomposed into matrices

T, S and O: $\underset{t\times o}{X} = \underset{t\times r}{T_o} \cdot \underset{r\times r}{S_o} \cdot \underset{r\times o}{O'_o}$

T_o and O_o are both column orthonormal. Their columns are of unit length and the inner product of any two different columns is zero. S_o is a diagonal matrix of singular values. All off-diagonal elements are zero and all diagonal elements (the singular values) are non-negative real values, typically ordered by decreasing value. $T^T T = I$, where I is the identity matrix. $TT^T = P$, where P is a projection matrix onto the space spanned by the columns of X. If the rank of X is *r*, then the dimensions of T are *t* by *r*, the dimensions of O are *o* by *r*, and the dimensions of S are *r* by *r* with no zero singular values.

SVD has the property that it provides the best lower rank approximation of a matrix X in terms of the Euclidian matrix norm. If T_k is the *t* by *k* (*k*≤*r*) matrix found by removing the *r*-*k* columns from T, then the *k* columns remaining in T_k correspond to the largest singular values in S. By keeping only the *k* largest singular values of S_o along with their corresponding columns in T_o and O_o, the resulting new matrix X' is the unique matrix of rank *k* that is closes to X in a least squares sense. This reduced matrix X' has eliminated much of the noise in the original matrix while retaining the critical associational structure of the original matrix. It is this reduced matrix X' that is used to compute similarities among terms and documents.

Any set of terms within the matrix can be represented by taking the centroid of the vectors representing the individual terms within the set. That is, vector representations of new sequences of terms can be derived by simply adding the vectors of their constituent terms. In this way the similarity between any set of terms within the original matrix can be computed. A similarity between units of text is typically measured by their cosines (cos) in the resulting high dimensional semantic space. The measured relation between words or phrases is assumed to reflect the extent to which they have similar phrase meanings, not simply the relative frequency with which they co-occur in the same documents.

LSA & Meaning. LSA assumes that a term's meaning is contained in the term's pattern of occurrences across a sample of text units. Underlying this idea is that meaning is contextually dependent. A term's meaning is ultimately determined by how (where) the term is used. Two terms that occur together within text units across a very large sample of texts would likely be semantically related. Further, two related documents may be represented similarly even though they do not share any keywords. This high similarity may occur, for example, if the terms used in each of these documents co-occur frequently in other documents. Hence, it is a derived or inferred similarity that is the basis for LSA similarity.

It is further assumed that there exists a higher-order structure in the association of terms across documents. It is this higher order semantic structure that LSA purports to extract through the use of SVD. LSA claims that the similarity between terms or phrases in the reduced space is a better indicator of true semantic similarity than similarity measured in the original term space. LSA presumes that the overall semantic content of a passage, such as a paragraph, abstract, or a fully coherent document, can be usefully approximated as a sum of the meaning of its words. One of the goals of the current work is to test some of these claims of LSA using the ASRS database of text.

In an attached pdf report we describe in detail a series of investigations aimed at better understand the role of weighting, transformations, dimensionality of SVD solutions, and higher order structure in text analysis. We were particularly interested in understanding how these various parameters behave in text analyses applied to ASRS reports. In addition, we often applied these tests to a large standardized text database (Reuters) in order to allow us to compare our results with results reported in the literature.

References

Advisory Circular 120-54 (1991). Advanced qualification program. U.S. Department of Transportation and Federal Aviation Administration.

Aggarwal CC, Hinneburg A, Keim DA (2001). On the surprising behavior of distance metrics in high dimensional space. *Proceedings of the Eighth International Conference on Database Theory, Volume 1973 of Lecture Notes in Computer Science,* 420-434.

Arthur, W., Jr., Bennett, W., Jr., Stanush, P. L, & McNelly, T. L. (1998). Factors that influence skill decay and retention: A quantitative review and analysis. *Human Performance, 11,* 57-101.

Berry, M. W. (1999). *Large-scale sparse singular value computations.* The International Journal Conference on Uncertainty in Artificial Intelligence (UAI'99), San Francisco,CA, 1999,

Childs, J. M. & Spears, W. D. (1986). Flight-skill decay and recurrent training. *Perceptual and Motor Skills, 62,* 235-242.

Deerwester S, Dumais ST, Furnas GW, Landauer TK & Harshman R (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science,* 41(6), 391-407.

Driskell, J. E., Copper, C., & Moran, A. (1994). Does mental practice enhance performance? *Journal of Applied Psychology, 79,* 481-492.

Driskell, J. E., Willis, R. P., & Copper, C. (1992). Effect of overlearning on retention. *Journal of Applied Psychology, 77,* 615-622.

Foltz PW & Dumais ST (1992). Personalized information delivery: an analysis of information filtering methods. *Communications of the ACM,* 35(12)(Dec), 51-60.

Husbands P, Simon H & Ding C (2001). On the use of singular value decomposition for text retrieval. *Proceedings of the SIAM Computer Information Retrieval Workshop, October 2000.*

Kintsch W (2002). The potential of latent semantic analysis for machine grading of clinical case studies. *Journal of Biomedical Informatics,* 35, 3-7.

Kirby, N. H. (1976). Sequential effects in two-choice reaction time: Automatic facilitation or subjective expectancy? Journal of Experimental Psychology: Human Perception and Performance, 2(4), 567-577.

Landauer TK & Dumais ST (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review,* 104(2), 211-240.

Landauer TK, Foltz PW, & Laham D (1998). Introduction to Latent Semantic Analysis. *Discourse Processes,* 25, 259-284.

Marmie, W. R., & Healy, A. F. (1995). The long-term retention of a complex skill. In A. F. Healy & L. E. Bourne, Jr. (Eds), *Learning and memory of knowledge and skills: Durability and specificity* (pp. 30-65). Thousand Oaks, CA: Sage Publications.

Mattes, S. Ulrich, R. & Miller, J. (1997). Effects of response probability on response force in simple RT. The Quarterly Journal of Experimental Psychology, 50A (2), 405-420.

McGreevy, M. W. (1996). *Reporter concerns in 300 mode-related incident reports from NASA's Aviation Safety Reporting System.* NASA TM-110413. Moffett Field, CA: Ames Research Center.

McGreevy, M. W. (1997). *A practical guide to interpretation of large collections of incident narratives using the QUORUM method.* NASA TM-112190. Ames Research Center, Moffett Field, California.

McGreevy, M. W. (2001). *Searching the ASRS database using QUORUM keyword search, phrase search, phrase generation and phrase discovery.* NASA TM-2001-210913. Ames Research Center, Moffett Field, California.

McGreevy, M. W. (2004). *Using Perilog to explore "Decision Making at NASA".* NASA TM-2004-XXXXXX. Moffett Field, CA: Ames Research Center.

Posner, M. I., Snyder, C. R. R. & Davidson, B. J. (1980). Attention and the detection of signals. ournal of Experimental Psychology: General, 109(2), 160-174.

ang B, Shepherd M, Milios E & Heywood M (2005), Comparing and combining dimension reduction techniques for efficient text clustering. *Proceedings of the SIAM International Workshop on Feature Selection for Data Mining - Interfacing Machine Learning and Statistics,* in conjunction with the 2005 SIAM International Conference on Data Mining, Newport Beach, CA.

Wolfe, J. M., Horowitz, T. S. & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature, 435,* 439-440.