Aviation Human Factors Division
Institute of Aviation

University of Illinois
at Urbana-Champaign
1 Airport Road
Savoy, Illinois 61874

The Effectiveness of a Personal Computer Aviation Training Device, a Flight Training Device, and an Airplane in Conducting Instrument Proficiency Checks

Volume 2: Objective Pilot Performance Measures

Esa M. Rantanen, Nicholas R. Johnson, and Donald A. Talleur

Final Technical Report AHFD-04-16/FAA-04-6

November 21, 2004

Prepared for

Federal Aviation Administration
Civil Aerospace Medical Institute
Oklahoma City, OK

Cooperative Agreement DTFA 2001-G-037

Foreword

This research was prompted by the FAA Advisory Circular (AC) No. 61-126 (1997), which authorized the use of a Personal Computer Aviation Training Device (PCATD) to be used for 10 of the 15 hours authorized for an approved ground training device, but not for Instrument Proficiency Checks (IPCs). The research was supported under Federal Aviation Administration (FAA) cooperative agreement DFTA2001-G-037 with the Institute of Aviation, University of Illinois at Urbana-Champaign, during September 2001-November 2004. The study was sponsored by FAA Headquarters Flight Standards Service, General Aviation and Commercial Division. Dennis B. Beringer, Civil Aerospace Medical Institute (CAMI), served as the contracting officer's technical representative.

This report is Volume 2 of a two-volume final report. It is in the process of review and approval and is not at present an official FAA document. Consequently, the views expressed herein do not necessarily represent official FAA positions. Volume 1 covered results obtained from subjective pilot performance measures by certified flight instructors, instrument (CFII), who conducted the IPC flights for the study participants. This volume (Vol. 2) will describe objective pilot performance measures developed for the project and the results that they yielded. Published reports and presentations of the work on development of objective pilot performance measures are listed in Appendix A.

Many people apart from the authors have contributed to the success of this project. We express our appreciation to Mary Wilson, who scheduled participants, and to Karen Ayers, who assisted with report formatting. We also thank Bill Jones, David Boyd, Sybil Phillips and Donald Talleur, who served as the check pilots. We also thank the Institute of Aviation flight instructors who provided familiarization training in the Flight Training Device, the PCATD, and the airplane, as well as the instrument pilots for their participation in the study.

Executive Summary

To maintain instrument currency, instrument pilots must meet the recent instrument experience requirements of FAR 61.57(c) or (d) every six months. The requirements may be conducted in an airplane or an approved flight training device (FTD). If an instrument pilot fails to meet recent experience requirements within the previous 12-month period, an instrument proficiency check (IPC) must be successfully accomplished to regain instrument currency. The present study compared the performance of pilots receiving an IPC in a (PCATD), a FTD or an airplane (IPC #1) with their performance in an airplane (IPC #2). The comparison of performance in a PCATD to that in an airplane investigates the effectiveness of the PCATD as a device in which to administer an IPC. Currently, the PCATD is not approved to administer IPCs. The comparison of performance in a FTD with performance in an airplane will help determine whether the current rule to permit IPCs in a FTD is warranted. The comparison of the performance in a PCATD and a FTD permits a comparison of the relative effectiveness of the respective devices. Finally, the comparison of performance of pilots receiving IPC #1 in an airplane and IPC #2 in an airplane with a second CFII will permit the determination of the reliability of IPCs conducted in an airplane.

This study involved 75 participants (25 participants in each group: FTD, PCATD and airplane). Each participant agreed to refrain from instrument flight (either in flight or in a ground-based device) between IPCs #1 and #2. They also flew a familiarization flight in the FTD, the PCATD and the airplane prior to being randomly assigned to one of the three groups (FTD, PCATD and airplane). The participating instrument pilots in the study were in one of four categories of instrument currency: (1) instrument current, (2) within one year of currency, (3) between one and two years of currency, and (4) between 2 and 5 years of currency and they were balanced among the three groups. Pilots in the 2 to 5 year category received up to five hours of instrument proficiency training in either a FTD or a PCATD prior to the experiment

Automated objective pilot performance measures potentially enhance and expand traditional proficiency evaluation methods by an instructor pilot. In this report, we describe the development of nine specific metrics of pilot performance derived from time series of different flight parameters and examine their descriptive power and sensitivity against data from pilots with known differences in performance, as judged by an expert instructor pilot. Two autocorrelation based metrics and seven Fourier analysis based metrics are evaluated. Our results show many of these metrics to be both sensitive and diagnostic in differentiating between good and poor pilot’s performances as determined by the instructor pilot. The findings are consistent with the hypothesis that a skillful pilot will control the aircraft with a greater range of frequencies of input than a less skillful pilot, making adjustments appropriate to the circumstances whereas a poor pilot appears to make the same adjustment regardless of the actual magnitude or frequency of the adjustment needed. The results show the potential usability of performance metrics derived from time series data and that it is possible to discriminate between good and poor pilot performance using this approach. Furthermore, analysis of objective metrics between device groups confirmed the conclusion presented in Volume 1 of this report that there were no differences between the three different devices in which IPC flights were administered (airplane, Frasca FTD, and PCATD).

Table of Contents

Acronyms and Definitions

ACM	Air Combat Maneuvers
ACS	Autocorrelation function slope; slope of the autocorrelation function regression for the first 10 data points, or lags 0–9, quantifying how quickly autocorrelations tend to zero.
AE	Absolute error
ALT	Altitude
ANOVA	Analysis of Variance
AOA	Angle of Attack
BAL	Ball, or slip indicator
BFRS	Bedford Rating Scale
CAMI	The Civil Aerospace Medical Institute of the Federal Aviation Administration in Oklahoma City, OK.
CDI	Course Deviation Indicator
CFII	Certified Flight Instructor, Instrument
CI	Control Inputs (E = elevator, A = aileron, R = rudder, T = throttle)
DCGC	Standard deviation of the normalized squared magnitude of the spectral components above a criterion value of the Fourier analysis
DH	Decision height
DSC	Standard deviation of the normalized squared magnitude of the spectral components of the Fourier analysis
FAA	Federal Aviation Administration
FAF	Final Approach Fix
FAR	Federal Aviation Regulation
FDCG	Standard deviation of frequencies of the spectral components greater than a criterion value
FDR	Flight Data Recorder
FMCG	Mean of frequencies of the spectral components greater than a criterion value
FTD	Flight Training Device
GS	Glideslope
GSI	Glide Slope Indicator
HDG	Heading
HUD	Heads Up Display

IAS	Indicated Airspeed
ILS	Instrument Landing System
IP	Instructor Pilot
IPC	Instrument Proficiency Check
LOC	Localizer
LPF	Low Pass Filter; cutoff frequencies were 0.005 Hz for LPF1, 0.01 Hz for LPF2, 0.05 Hz for LPF3, and 0.1 Hz for LPF4.
MAE	Mean Absolute Error
MAP	Missed Approach point
MCGC	Mean of the normalized squared magnitude of the spectral components above a criterion value.
MEDF	Median frequency of the power spectrum
MSC	Mean of the normalized squared magnitude of the spectral components.
MTE	Mean time to exceed tolerance
NASA	National Aeronautics and Space Administration
NCGC	Number of spectral components above a criterion value
ND	Number of deviations outside tolerance(s)
NDB	Non-Directional Beacon
PCATD	Personal Computer Aviation Training Device
PIT	Pitch
PSD	Power spectal density
RMSE	Root Mean Square Error
RNG	Range
ROL	Roll
RV	Radar vector (-s, -ing)
SA	Situation Awareness
SD	Standard Deviation
SVS	Synthetic Vision System
T/O	Take off
TD	Cumulative time (duration) of deviations outside tolerance(s)
TLX	Tasl load index
TOT	Time on Target
TR	Turn Rate

UPT	Undergraduate Pilot Training
VOR	Very High Frequency Omnidirectional Range
VS	Vertical Speed

Introduction

Objective pilot performance measures are very desirable for a multitude of purposes and for many reasons. Automatic data collection has the potential to enhance and expand traditional proficiency evaluation methods by an instructor pilot by alleviating the time constraints and information overload often associated with direct observation. Furthermore, quantitative performance data can be utilized in research and subjected to various statistical analyses to reveal underlying, covert patterns in pilots' performance. Not surprisingly, objective pilot performance measures derived from flight data recorders (FDRs) or data output from simulators have a relatively long history of research (e.g., Gerlach, 1972; Vreuls et al. 1975; Stave, 1977; De Maio, Bell, & Brunderman, 1985; Benton, Corriveau, & Koonce, 1993). In spite of this, however, it appears that a relatively small number of distinct objective metrics have been utilized in research, and routine use of objective measures is still rare. There are a several notable obstacles to application of these measures. For example, Vreuls and Obermayer (1985) noted that the internal processes that drive operator actions are not observable and that few theories of human performance exist to predict what should be measured and the relative importance of each measure. Furthermore, task segmentation is necessary for automated performance measurement, making the process difficult. For maximum utility in training, performance measures need to be available and discernible in real time or as close as possible to the completion of the training session as well (Vreuls & Obermayer, 1985). However, while these problems are undeniable and difficult to overcome (cf Rantanen & Talleur, 2001), they are arguably outweighed by the potential benefits of objective measures, making continued research on the latter important.

The following review of related literature is organized in two ways: we first review the past research involving objective pilot performance measurement in a chronological order, thus providing an overview of these efforts from a historical perspective and highlighting the uneven distribution of research on this topic throughout time. The second part of the review consists of a catalog of objective performance metrics used in the past. We will review the pros and cons of each metric and evaluate their utility in measuring various aspects of pilot performance, both alone as well as in conjunction with other measures. Although our review was aimed to be exhaustive, much of past research may not have been published and hence not been available for review. Other reviews of objective pilot performance measures include Mixon and Moroney (1982), who listed 189 articles on objective pilot performance measurement, broken down into fixed/rotary wing aircraft and simulator/field studies. No attempt was made to review or critique the articles, but the number of subjects, equipment, scenarios and measures were listed. Also Benton, Corriveau, and Koonce (1993) summarized some of the practical implications for designing and implementing an automated performance measurement system to be used for basic flight maneuvers training. The authors identify the flight parameters to be measured directly (IAS, ALT, VS, pitch, bank, yaw, throttle, flaps) but did not give any details about what should be done with these direct measures in order to evaluate performance. Finally, Gawron's (2000) handbook of performance measures includes objective measures of performance and objective and subjective measures of workload, presenting various 'standard' measures (e.g., AE, RMSE) with references and a novel landing measure based on an intergral equation. The major focus of the review is on workload measures, however.

Objective Pilot Performance Measures in Aviation Research

The first published accounts on objective pilot performance measurement appear in the early 1970s from the U.S. Air Force laboratories. Consequently, many of the flight maneuvers evaluated were common to military flying. For example, Knoop (1973) studied Lazy-8 and barrel roll maneuvers in a T-37 simulator, recording IAS, ALT, HDG, vertical acceleration (±g), pitch, roll, pitch-, roll-, and yaw rates, control inputs (elevator, ailerons, rudder, and throttle), flaps, landing gear, speed brakes, and trim. Boolean measures (i.e., 1 or 0, within certain tolerance bands) for the flight parameters and deviations from reference values, (as in Knoop & Welde, 1973) were calculated. Linear combinations of Boolean measures and deviations from reference values at certain points of the flight maneuver were then compared to subjective IP evaluations (4 levels) for the flight maneuver. Using the IP generated data, Knoop showed that there was good agreement between the IP subjective ratings and the linear combinations measures. While there was still overlap in metric value between adjacent IP evaluation levels, the metric value trended with performance level (no quantitative results were given). However, when student-generated data were used, the linear combinations measures did not correspond with IP evaluations, suggesting that issues of intra-rater and inter-rater reliability had not been addressed properly. In addition, some of the standards the IP's used for evaluation were called into question. The author presented limited data to show that the measures were sensitive to training time. In a parallel study, Knoop and Welde (1973) evaluated deviations from reference/criterion values for the same flight parameters as above (Knoop, 1973) and a sum of 12 ‘absolute value of deviations from criterion values’ at four points in the maneuver was used as an index of pilot performance. This index was compared to a single instructor pilot’s subjective evaluations on a four-point scale for 47 lazy-8 maneuvers. The performance index measure accounted for 67% of the variance of the IP's own subjective evaluations. Evaluation by IPs as criteria for objective metrics was also used by Carter (1977), who reported high correlations between derived objective measures and subjective IP ratings.

Connelly et al. (1974) performed a study which was concerned with developing candidate measures for pilot performance evaluation in the T-37. They studied lazy-8, approach and landing, barrel roll, split-S, and cloverleaf maneuvers in an airplane (T-37), recording ALT, IAS, VS, Roll, Pitch, HDG, acceleration, control inputs (elevator, ailerons, rudder, and throttle) and input forces (elevator, ailerons), roll, pitch, and yaw rates. These measures were discussed and formulated in terms of continuous differences from a reference trajectory (e.g., RMSE, MAE), where this trajectory could be empirically derived, and tolerance deviation measures were computed from either external criterion or SDs from empirical data. Linear combinations of weighted errors (cf Knoop & Welde, 1973); vector combination of error terms (allowing simultaneous comparison of all error terms) were discussed.

It is evident that complex tasks—such as flying—involving multiple dimensions (exemplified by flight parameters in our discussion) yield a vast number of measures. Much effort has therefore been expended to data reduction and development of performance indices that combine many of the objective metrics in some meaningful and informative way. For example, a study by Hills and Eddowes (1974) yielded a total of 2436 measures per subject, which were broken down over the 10 flying tasks/segments. The effect of pilot experience on performance in one-, two, and three-dimensional tracking (roll, pitch and roll, and pitch, roll and yaw, respectively) while straight and level, climbing and descending turns, and in ILS approach was examined in a simulator. The parameters measured included ALT, IAS, VS, Roll, Pitch,

HDG, GS, LOC, TR, as well as control inputs (elevator, aileron, rudder, and throttle). In addition to means, standard deviations, and correlations, tracking measures (roll and pitch gain, phase, and cross-over, based on frequency analysis) The authors attempted to distinguish the three pilot experience groups based on the objective measures derived from the flight tasks. Because the authors used a tracking task with an applied sum of sinusoids error generator, traditional tracking measures could be employed (i.e., Bode plot measures/describing functions). Two experiments were performed with the first being a simpler version of the second. One-way ANOVAs were used to determine the ability of each measure to independently predict group membership. Only a little over 17 % (420) of the variables were found to be statistically significant ($p < 0.05$l) in separating groups. Standard deviations of variables produced the highest proportion of significant variables (32%), followed by tracking (20%), means (18%) and correlations (11%).

Unfortunately, Hills and Eddowes (1974) did not summarize the numeric values of metrics between groups. In general it appears that SDs decrease with pilot group skill and the high-frequency crossover point from the describing function analysis increases with pilot group skill for those parameters that show a difference between groups. The discriminant function from the first experiment was used to classify performance in the second experiment and the classification was significant at predicting group membership ($p<0.005$). However, the misclassification proportion using this discriminant function was still 33%. Because of the results of the cross-validation, the authors concluded that the idea of using a combination of measures from a large number of aircraft state variables to predict pilot performance was not viable.

Also Vreuls et al. (1975) sought to limit the amount of measures to those that were sensitive to training progress and utilize them in an automated IFR training simulator. Presuming that performance would improve with training, early training and late training were used as independent variables and the flight tasks included straight and level, standard rate climbs and descents, level turns, climbing and descending turns, plus control inputs (elevator, aileron, rudder, and throttle). The flight parameters measured were AOA, Pitch, VS, ALT, IAS, Roll, TR, HDG, sideslip (ball); from these, minima, maxima, MAE, SD, RMSE, TOT, RNG, zero crossings, autocovariance, and frequency analysis measures were derived as secondary measures. These were further used to form a discriminant function that best predicted late training performance from early training performance with a minimum number of measures. Across the various flight maneuvers, the discriminant function on average contained 9 secondary measures. In addition, the authors found that control inputs were important in distinguishing training stages. Using these discriminant functions in automated feedback training scenarios reduced training time to set criteria by 34–40% compared to the original scoring algorithm that included analytically derived measures only.

McDowell (1978) used a number of objective measures to study the effect of pilot experience (beginner, intermediate, advanced) on flight performance in a simulator. Here, experience served as a handy variable to evaluate objective performance metrics, as it was presumed that more experienced pilots would perform better. The flying tasks were straight and level, turn to heading, vertical S and formation flight. Control inputs (elevator, aileron, throttle) were recorded and from these minima, maxima, moments (1-4), and relative power spectra were derived. Subjects from each experience group flew four maneuvers from an UPT syllabus from which the objective measures were generated to discriminate between the groups. Discriminability increased with maneuver difficulty and there were changes in pilots' control input power spectra with skill level. Aileron measures proved to be better than elevator and throttle measures in

generating significant differences between group performances. For the maneuvers that produced significant group differences, skilled pilots' power spectra were shifted to higher frequencies relative to less experiences pilots. Also Childs (1979) found that mean (n = 4) scores for maneuver performance increased with training day, indicating that the measure was sensitive to performance improvement

Swink et al. (1978) identified tasks and performance variables essential to effective operation and developed functional specifications of an airborne performance measurement system to quantify transfer of training from a C-5 simulator to aircraft. Sorties involving T/O, emergency procedures, and VOR and ILS approaches were simulated and IAS, ALT, VS, HDG, Position, Pitch, Roll, Yaw, control inputs (elevator, aileron, rudder, and throttle), CDI, and GS were measured. From these, RMSE were calculated. Tolerances of continuous flight variables were mentioned but the authors did not explicitly develop any secondary measures beyond RMSE.

Hennessy, Hockenberger, Barnebey, and Vreuls (1979) developed automated performance evaluation system for UH-1 helicopter simulator for climbing and descending turns, NDB, VOR and ILS approaches (including ILS backcourse), and holding maneuvers. The flight parameters recorded included ALT, IAS, Bank, TR, VS, HDG, Trim, NDB tracking, VOR tracking, and ILS (LOC & GS) tracking. Performance was measured on a scale from 1-6 for primary measures within three specified tolerance bands (analytically derived) within a segment. Kelly, Wooldridge, Hennessy, and Reed (1979) used pilot experience and an independent variable to evaluate measures specific to air combat maneuvering (i.e., time in gun range, gun kill success, offensive time etc.) and up to 67 flight variables (including ALT, IAS, HDG, and control inputs). From these, RMS, MAE, Mean, SD and RNG were calculated. Multiple regression and discriminant analysis reduced the set of 31 candidate measures to a set of 13. The linear discriminant function, made up of 13 secondary measures, accounted for 52% of the variance in the performance data. The discriminant function predicted skill group membership (high or low) with 92% accuracy. However, of the three levels of pilot experience only two skill levels were used as classification variables, making it difficult to judge the success or significance of the study. Wooldridge et al. (1982) reanalyzed these data (Kelly et al., 1979) using a different segmentation process and discriminant models. Control inputs were found to be an important component of these new discriminant models. Results varied but were not 'significantly' (according to the authors) different from the previous study.

Interest in practical applications for automated performance measures is exemplified by Semple, Cotton, and Sullivan (1981), who sought to develop automated pilot performance measures for simulators. The proposed measures included means, SD, RNG, maxima, minima, RMSE, MAE, TOT, time on target, zero or average-value crossings of time histories of data, autocorrelation, power spectral density function (bandwidth, peak power, low/high frequency power), Bode plots and describing functions of several parameters. However, no data on any particular flight parameters are reported. The general requirements of an automated performance measurement system are summarized within a broad discussion of aircrew training devices. Their general conclusion was that sophisticated tools needed to be developed that could weight various combinations of measures to generate performance scores that could be easily interpreted by instructor pilots. They noted the progress that had been made in this area (as above) but that there were still open questions regarding how best to select and combine measures.

Automated performance measures have also been used in research on various topics. Martin and Rinalducci (1983) studied simulator acuity (cue density and shade) in a simulator and the

effect of these on pilots' altitude maintenance performance. Also airspeed was manipulated in the simulated low-level military sortie. Altitude RMSEs were calculated. Performance was better in higher density simulated surroundings, higher contrast surroundings and at slower airspeeds. De Maio, Bell, and Brunderman (1985) examined the impact of visual cue quality (5 levels, low to high) and task difficulty (straight flight, turning flight) on pilot performance using objective measures of ALT, VS, and control inputs (elevator, aileron). Median ALT, median ALT RNG, and Critical Control Inputs (defined as control input that changes the sign of vertical acceleration) were calculated, and from these, Smoothness (ratio of critical control inputs to total number of control inputs) and Critical Error Rate (distance traveled from Critical Control Input to Vertical Acceleration sign change) were derived. These measures were indeed sensitive to changes in pilot performance across the experimental conditions. Median ALT and Smoothness were influenced by visual acuity of the simulated world, but not in a simple linear manner. ALT RNG and Critical Error Rate both increased with increasing flight task difficulty. The authors suggested that these measures could be used to distinguish between the perceptual and task difficulty components of flight control performance.

The above studies appear within a twelve-year period that can be considered as a first wave of research on automated and objective performance measure development. It is interesting that a lengthy gap preceded the next wave, or 'cluster' of research in the mid-1990s. The following studies also show evidence of usage of the measures merely as research tools rather than interest in further development and evaluation of objective pilot performance measures. Furthermore, the measures used tended to be relatively simple.

Sirevaag et al. (1993) used objective pilot performance measures to study communication modality (verbal–digital) and communication load (high–low) in a helicopter simulator, recording IAS, ALT, Position, Roll, Pitch, Sideslip, and control inputs, and calculating means, SD, and TOT for each. In addition to aircraft control, performance on secondary tasks was measured and subjective (NASA TLX) and physiological measures were taken. Control performance measures showed that pilot performance was best in the low load, verbal communication condition. However, performance on secondary tasks was better in the digital mode condition. The TLX ratings of workload were constant across conditions. Two of the physiological measures discriminated between levels of communication load, but no correlation of these measures with other measures were made.

Examples of simple metrics used to evaluate new aviation technologies and their impact on pilot performance include Reising, Ligget, Solz, and Hartsock's (1995) study of HUD format in simulated ILS approaches; RMSE measurements of ALT, IAS and CDI showed pilot performance was better (more accurate) using the pathway HUD than the standard HUD format. Fox, Merwin, Marsh, McConkie and Kramer (1996) studied instrument panel peripheral information during simulated basic flight maneuvers. RMSE were calculated for ALT, HDG, IAS, VS, and TR. They found that pilot performance was degraded when peripheral information on the instrument panel was removed. In another study done at the University of Illinois, Ververs and Wickens (1996) used MAE on ALT, IAS, and HDG, and determined that pilot performance was better (smaller MAE values) in a clear sky condition than a cloudy condition, indicating better extraction of aircraft pitch and roll information from outside the aircraft; either from the real horizon or the HUD. Response time to detect a target showed that reducing clutter and lowlighting non-essential flight information resulted in faster detection.

Svensson, Angelborg-Thanderz, Sjoberg, and Olsson (1997) examined the effect of information complexity (low to high, 16 levels) in a low level military sortie in a simulator, recording ALT, IAS, HDG, Pitch, and Roll data. From this data, a flight performance index was formed by combining subjective IP ratings of four aspects of a subject's flight: altitude precision, IAS precision, timing of a turn and bank performance in this turn. The study examined the effect of information complexity on pilot performance, as measured by the flight performance index, and pilot mental workload. There was a significant correlation between pilot performance and information handling ($r = 0.59$, $p < 0.001$). Objective performance measurements (altitude deviations) were also used to show how flight performance began to decline once a certain level of information complexity had been reached. The flight performance index (based on the subjective ratings of IPs) was correlated with the three measures of pilot mental workload used (NASA TLX, Bedford Rating Scale [BFRS], Subjective Workload Assessment Technique) with the NASA-TLX and BFRS producing significant negative correlations of –0.37 and –0.43 respectively. There were no correlations performed between *completely* objective performance measures and workload assessments or physiological measures.

Hughes and Takallu (2002) recorded 68 unspecified flight parameters, including flight control inputs to study SVS Terrain Display Models during simulated basic flight maneuvers and ILS approaches. In addition to subjective SA measures, the minima, maxima, range, RMSE, and SD of the flight parameters were calculated. Only subjective SA data were reported with further analysis using objective pilot performance data still to be carried out. In a parallel study, Takallu, Wong, and Uenking (2002) examined three different types of synthetic vision displays' impact on pilot performance in a simulator. The flying tasks involved straight and level, 180 ° turn, descent and climb. The main flight parameters recorded included ALT, IAS, HDG, Pitch, Roll, and CI (Elevator, Aileron, Rudder, Throttle); additional parameters brought the total to 62. Deviations from reference/criterion values, RMSE, SD, and TOT ratios were calculated and a normal of vector containing tolerance-normalized errors of IAS and HDG, total scanning area (area under normal vs. time graph) was derived from these. The RMSE and TOT measures showed that pilot performance was better when using the SVS display. The total scanning error measure showed 8 of the 16 pilots had improved performance (lower scores) when using the SVS display. The other pilots showed similar errors across all three displays.

In sum, since the early 1980's, the literature shows little evidence of interest in developing objective pilot performance measures. The more recent reports merely use relatively simple measures as tools among many other measures to examine various research questions, usually to evaluate new aviation technologies or their impact on pilot performance. One explanation for the waning interest in objective pilot performance measures that may be offered is based on their complexity: the more complex the metrics become, the farther they are removed from the actual processes they are representing (i.e., pilot performance) and the harder they will be to interpret (cf Vreuls & Obermayer, 1985). Also, the effort to produce metrics based on power spectra and discriminant functions is often substantial and must be weighed against the validity and diagnosticity of such measures.

An Inventory of Objective Pilot Performance Measures

Measures of central tendency: Mean, median, and mode

Measures of central tendency are used frequently in data analysis in all application areas. The purpose of these measures is simply to reduce—sometimes very large—data sets to a single

number, for example, mean. While such measures are absolutely necessary to make data manageable and understandable, it is also critical to consider all the aspects of data they obscure. In the literature, measures of central tendency have never been used alone; rather, they appear to be used as initial, exploratory measures of pilot performance data. It is also noteworthy that these measures by themselves do not afford any conclusions to be made about pilot performance.

Minima, maxima, and range

Minima and maxima of given data, and range they yield, complement the measures of central tendency. What is said about the measures above also applies to these measures, and indeed, in the literature minima, maxima, and range appear to have been used only in an exploratory manner.

Standard deviation (SD).

One of the most common objective metrics is the standard deviation (SD) of selected flight parameters. This metric describes the amount of variability around the mean of any series of values. In contrast to measures of central tendency, small SD in the case of piloting an aircraft is usually indicative of good performance. For example, Svensson, Angelborg-Thanderz, Sjoberg, and Olsson (1997) examined the effects on information complexity on pilot mental workload and pilot performance in a simulator and found that altitude deviations increased and correction of errors were delayed as a result of increased workload. Also Hills and Eddowes (1974) found that SD variables produced the highest proportion of statistically significant differences between experience groups (32%). It is important to note, however, that SD does not provide any information about possible error relative to given criteria.

Root mean square error (RMSE) and Mean Absolute Error (MAE)

RMSE is a widely used measure of tracking performance (e.g., Scallen, Hancock, & Duley, 1995). It can be used to reduce the tracking performance along a specified parameter value, or criterion (e.g., a given altitude, or VOR radial) during an entire segment of a flight into a single number. A low number typically indicates good performance. The RMSE is calculated by squaring individual errors (sampled at certain rate), adding them together, dividing this sum by their total number, and then taking a square root of this quantity. The RMSE hence summarizes the overall error. In a study by Reising, Ligget, Solz, and Hartsock (1995) the RMSE measurements were successfully used to reveal pilot performance differences when using two different types of head-up displays (HUDs). Subjective feedback from the pilots corroborated the results. Ververs and Wickens (1996) measured mean absolute errors (MAE) of altitude, heading and airspeed, along with reaction time to a stimulus event to investigate the effect of clutter and low lighting on HUD assisted flight in a high fidelity flight simulation environment. Tracking error was used to determine that pilot performance was better in a clear sky condition than a cloudy condition, indicating better extraction of aircraft pitch and roll information from outside the aircraft; either from real horizon or the HUD display. Also Stave (1977) measured performance in a simulated helicopter flight task by RMSE from navigational course and an angular deviation from the Instrument Landing System (ILS) approach.

The RMSE has a number of shortcomings, however. It does not contain information about the direction of deviations or the frequency of deviations from the criterion. The latter is particularly important dimension of tracking performance, as it would allow for detection of high velocity error in tracking while the position error (measured by the RMSE) might be minimized

(Wickens & Holland, 2000). To overcome these limitations, additional measures of tracking performance are available.

Number of deviations (ND)

The number of deviations outside tolerance (ND) is a measure that tallies the occurrences of the aircraft straying outside predetermined tolerances (Reynolds, Purvis, & Marshak, 1990). This is essentially a measure of velocity error in tracking and it complements the RMSE, which contains the error magnitude information. A low number typically indicates good performance. A low value, however, can also be obtained if the pilot makes few aberrations outside the tolerances but stays there for a substantial proportion of the flight segment of flight. The ND measure must hence be considered together with the total time spent outside tolerance in a given segment.

Time outside tolerance (TD)

The cumulative time the aircraft spends outside a given tolerance provides an indication of tracking performance beyond the RMSE and number of deviations. This measure is computed simply by summing the time the pilot spends outside of a given tolerance and divided by the total time in the segment (i.e., percent time outside tolerance). A small number indicates good performance. Sirevaag et al. (1993) took aircraft control measures from a helicopter simulator in a study investigating the effects of verbal and digital communication loads on pilot performance. The measurement of time above an altitude criterion produced significant differences between experimental task conditions.

Mean time to exceed tolerance (MTE)

Rantanen and Talleur (2001) developed a metric labeled mean time to exceed tolerance (MTE). The MTE is computed from the rate of change between successive data points and the aircraft's position relative to a given tolerance. Based on this information, the measure extrapolates the time the aircraft will remain within the tolerance region, as opposed to the number of deviations and time outside tolerance measures described above. Because this measure could potentially yield very large values, it was truncated at 60 s. Thus, if the pilot was 60 s or more from exceeding tolerance throughout the flight segment, his or her performance was considered good. In subsequent analysis, the MTE on ILS localizer tracking showed a significant difference between pilots who passed an IPC flight and those who failed, by flight instructor evaluation (Rantanen & Talleur, 2001).

Critical control input

Other objective metrics include Critical Control Input, which is defined as a pilot input that changed or led to a change from positive vertical acceleration to negative vertical acceleration (or other flight parameter) or vice versa (De Maio, Bell, & Brunderman. 1985). A non-critical control input did not cause the vertical acceleration to change from positive to negative or vice versa. De Maio, Bell, and Brunderman (1985) hypothesized that “efficient” control would be characterized by a relatively large proportion of critical control inputs indicating that pilots were canceling small errors in altitude frequently.

Smoothness

Another metric, “Smoothness,” was defined as the proportion of critical control inputs from the total number of inputs (critical + non-critical). The critical error rate is the horizontal distance traveled from critical control input to vertical acceleration sign change divided by the time from

critical control input to vertical acceleration. This metric was designed to measure the effectiveness of a critical control input; low values for critical error rate would indicate a slow accumulation of error following the pilot control input. De Maio, Bell, & Brunderman (1985) found that that smoothness and critical error rate were affected by flight task difficulty (straight vs. turning, both while level).

Moments

The n^{th} moment of a series of data is the summation of individual series values raised to the n^{th} power and then divided by the number of sample points. Thus, the first moment is simply the average of a series of data. Average values have been commonly used as measures of pilot performance; for example, Hills and Eddowes (1974), McDowell (1978) and Sirevaag et al. (1993). However their use may be limited in certain circumstances given the way averages can mask important patterns and deviations in performance. The use of higher order moments appears to have been limited to De Maio and Eddowes (1978), where the aileron second moment showed differences between pilot experience groups in the study.

Power spectra

While frequency analysis has been identified as a useful tool to aid in performance measurement (Semple, et al., 1981), actual implementations of such frequency-based measures have been limited. Hills and Eddowes (1974) and Vreuls et al. (1975) used measures based on a manual tracking approach. Given a known disturbance function that was applied to the simulator aircraft, the researchers were able to use control inputs and derive describing functions and Bode plots of pilot performance. From this, measures such as cross-over power and high and low frequency gains were generated. Hills and Eddowes used these measures as part of a battery of over 2000 measures to derive discriminant functions that attempted to classify pilot experience groups (beginning, intermediate and advanced). Vreuls et al. (1975) performed a similar analysis using discriminant functions, however the exact nature of the frequency-based measures that were included is not clear. De Maio and Eddowes' (1978) did not use a manual tracking approach by contrast and instead used several measures to quantify pilots' control input power spectra. Several "digital filter" type measures were developed that estimated the relative power spectra below various frequency cut-off points. The 1/8Hz cut-off filter measure from the aileron control inputs produced the greatest separation between pilot experience groups of these filter metrics. In this case the more skilled pilots had their power spectra shifted towards higher frequencies.

Summary

Development of objective pilot performance metrics is inextricably linked to methods of their validation. The only alternative method of pilot performance evaluation is expert judgment, that is, evaluation by an instructor pilot, and hence subjective IP evaluations are the only method for validation of objective metrics. Unfortunately, poor inter-rater (and in some cases, also intra-rater) reliability (cf Knoop, 1973; Knoop & Welde, 1973) severely encumber efforts to develop and use objective pilot performance measures. Another aspect of objective metric development is the typically large number of measures and the need to somehow reduce these data to a manageable set that can be interpreted in an effective and meaningful way. These efforts have seldom been thoroughly successful (cf Hills & Eddowes, 1974). Such difficulties associated with data reduction and performance indexing might in part explain the apparent lack of interest in

objective performance metrics after the 1970s and the initial push for objective pilot performance evaluation by the Air Force.

The present research offered an opportunity to revisit the topic of objective pilot performance measurement. We had a several advantages on our side compared to many of the previous efforts, including established inter-rater reliability (please see Vol. 1 of this report), a revised IP scoring sheet to facilitate effective segmentation of the IPC flights and direct comparison of IP evaluations and objective measures, and availability of flight data from all three devices, that is, from an airplane, a Frasca FTD, and a PCATD. Unfortunately, however, control inputs could not be recorded by the FDR employed in this study and hence not included in the analyses. Despite the lack of control input data, this research represents—to the best of our knowledge—the first systematic evaluation of objective pilot performance measurement in the general aviation domain.

Method

Data Collection

The experimental design and data collection methods have been detailed in Volume 1 of this report and will not be repeated here. Data from which objective pilot performance measures were derived were collected by a flight data recorder (FDR) on-board the Beechcraft BE-C23 Sundowner aircraft used in this project. The FDR recorded 11 flight parameters, including aircraft position using Wide Area Augmentation System (WAAS) corrected Global Positioning System (GPS) receivers, altitude, pitch, roll, yaw, magnetic heading, vertical speed, and airspeed, as well as VHF Omnidirectional Range receiver and Localizer (VOR/LOC) and glideslope (GS) indications. For a detailed technical description of the FDR, see Rantanen and Talleur (2001) and Lendrum et al. (2000). Note that no control input data were available. Data streams form the Frasca FTDs and Elite PCATDs were also recorded. Both devices recorded the same flight parameters as the BE-C23 aircraft. The aircraft and FTD recorded data at a 1Hz rate, while the PCATD data, recorded at approximately 50Hz rate, was decimated to 1Hz prior to data analysis. All data from the three devices were brought to a standard format for preprocessing by two custom-written computer programs. These data processing steps are described in detail by Rantanen and Talleur (2001). A separate Matlab program was created to compute the time series metrics, as described below.

Development of Time Series Based Pilot Performance Metrics

The metrics described in this section were developed to supplement existing performance metrics (Rantanen & Talleur, 2001) with analyses that examine underlying patterns in the pilot-generated time series of data (see Figure 1 for an example). The new metrics utilize spectral (Fourier) and autocorrelation analyses and will be described in detail below. Two guiding hypotheses were used to develop these metrics: First, there may be a difference in the frequency of observed flight characteristics (based on pilot's control inputs). Better pilots are expected to exhibit a larger range of frequencies of aircraft control than less able pilots, who may only control the aircraft with low frequency control inputs. Using Fourier analysis, a time series of data can be decomposed into spectral or frequency components. This decomposition allows an explicit representation of the underlying frequencies occurring in the time series. The second hypothesis is that more skillful pilots will exhibit a better awareness of the airplane's constantly changing

state and be able to predict what control inputs will be required to maneuver the airplane to the future desired state. This may be manifested in the degree of correlation between flight parameter values in a time series. That is, better pilots may exhibit a greater correlation between a previous time point and the present time point than less skilled pilots who may exhibit a greater randomness of control on flight parameter values. By taking the autocorrelation of a time series for a particular observed flight variable, the degree of randomness between successive measurements can be investigated. Derivation of specific metrics from time series data is described next.

Fourier Analysis Metrics

To examine the periodic components of a time series, Fourier analysis was used. Taking the Fourier transform of time series data, Y_k, gives the spectral decomposition:

$$Y_k = \frac{1}{N}\sum_{j=1}^{N}\tilde{Y}_j e^{\frac{2\pi i}{N}(k-1)(j-1)}$$

where the Fourier coefficients $\tilde{Y}_j$ are given by

$$\tilde{Y}_j = \sum_{k=1}^{N} Y_k e^{\frac{-2\pi i}{N}(k-1)(j-1)}$$

and where N is the number of time series data points and $i = \sqrt{-1}$.

The original time series is then expressed as a weighted sum over all frequencies contained in the Fourier transform. The weights, $\frac{\left|\tilde{Y}_j\right|^2}{N}$, represent the contribution a particular frequency makes to the original time series and are termed *power spectral densities* (PSD). The $\frac{\left|\tilde{Y}_j\right|^2}{N}$ can be plotted against frequency, $f = \frac{j}{N}$ (for a 1Hz sampling rate), in a periodogram.

We hypothesized that a good pilot's time series may contain a greater range of frequencies that contribute significantly to the time series, that is, a greater proportion of components that have a large PSD, compared to a poor pilot's time series (see Figure 2). The metrics that were developed with Fourier methods are used to quantify both the range and magnitude of these significant frequency components.

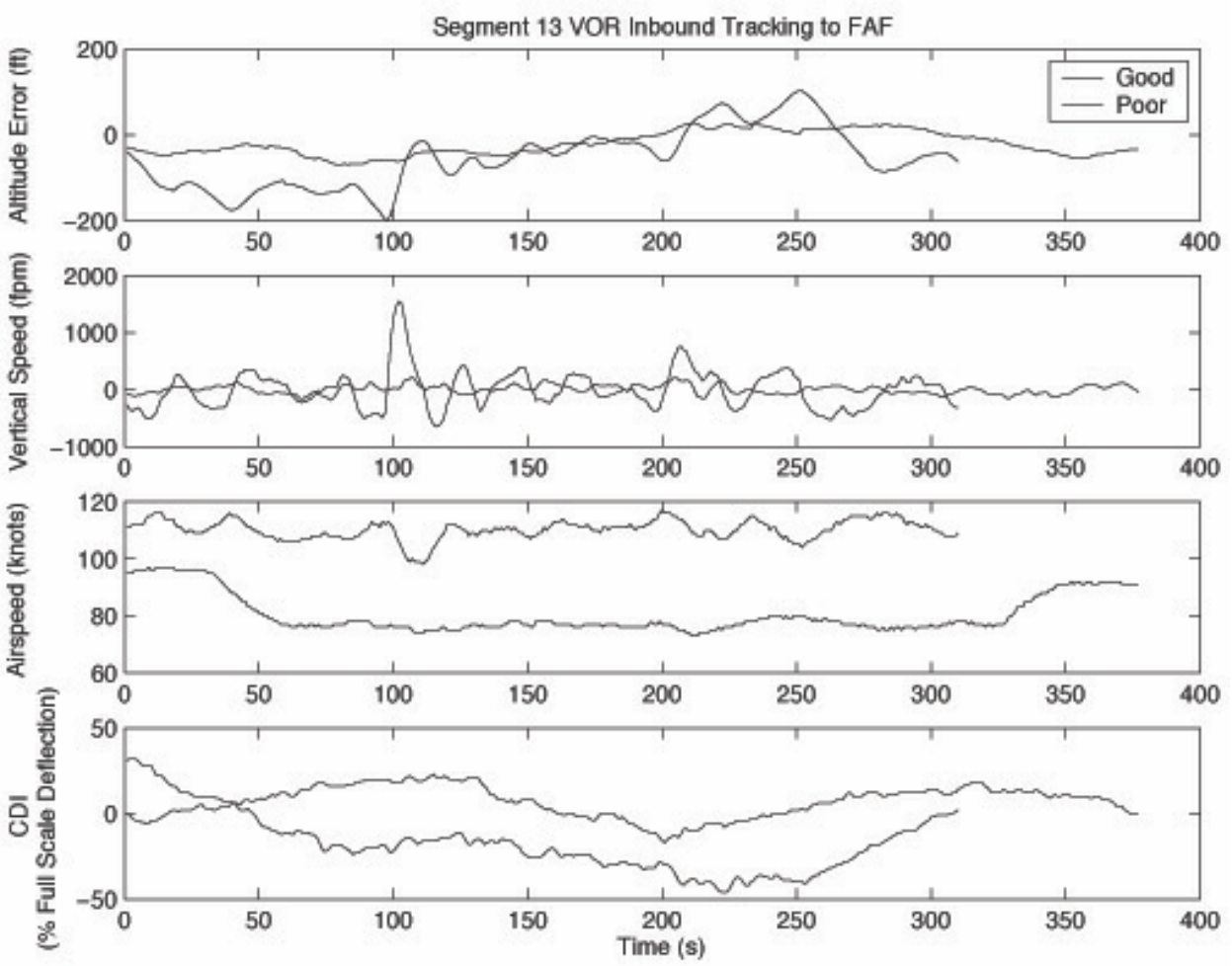

Figure 1. Time series of four flight parameters from a VOR approach segment in an IPC flight. The data are from two pilots, one who passed the IPC (good pilot) and one who failed (poor pilot). The poor pilot exhibits much larger variability in the depicted parameters, and since Fourier analysis provides a sensitive method to quantify and examine such variability, Fourier-based metrics were pursued for the purposes of this project.

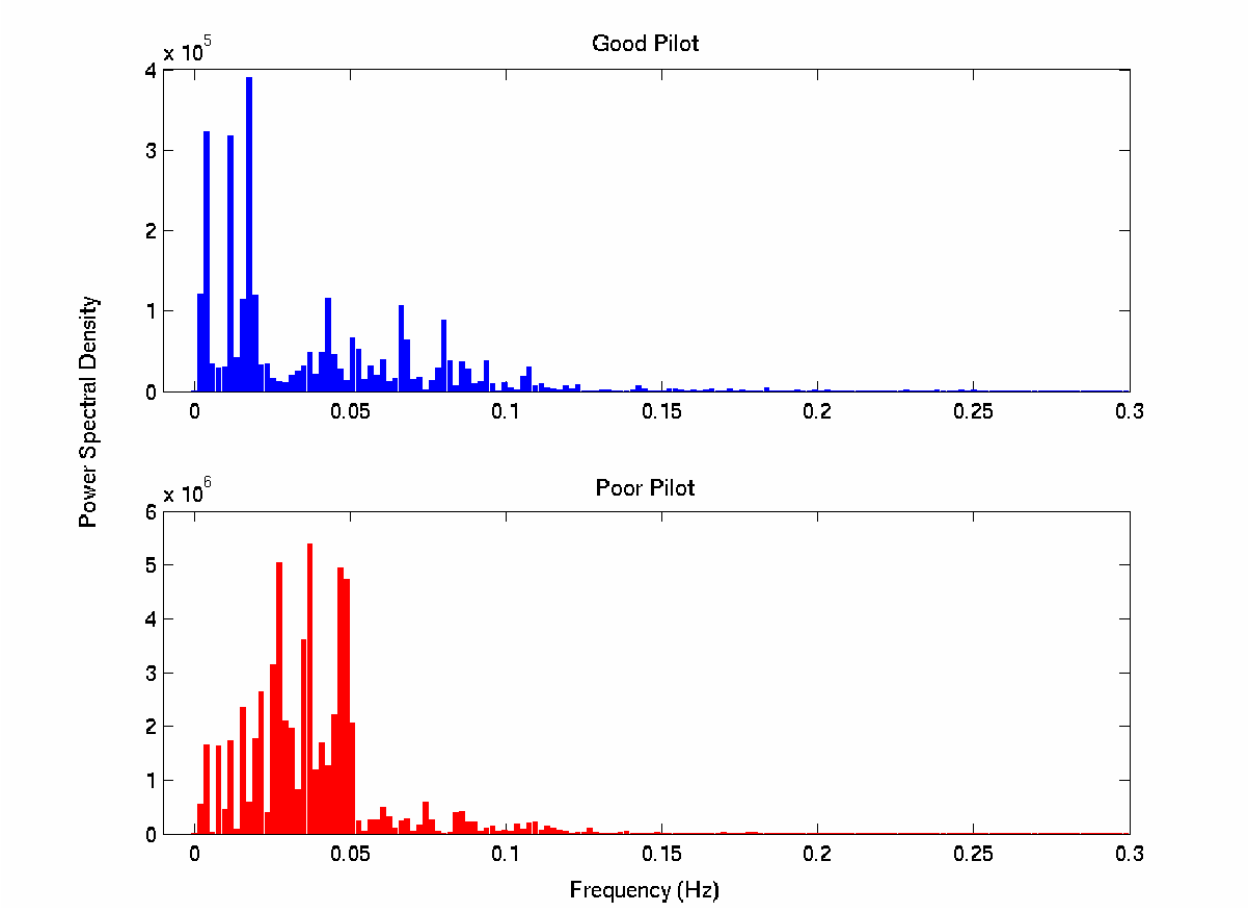

Figure 2. Comparison of a good pilot's and a poor pilot's power spectral density plots, known as *periodograms*, lend support to the initial hypothesis that a good pilot's time series may contain a greater range of frequencies that contribute significantly to the time series, that is, a greater proportion of components that have a large PSD, compared to a poor pilot's time series. Note, too, that scales of the y-axes in the above plot are different by an order of magnitude, the good pilot exhibiting much smaller PSDs than the poor pilot.

In determining what spectral components of the Fourier decomposition were significant, a critical value v_c was set. Components with PSD greater than v_c were counted and used in the subsequent metrics described below. Setting v_c involves some difficulties, however. Because the data ranges of the time series vary greatly between flight parameters (altitude and airspeed for example) and individual pilots, PSD magnitudes in the Fourier decomposition will also vary greatly between parameters and pilots. Thus setting a single critical value to be used across all pilots' flight parameters will not achieve the desired level of sensitivity. Therefore, a relative v_c was set to a fraction of the mean or maximum value of the spectral components. This approach will also allow for manipulation of v_c in order to find the value that produces maximum sensitivity in distinguishing good and poor pilots.

Seven Fourier-analysis based metrics were developed; (1) mean and (2) standard deviation of the spectral components $\frac{|\tilde{Y}_j|^2}{N}$, (3) the number of spectral components that are greater in magnitude than a critical value v_c , (4) the mean and (5) standard deviation of spectral components greater than v_c , and (6) the mean frequency and (7) standard deviation of the frequencies of spectral components with magnitude greater than v_c (see also Table 1).

Autocorrelation Metrics

The autocorrelation coefficient (r_h) gives a measure of the correlation between data points Y_k and Y_{k+h} of the time series $Y = \{Y_1, Y_2,\ldots, Y_N\}$ and is given by:

$$r_h = \frac{\sum_{k=1}^{N-h}\left(Y_k - \bar{Y}\right)\left(Y_{k+h} - \bar{Y}\right)}{\sum_{k=1}^{N}\left(Y_k - \bar{Y}\right)^2}$$

where

$$\bar{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k$$

is the mean of time series data and $-1 \le r_h \le 1$

A plot of r_h versus lag, h, is termed a *correlogram*; $r_0 = 1$ by definition. The autocorrelation coefficient gives a measure of how well a subsequent measurement can be predicted from a previous value in the time series. Values of r_h close to zero indicate little correlation between data points and values close to –1 or 1 indicate a strong negative or positive correlation respectively between data points. The time series of flight parameter values from a pilot who is aware of the state of the aircraft and can predict its future state may generate autocorrelations that are greater than those from a pilot who is not as aware. Since autocorrelations at large lag h tend towards zero, a useful measure that may be indicative of pilot performance is how quickly the series autocorrelations tend to zero.

We hypothesized that time series of less skillful pilots would produce autocorrelations that decay more quickly to zero than those of more skillful pilots (see Figure 3). Consequently, two specific metrics were developed: (1) To quantify the decay of autocorrelation coefficients, the slope of the first 10 autocorrelation coefficients (from lag = 0 to lag = 9) was determined by regression analysis, and (2) the sum of squares error of the fitted regression line was also included as a second autocorrelation based metric.

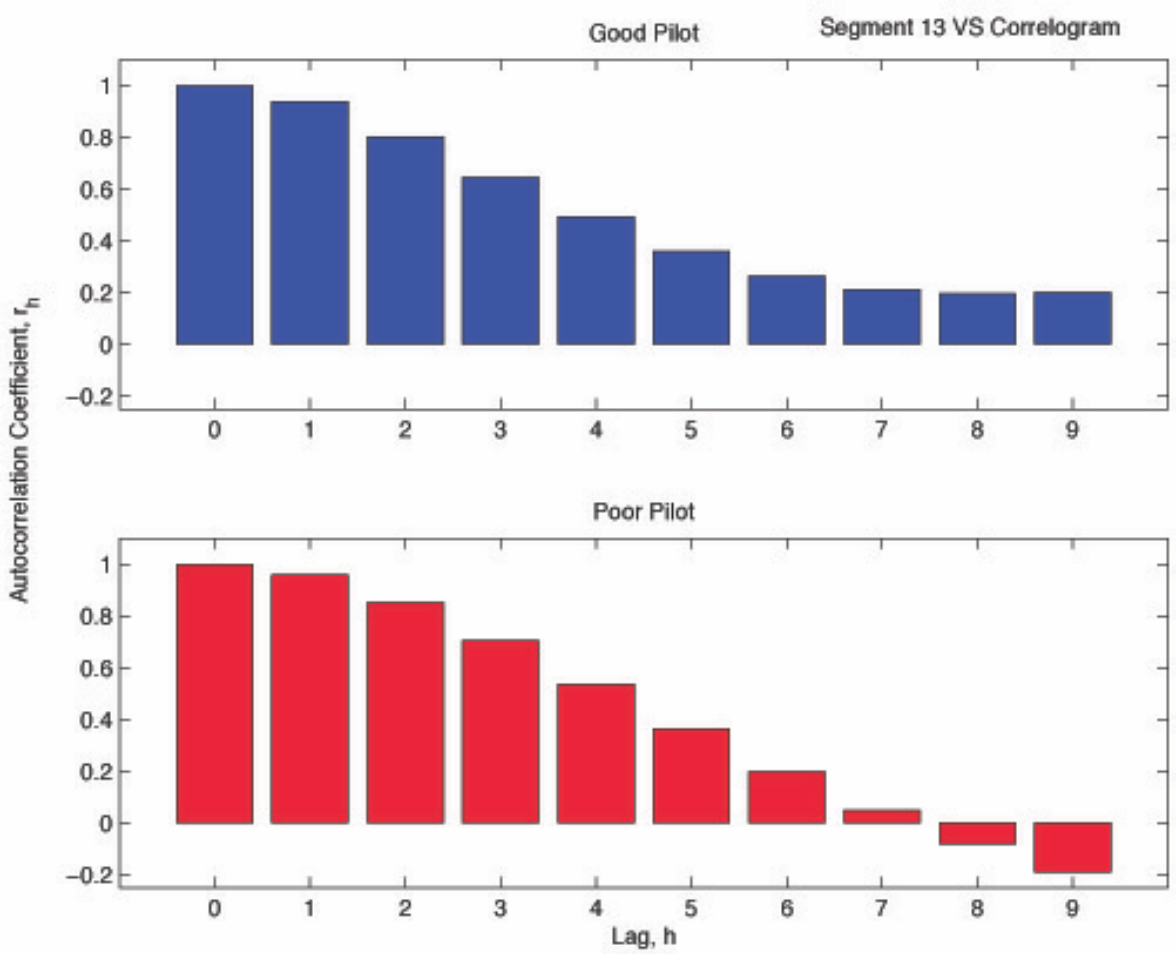

Figure 3. Correlograms of good and poor pilots' vertical speed from the same segment (cf the second time series plot in Figure 1). Note that the good pilot's correlation coefficient decays from 1 much slower than the poor pilots (i.e., it has a shallower negative slope).

The objective pilot performance measures used in this project are summarized in Table 1. Development of the time series based metrics has been described above. Some of the static metrics (RMSE and SD) are well known and have been used extensively in the past. Development of the other static metrics (i.e., ND, TD, and MTE) has been described in Rantanen and Talleur (2001).

Table 1
The objective pilot performance metrics used in the project.

	Metric	Description
1.	RMSE	Root mean square error
2.	SD	Standard deviation
3.	ND	Number of deviations outside tolerances
4.	TD	Time outside tolerance range
5.	MTE	Mean time to exceed tolerance values
6.	ACS	Slope of the autocorrelation function regression (first 10 data points, lags 0 to 9); quantifies how quickly autocorrelations tend to zero.
7.	MSC	Mean of the $\frac{\left\|\tilde{Y}_j\right\|^2}{N}$, the normalized squared magnitude of the spectral components incorporates the magnitude of deviations in the time series.
8.	DSC	SD of the $\frac{\left\|\tilde{Y}_j\right\|^2}{N}$; quantifies the magnitude spread of the spectral components.
9	NCGC	Number of $\frac{\left\|\tilde{Y}_j\right\|^2}{N} > v_c$.; number of spectral components greater than a criterion value
10.	MCGC	Mean of the $\frac{\left\|\tilde{Y}_j\right\|^2}{N} > v_c$; mean magnitude of the spectral components greater than a criterion value.
11.	DCGC	SD of the $\frac{\left\|\tilde{Y}_j\right\|^2}{N} > v_c$; magnitude spread of spectral components greater than a criterion value
12.	FMCG	Mean frequency of the $\frac{\left\|\tilde{Y}_j\right\|^2}{N} > v_c$; mean frequency of spectral components greater than a criterion value
13.	FDCG	SD of frequencies of the $\frac{\left\|\tilde{Y}_j\right\|^2}{N} > v_c$; the frequency spread of the spectral components
14.	MEDF	Median frequency of the power spectrum
15.	LPF1	Low pass filter, 0.005 Hz cutoff frequency
16.	LPF2	Low pass filter, 0.01 Hz cutoff frequency
17.	LPF3	Low pass filter, 0.05 Hz cutoff frequency
18.	LPF4	Low pass filter, 0.10 Hz cutoff frequency

Results

Data Reduction and Analyses

There were a total of 18 objective pilot performance metrics used in this project (see Table 1), 5 of which may be called static (RMSE, SD, ND, TD, and MTE), and 13 metrics derived from time series of the data and that will be referred to as dynamic. One of the latter metrics was based on autocorrelation and the remaining 12 on Fourier transforms of the time series. All these performance metrics were derived from 9 different flight parameters: (1) altitude (ALT), (2) heading (HDG), (3) airspeed (IAS), (4) ball (BAL), (5) roll (ROL), (6) pitch (PIT), (7) vertical speed (VS), (8) course deviation indicator (CDI), and (9) glide slope indicator (GSI). Furthermore, each IPC flight was divided into 14 separate segments (see Table 2 for segment descriptions). Hence, there was initially a total of 9 (flight parameters) x 18 (metrics) x 14 (segments) = 2268 objective performance measures for each subject and each IPC flight. Such large number of dependent variables was clearly too much for meaningful analyses; consequently, several steps were taken to reduce the number of variables to a manageable level.

Table 2.
IPC flight segmentation

1. VOR 36 Course Intercept
2. VOR 36 Outbound Tracking
3. VOR 36 Procedure Turn
4. VOR 36 Inbound Tracking to FAF
5. VOR 36 Final Approach Segment to MAP
6. Holding Pattern Entry
7. Left 360˚ Steep Turn
8. Right 360˚ Steep Turn
9. ILS 6 Intercept (RV to FAF)
10. ILS 6 Inbound Tracking to FAF
11. ILS 6 Glideslope Tracking to DH
12. Partial panel VOR Approach Intercept
13. Partial panel VOR Approach Inbound Tracking to FAF
14. Partial panel VOR Final Approach Segment to MAP

Elimination of meaningless variables.

Many of the measures were simply nonsensical; for example, metrics related to altitude from segments where altitude was not constant, or GSI-related metrics from segments not involving glide slope tracking are meaningless. Hence, the first step in the data reduction process was to select variables that would provide for meaningful information about pilots' performance in the given segments. This was done by the expert judgment and consensus of the research team. The results are depicted in Table 3 below.

Table 3

Metrics relevant to each segment; the grayed-out measures were eliminated from analyses.

		Segment													
Parameter	Metric	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Alt	ACS														
Alt	FT														
Alt	SD														
Alt	RMSE														
Alt	ND														
Alt	TD														
Alt	MTE														
VS	ACS														
VS	FT														
VS	SD														
VS	RMSE														
VS	ND														
VS	TD														
VS	MTE														
IAS	ACS														
IAS	FT														
IAS	SD														
IAS	RMSE														
IAS	ND														
IAS	TD														
IAS	MTE														
HDG	ACS														
HDG	FT 1:9														
HDG	SD														
HDG	RMSE														
HDG	ND														
HDG	TD														
HDG	MTE														
Pitch	ACS														
Pitch	FT														
Pitch	SD														
Pitch	RMSE														
Pitch	ND														
Pitch	TD														
Pitch	MTE														
Roll	ACS														
Roll	FT														
Roll	SD														
Roll	RMSE														
Roll	ND														
Roll	TD														
Roll	MTE														
Ball	ACS														
Ball	FT														
Ball	SD														
Ball	RMSE														
Ball	ND														
Ball	TD														
Ball	MTE														
CDI	ACS														
CDI	FT														
CDI	SD														
CDI	RMSE														
CDI	ND														
CDI	TD														
CDI	MTE														
GSI	ACS														
GSI	FT														
GSI	SD														
GSI	RMSE														
GSI	ND														
GSI	TD														
GSI	MTE														

Univariate ANOVAs

Data were analyzed by both the IP evaluation outcome (pass and fail groups) and by the device (airplane, Frasca FTD, PCATD) groups. Univariate ANOVAs were performed for all objective pilot performance metrics that had an IP evaluation on the corresponding element as well as for all metrics by device group. The significance level (or, rather, the *F*-value) was used as a criterion for selecting metrics for further analysis; the cutoff *p* –value was set at .05.

Correlation between variables.

The next step in the data reduction process was to calculate correlations between variables. All variable pairs with a correlation of .7—an arbitrarily selected threshold value—or above (at $p < .001$) were examined separately. Selection between highly correlated variables by the above threshold was done according to the following criteria:

1. In general, a variable that is simpler is given preference over a more complex variable. For example, while the dynamic Fourier transform-based metrics may well describe the general variability of a time series quite well, it would seem prudent to use a metric that explicitly measures variability (e.g., SD, RMSE) and is simpler to interpret.
2. When given the choice between MSC and MCGC metrics, choose MSC. The MCGC metrics are based on the criterion value that was arbitrarily set. By using the MSC metrics, we do not have to invoke a criterion value to derive the metrics. In essence, using the MSC metrics is simpler.
3. When given a choice between mean and standard deviation of spectral components, choose standard deviation. The key difference in our hypothesis on pilot control between poor and good pilots is that the more skilled pilots will exhibit a greater range of aircraft control frequencies. While the mean measure will contain this information, it seems that using the SD measure will be a more explicit way of testing our hypothesis.
4. Of the metrics of the percentage of the total Power Spectral Density of the time series below a certain frequency cutoff (LPF), choose one with lower cutoff.
5. Given a choice between the above and median of the power spectrum (MEDF), choose MEDF because it is more intuitive (the median frequency of the total Power Spectral Density, as opposed to the percentage of the PSD below a certain frequency).
6. When given the choice between the above (3 and 4) metrics and mean and SD of frequencies, choose the percentage of power spectrum measures The two groups of metrics are basically trying to quantify the same thing (spread of frequencies in the spectral distribution) but the latter metrics are slightly more robust in the sense that there is a long history of using those type of measures in frequency analysis in other fields.
7. Mean and SD of frequencies should be chosen over MCGC and over NCGC. FMCG and FDCG should be more sensitive to performance differences because they also contain frequency information of the spectral components greater than the criterion value.
8. Choose ACS over LPF, as LPF is highly correlated with a number of other metrics in general.

In the following sections, the results are presented by the analysis, first by IP evaluation outcome to establish the reliability of the objective metrics in comparison to IP judgment, and

then by device group, to examine the potential impact the differences between the devices might have had on pilot performance.

Evaluation of Objective Performance Metrics by Pass/Fail Groups

We shall first evaluate the objective pilot performance metrics as far as they correspond to IP evaluations as well as examine the participating pilots' performance in light of the metrics. Note that in the following analyses IPs provided scores on each element in the segment, for example, altitude was scored separately from airspeed, etc. In the following, including Figures 4-18, only those results that were statistically significant, ($p < .05$) for the pass/fail comparison are reported.

RMSE and SD

A good pilot, one who passes a segment element (variable), should have smaller RMSE and SD values than the pilot who fails on the segment element. This is necessarily so since the magnitude of error for a variable will be less for the good pilot. Although SD and RMSE may provide redundant information in most cases, in other cases, such as ALT in segment 13 (S13) for example, which is significant for SD but not RMSE, the RMSE variable did not adequately capture differences between good and poor pilot performance (Figures 4 and 5). The reason for this may be that a good pilot may produce the same RMSE with more small deviations around the criterion ALT as the poor pilot does with just a few large deviations outside the ALT criterion. RMSE is not as sensitive to the magnitude of the error as is SD. In this example, the SD differed for the two pilots to the point where a significant difference was found.

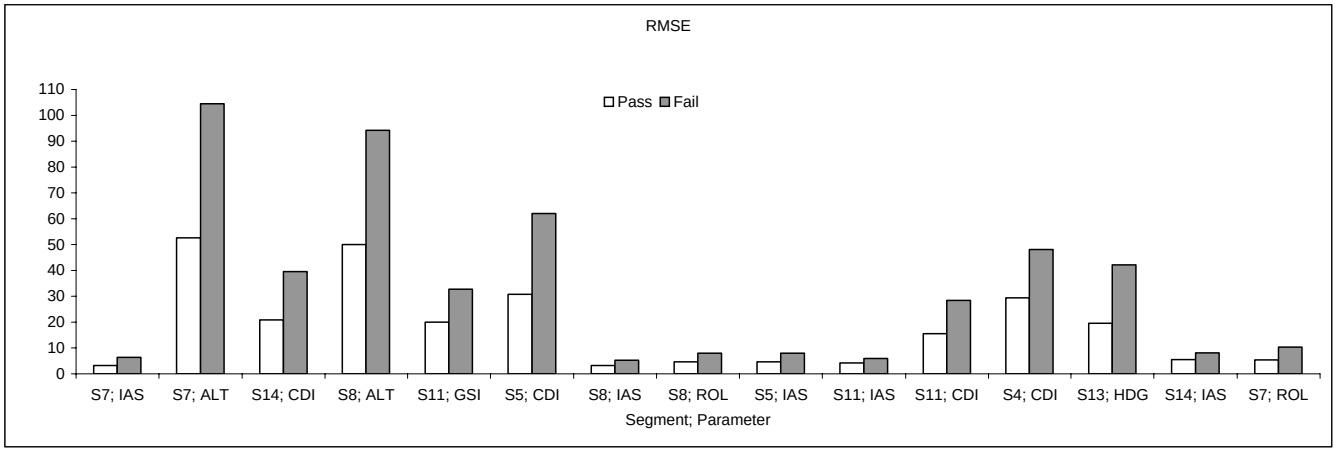

Figure 4. RMSE measures showed significant ($p < .05$) differences between pass and fail groups for many flight parameters in many segments, making it one of the best metrics to evaluate pilot performance by. In all cases, pilots who failed the element in the IPC flight exhibit substantially larger RMSE values.

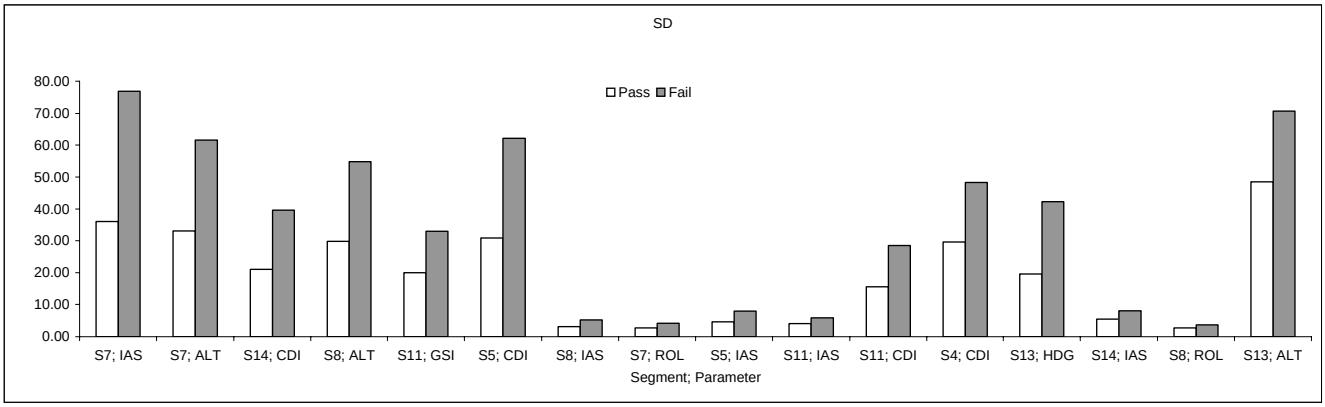

Figure 5. SD measures are also consistent across segments and flight parameters. All of the above differences between pass and fail groups (for the element in the segment) were significant, $p < .05$.

ND and TD

These metrics, while less sensitive than SD and RMSE, nevertheless differentiated between pass and fail groups by a number of flight parameters and segments. In particular, TD and ND appear well suited for use in segments 7 and 8 (steep turns) due to the nature of the maneuver. However, the time period of the maneuver may be too small for SD or RMSE to adequately capture differences at only 1 sample per sec. The TD and ND data show exactly what we would expect for the variables found to be different between pass/fail performance; failed pilots exhibiting both higher number and duration of tolerance exceedances (Figures 6 and 7).

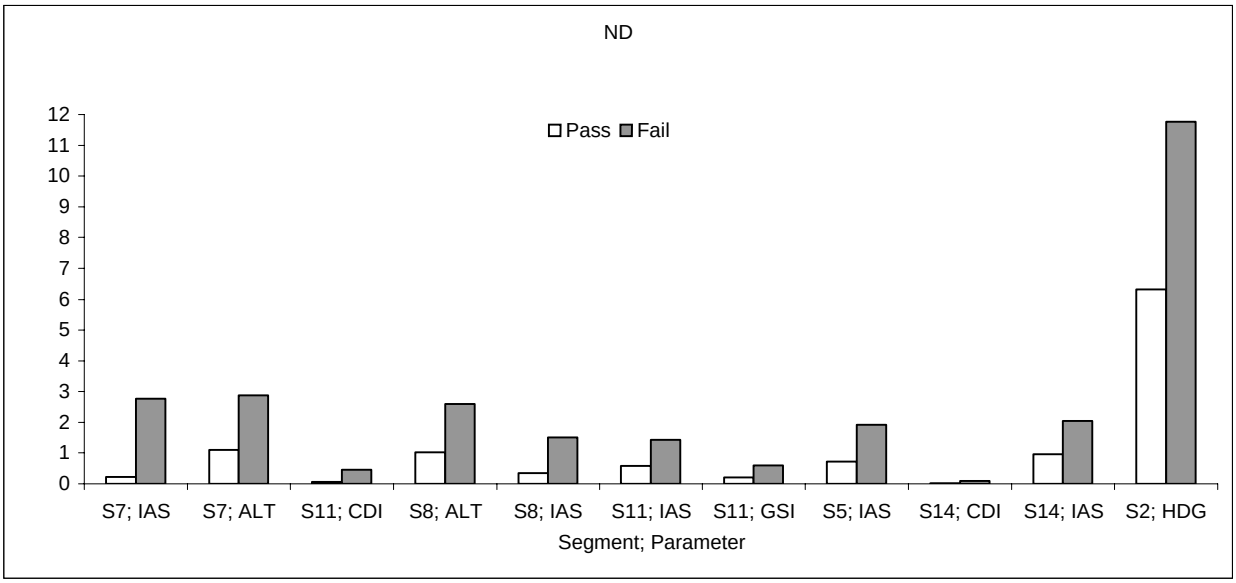

Figure 6. Failed pilots had consistently and significantly ($p < .05$) higher number of tolerance exceedances than pilot who got a passing score for the element.

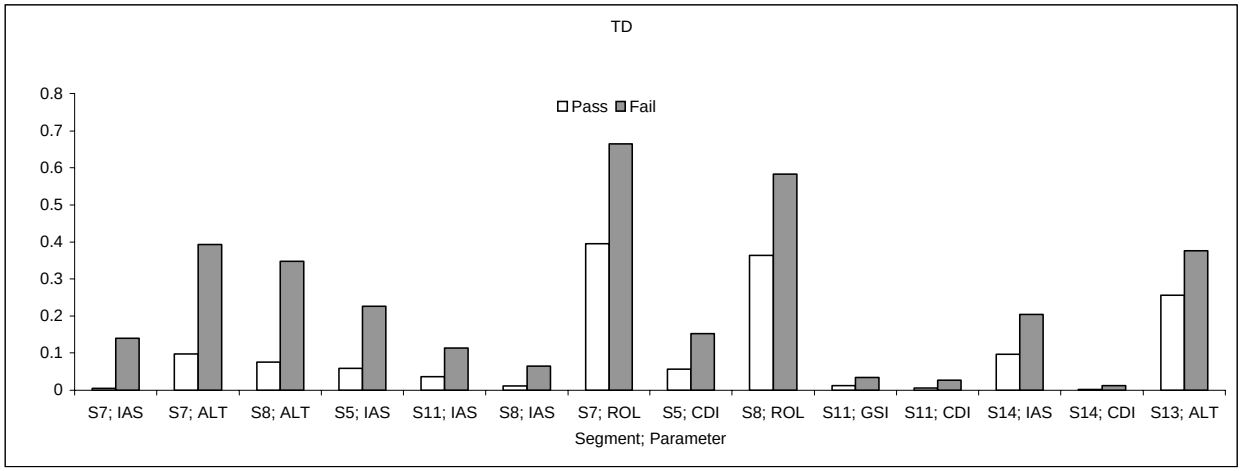

Figure 7. The durations of exceedances of tolerances (TD) are consistent with their number (ND; Figure 6). Also here, failed pilots show consistently higher durations than passed pilots. All differences between the groups depicted above were significant (p <.05).

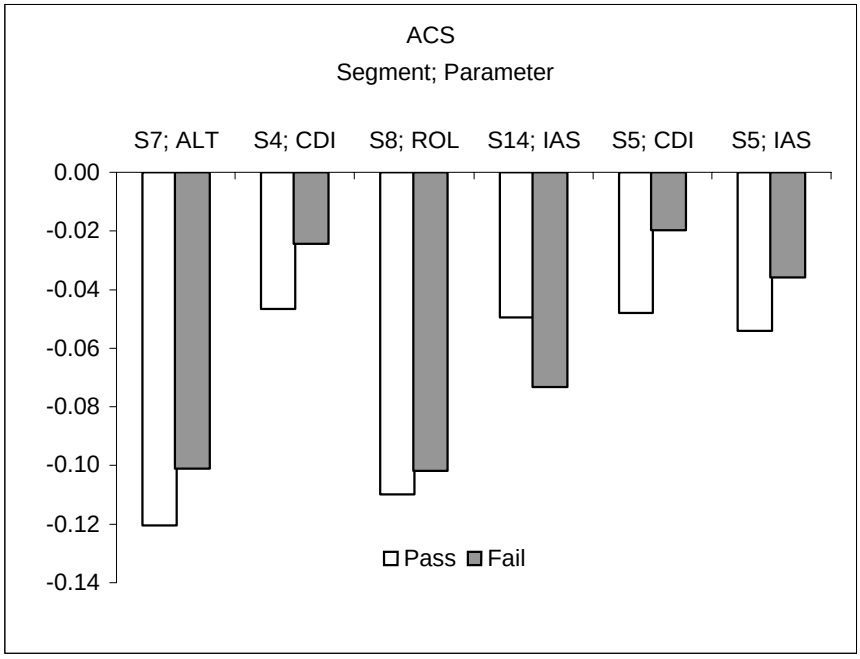

Figure 8. The ACS results were predominantly opposite to our hypothesis. These results could be partially explained by the nature of the segments, however**.** It is also noteworthy that such a small number of ACS metrics reached significance between pass/fail groups, raising questions about the relevance and validity of this metric.

ACS

We had hypothesized that good pilots would have shallower slopes of the autocorrelation function, but in the data, the opposite seems to be true. In Figure 8, segment 7 is a steep turn in which the good pilot is likely to make quick changes to maintain altitude, whereas the poor pilot will slowly drift off altitude as the maneuver progresses. In segment 8 (another steep turn but to the opposite direction) the poor pilots probably failed to reach the appropriate bank angle while good pilots makes aggressive changes in roll to maintain the proper angle of bank, again producing a steeper negative slope. However, due to the nature of the maneuver, steep turns may not be well suited for this kind of analysis. In segments 4 and 5, CDI, the good pilot is more likely than the poor pilot to make more aggressive changes to correct for off course indications as they approach the VOR, as well as immediately after passing the VOR, resulting in the higher negative slope. In segment 14, IAS does correspond to our hypothesis, and might be explained by poor pilots having difficulty flying the partial panel approach and sacrificing airspeed control as they attempt to hold altitude without a direct attitude reference.

MSC and MCGC:

These spectral components metrics seem particularly suited to differentiate good from poor CDI and GSI tracking skills. MSC and MCGC appear highly correlated on the significant segment/parameter pairings. An important result however is that MSC captures two segments where IAS has significant differences in spectral components; the MCGC metric does not capture this difference. One possible explanation is that the criterion value masked those two significant results.

These results are not surprising; the greater mean values of poor pilots' MSC and MCGC result from the fact that poor pilots have greater RMSE values for a particular flight parameter than good pilots' (Figure 4). A greater range of values in a time series will result in greater coefficient values in the Fourier decomposition and hence greater values in the power spectra (values proportional to the square of the Fourier coefficients). There are two anomalies to this generalization shown in Figures 9 and 10. Firstly, the good pilot-poor pilot pattern described above is reversed in segment 8 altitude. It is likely this is due to a failure to subtract the mean value of the altitude time series in this (and other) segment. Because raw altitude values are significantly different from zero (i.e., in thousands of feet), there will be an extremely large zero-frequency component in the power spectrum. The larger mean values of metrics MSC and MCGC for the good pilot may be an artifact of this large zero-frequency term that resulted from good pilots maintaining a higher mean altitude over the course of the flight segment. The second anomaly is the two steep turn segments (7 and 8) where Figures 9 and 10 show good pilots' MSC and MCGC values greater than poor pilots'. Similarly to the discussion above for the ACS metric, this could indicate that good pilots do not act in accordance to our original hypothesis. It could also be the case that the same problems are occurring here as in the altitude measures. Both steep turn roll angles are supposed to be held at 45 degrees by the pilot and the failure to subtract the mean value of the time series may result in artifacts in the power spectrum calculations.

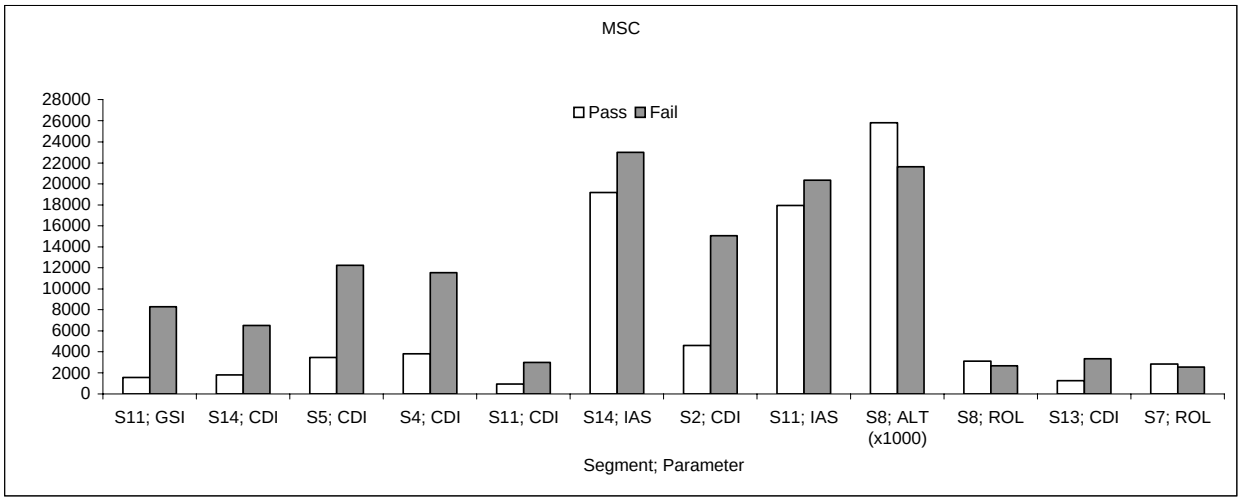

Figure 9. Means of spectral components (MSC). The results are consistent across parameters and segments, with good pilots exhibiting significantly ($p < .05$) smaller means. The only exceptions are steep turns (segments 7 and 8), but these maneuvers are clearly a special case.

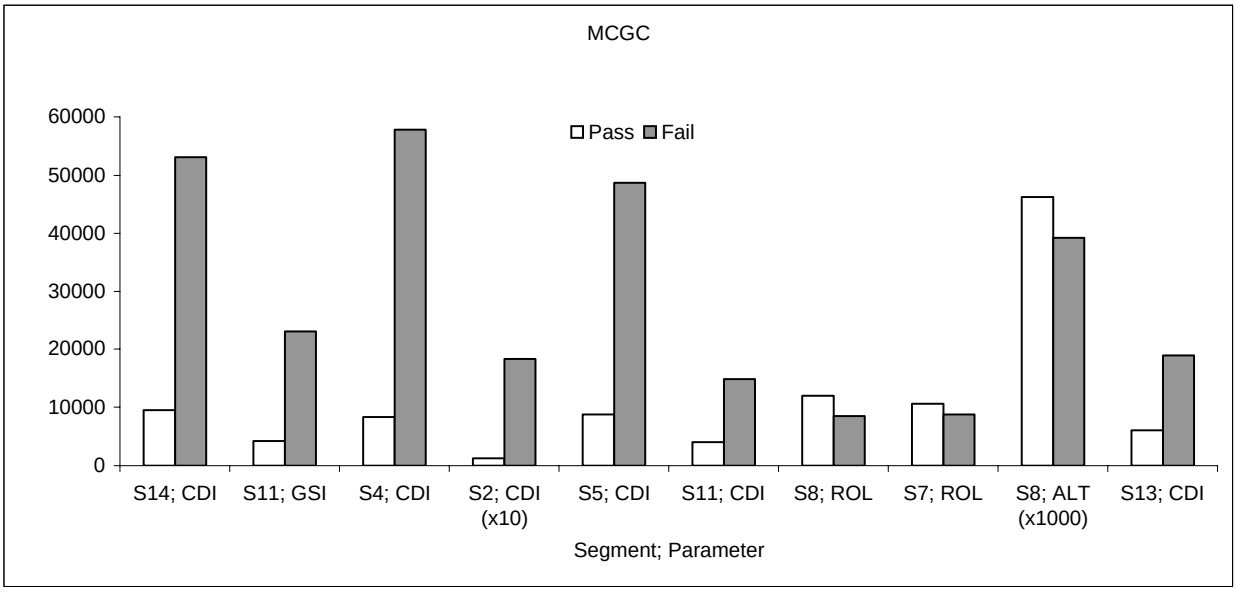

Figure 10. It was not surprising to find similar results for the means of spectral components above a criterion (MCGC); however, it is noteworthy that IAS measures did not reach significance with this measure. The results are also entirely consistent in that poor pilots exhibited significantly higher MCGC values than good pilots, with the exception of steep turns (segments 7 and 8).

DSC and DCGC

The same 11 segment/parameter pairings are represented in both the DSC and DCGC charts (Figures 11 and 12). In addition, the relationship between pass and fail groups on each pair is virtually identical between DSC and DCGC. As a result of the almost perfect correlation, the standard deviation measures should be preferred over measures of central tendency. The general conclusion is that the SD of the Spectral Components is particularly suited to differentiate good from poor pilot performance on navigation course tracking accuracy. Seven out of 11 significant segment/parameter pairings deal with course tracking. Segments 7 and 8 (altitude DSC and DCGC values are reversed for good and poor pilots when compared to the remainder of the metrics) should probably be ignored as suggested earlier, due to the unique characteristics of the steep turn maneuvers. In addition the IAS (segment 7 and 14) metric values could be affected in the same way. The airspeed time series did not have the mean value subtracted before frequency analysis and this could lead to zero-frequency components of the power spectrum that obscure predicted performance trends under out original hypothesis.

FMCG and FDCG:

As would be expected, the FMCG and FDCG measures show similar significant segment/parameter pairings for pass and fail pilots, with the exception of HDG in segment 4 for FMCG. The results for the CDI measures shown in Figures 13 and 14 support the hypothesis that good pilots' power spectra will be shifted towards higher frequencies compared to poor pilots. For the non-CDI measures (segment 7 and 14 IAS and segment 8 Roll), the same issues that are addressed above for MSC, MCGC, DSC and DCGC are relevant here: the failure to pre-process the raw time series data may lead to unanticipated results or a reduction of metric sensitivities.

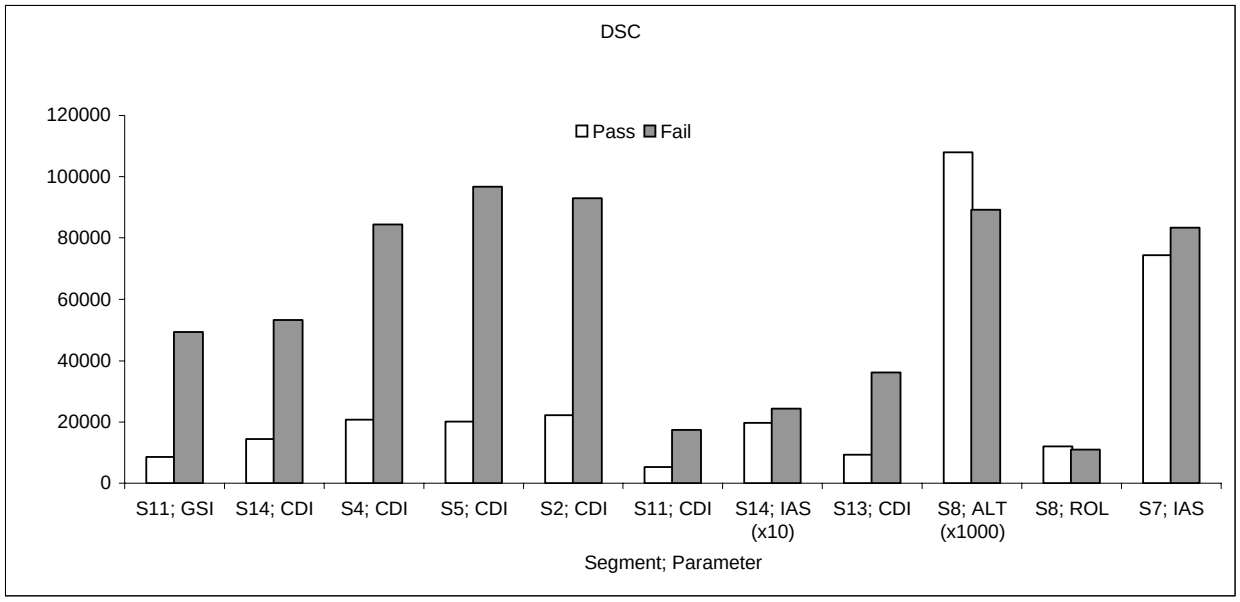

Figure 11. The results on standard deviation of spectral components (DSC) were consistent with our hypothesis across many parameters and segments; poor pilots exhibited significantly ($p < .05$) larger variability than good pilots. Segments 7 and 8 should again be considered separately as special cases.

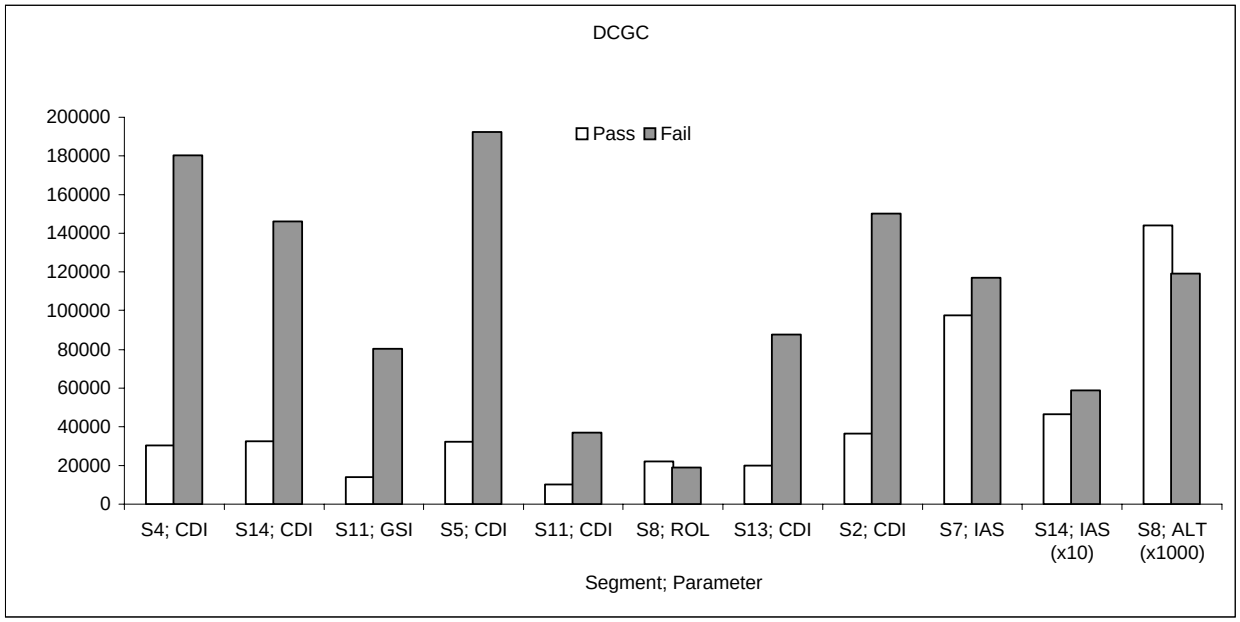

Figure 12. The results for standard deviations of spectral components over criterion are similar to those of DSC and consistent across segments and parameters (except for steep turns, segments 7 and 8)

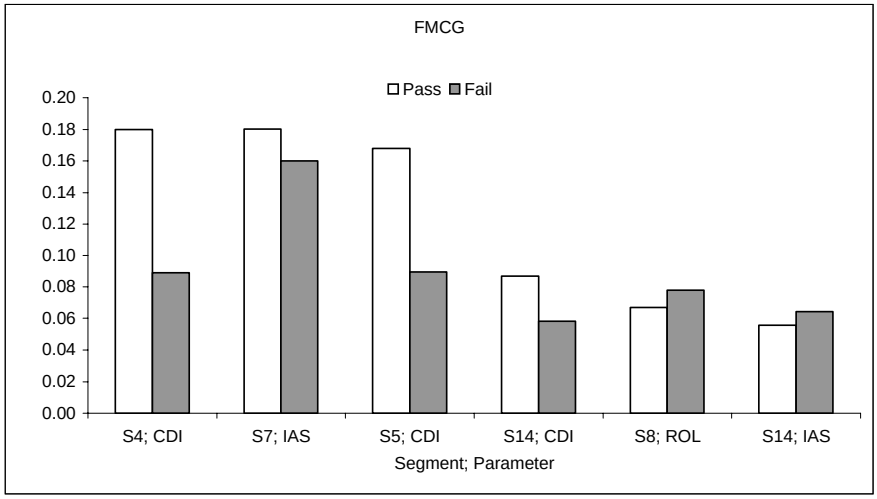

Figure 13. The results concerning the frequency distribution of spectral components were also consistent with our hypotheses and across segments and parameters. Good pilots had significantly ($p < .05$) larger number of frequencies than poor pilots, with the exception of steep turn roll control.

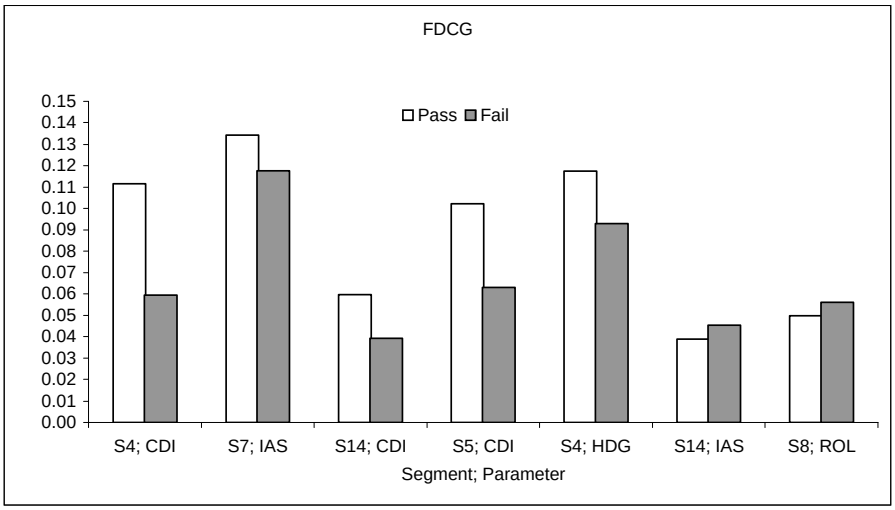

Figure 14. The standard deviation of spectral frequencies also supported our hypotheses, good pilots exhibiting significantly ($p < .05$) larger variability than poor pilots.

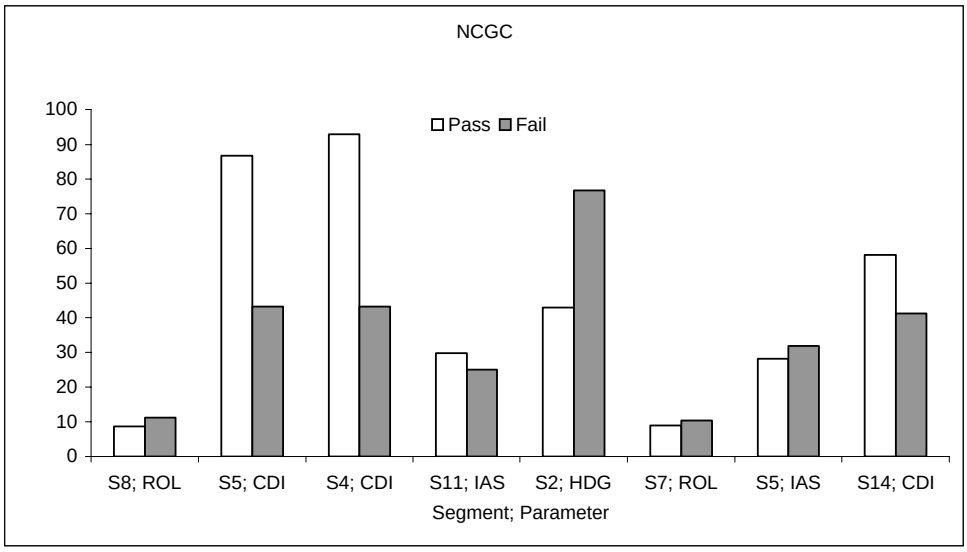

Figure 15. While the number of significant spectral components can be viewed as an auxiliary measure primarily used to determine criterion values for other metrics (MCGC, DCGC, FMCG, FDCG) it nevertheless allows for an additional look at pilots' performance. Good pilots should have higher number of significant components, which seems to be supported by the majority of the data. Steep turns are again an exception, as well as HDG in segment 2.

NCGC:

The NCGC measure is used as a simple attempt to quantify a pilot's power spectrum. Its effectiveness is limited, however, because it does not take into account the frequency and magnitude values of the spectral components. It is nevertheless useful as a guide in setting the criterion value that is used in the MCGC, DCGC, FMCG and FDCG metrics. We hypothesized that good pilots would have a greater number of "significant" spectral components; that is, a larger value for NCGC. This is supported by CDI measures in Figure 15, but as in previous measures, IAS and roll measures show that good pilots have smaller NCGC values. The segment 2 HDG result is also surprising, but given the lack of significant results for HDG measures in the Fourier-based measures, it suggests that HDG measures are not a sensitive measure of pilot performance using these techniques.

MEDF:

There were few significant Median Frequency of Power Spectrum segment/parameter pairings and these were generally less significant than other metrics for the same segment/parameter pair. It is doubtful that MEDF inclusion in any explanatory model adds any real value. It appears that the median frequency is simply not sensitive to differences in power spectra. Note that in Figure 16 we see the effect of small airspeed variations around a relatively large mean value. The segment 14 IAS MEDF values for both good and poor pilots is basically zero. This would indicate the need to subtract the mean value from such time series that do not naturally oscillate around the vale zero.

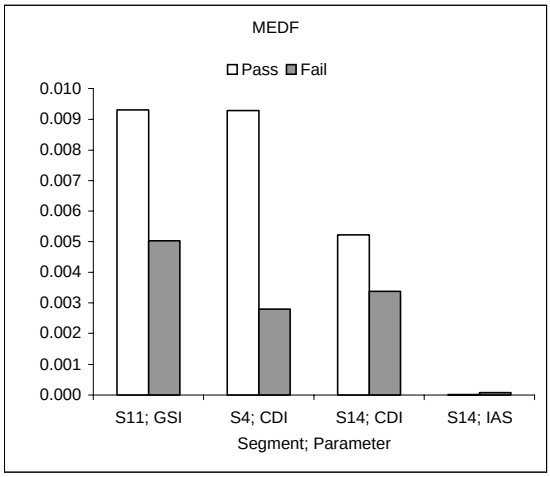

Figure 16. Very sparse results were obtained for the median frequency of spectral components.

LPF:

Low Pass Filters 1 and 2 seem particularly good at differentiating performance on tracking (CDI, GSI) and IAS segments with descents while tracking; (e.g., S11, S14, see Figures 17 and 18). LPF 3 and 4 rarely make the .05 level, however, indicating that we might have looked for differences at too high a frequency cut off. That is, the vast majority of *all* pilots' power spectrum is at frequencies less than 0.05Hz. Segments 7 and 8 (steep turns) show up on IAS and ROL in LPF 1 and 2 metrics but it is difficult to draw any conclusions on these due to the nature of the maneuver. The relationship of the dynamics of the maneuver to these patterns of these metrics needs more study. The LPF results are stable in that for CDI and GSI variables, the good pilots' values are always less than for the poor (fail) pilots'. The same holds true for ROL, but the effect is reversed. The direction of LPF effect for IAS is not stable across maneuvers. It favors the pass performance for S7 and S8, but the fail performance for S11 and S14, which might serve to underscore the unique characteristics of steep turn maneuvers. This could be either the fact that we're looking at time series values around 45 degrees (+ or -) or that good pilots do actually control the airplane in a more aggressive manner. The IAS problems may also be due to the lack of pre-analysis data formatting (i.e., subtracting the mean IAS value before taking the Fourier transform)

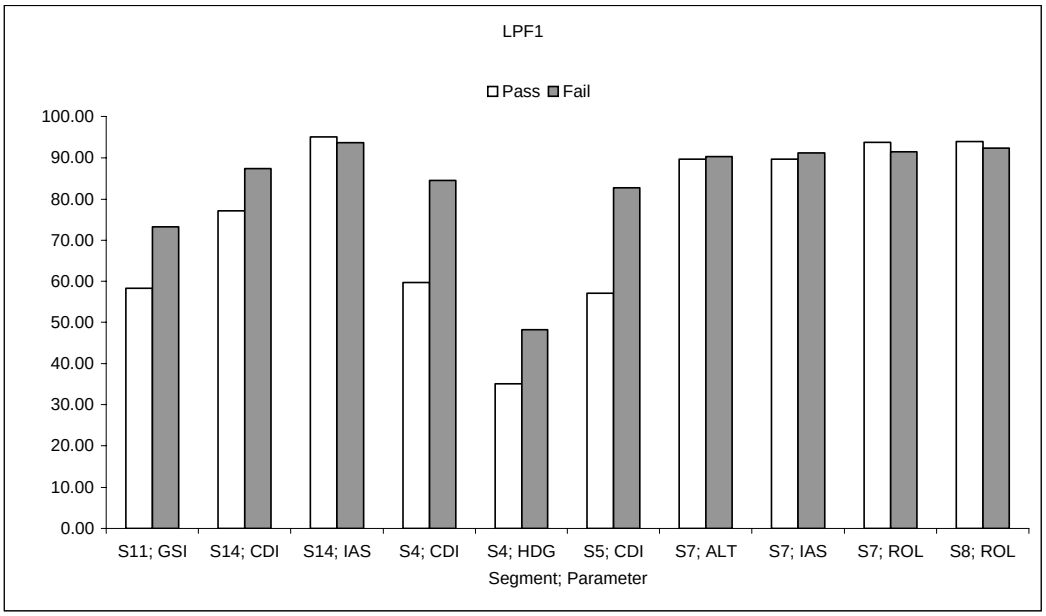

Figure 17. Low Pass Filter (0.005 Hz cutoff frequency). With the exception of steep turns and S14 IAS, this measure consistently shows significantly higher values for failed pilots than for passed pilots

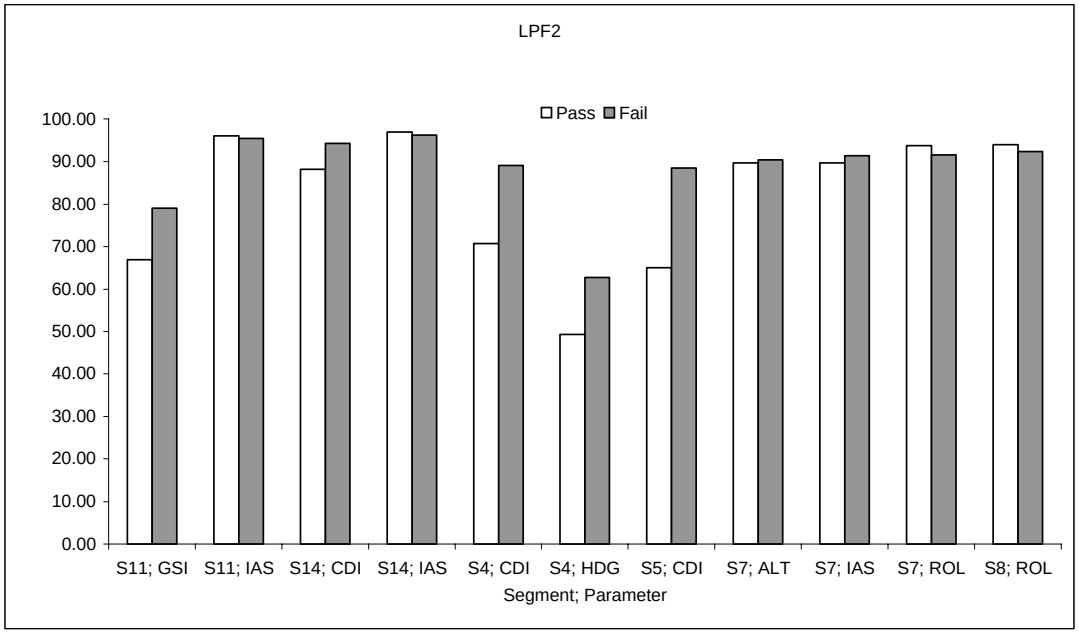

Figure 18. The results are similar for LPF2 (0.01 Hz cutoff frequency). Interestingly, this metric shows differences between pilot groups also for S11 IAS, which was missing from LPF1 results.

Segments 2, 4, and 13: VOR tracking

Another way to evaluate objective pilot performance metrics is to examine them in the context of the actual flying tasks, that is, by IPC segment. In the following tables, we examine the differences between pilots who passed and those who failed particular task elements within a segment, as judged by IPs, but in the light of objective performance metrics.

Segment 2 involved outbound tracking and segment 4 inbound tracking during a VOR approach. In both segments, CDI and HDG parameters showed significant differences between the two pilot groups (pass and fail), but there were large discrepancies between the actual metrics reaching significance. The fact that inbound tracking (segment 4, Table 5) yielded differences between the pilot groups (pass/fail) by many more measures than outbound tracking (segment 2, Table 4) may be explained by the relatively short period of time spent tracking outbound in which pilots had less opportunity to deviate from course. As with any short time-period maneuvers, many time series metrics are rendered ineffective as tools for discriminating performance.

Table 4.
Six measures from segment 2, VOR outbound tracking, showed significant ($p < .05$) differences between pass and fail groups of pilots. After rejecting highly correlated (Pearson $r > .8$, $p < .001$) measures, four metrics remain. All of these metrics also exhibit substantial effect size.

Parameter	Metric	F	P	Pass	Fail	%Diff
CDI	MCGC	44.40	0.000	12393.00	182930.00	93.23
CDI	DSC	18.53	0.000	22199.00	92892.00	76.10
HDG	NCGC	6.01	0.016	42.93	76.75	44.07
HDG	ND	4.70	0.032	6.314	11.763	46.32

Table 5.
VOR tracking inbound. Only CDI showed significant differences (except for a single HDG measure) between pilot groups, but by multiple measures.

Parameter	Metric	F	P	Pass	Fail	%Diff
CDI	DSC	49.93	0.000	20752	84456	75.43
CDI	FDCG	20.45	0.000	0.11163	0.0594	46.79
CDI	FMCG	19.89	0.000	0.18005	0.0891	50.51
CDI	RMSE	15.22	0.000	29.44	48.16	38.87
CDI	SD	15.15	0.000	29.54	48.29	38.83
CDI	LPF1	12.00	0.001	59.76	84.54	29.31
CDI	NCGC	7.83	0.006	92.89	43.2	53.49
CDI	ACS	6.98	0.009	-0.05	-0.02	47.57
CDI	LPF3	6.25	0.014	89.88	97.72	8.02
CDI	MEDF	6.24	0.014	0.00928	0.0028	69.83
CDI	LPF4	4.57	0.034	94.3	98.88	4.63
HDG	FDCG	4.97	0.027	0.117	0.093	20.79

Table 6.
Segment 13: VOR tracking inbound. Also altitude and heading yielded significant differences between pilots in this segment.

Parameter	Metric	F	P	Pass	Fail	%Diff
ALT	SD	4.87	0.029	48.55	70.68	31.31
ALT	TD	3.35	0.070	0.26	0.38	31.91
CDI	DCGC	11.65	0.001	19945	87728	77.26
CDI	DSC	10.31	0.002	9307	36142	74.25
CDI	MCGC	4.93	0.028	6015	18944	68.25
CDI	MSC	4.18	0.043	1264.8	3319.6	61.90
HDG	RMSE	11.17	0.001	19.541	42.162	53.65
HDG	SD	11.14	0.001	19.597	42.242	53.61

Segment 5 and 14: VOR final approach to FAF

The trend in VOR approach evaluation seems to be that the closer the pilots are to the runway, the clearer the differences between those who pass and those who fail. In segment 5, final approach to FAF, the number of significant ($p < .05$) metrics increased substantially over segment 4 and also now include IAS in addition to CDI (Tables 7 and 8).

Table 7.
Final VOR approach. A large number of metrics on CDI showed significant differences between the pilot groups. Also IAS differentiated between those pilots who passed and those who failed this element in this segment.

Parameter	Metric	F	P	Pass	Fail	%Diff
CDI	DCGC	47.37	0.000	32271	192296	83.22
CDI	DSC	46.21	0.000	20055	96747	79.27
CDI	MCGC	40.69	0.000	8774	48748	82.00
CDI	MSC	36.38	0.000	3444.7	12241.4	71.86
CDI	RMSE	34.53	0.000	30.789	61.932	50.29
CDI	SD	34.52	0.000	30.867	62.09	50.29
CDI	NCGC	10.31	0.002	86.78	43.17	50.25
CDI	TD	9.16	0.003	0.05678	0.15267	62.81
CDI	FMCG	8.84	0.003	0.1681	0.0895	46.76
CDI	FDCG	6.47	0.012	0.10217	0.063	38.34
CDI	LPF1	6.12	0.015	57.142	82.728	30.93
CDI	LPF2	5.48	0.021	65.065	88.53	26.51
CDI	ACS	4.56	0.034	-0.04794	-0.01983	58.64
IAS	RMSE	18.87	0.000	4.622	8.023	42.39
IAS	SD	18.85	0.000	4.63	8.04	42.38
IAS	TD	16.85	0.000	0.06	0.23	73.86
IAS	ND	12.63	0.001	0.73	1.9259	62.10
IAS	NCGC	4.21	0.042	28.15	31.80	11.49

Table 8.
Final VOR approach. The results were similar to thse from segment 5 but with substantially larger number of metrics differentiating between pilot groups.

Parameter	Metric	F	P	Pass	Fail	%Diff
CDI	MSC	62.80	0.000	1785.5	6491.5	72.49
CDI	DSC	54.46	0.000	14465	53379	72.90
CDI	MCGC	53.73	0.000	9536	53138	82.05
CDI	DCGC	52.74	0.000	32512	146178	77.76
CDI	SD	48.18	0.000	21.004	39.668	47.05
CDI	RMSE	48.15	0.000	20.961	39.587	47.05
CDI	ND	12.42	0.001	0.01567	0.09758	83.94
CDI	FDCG	7.85	0.006	0.059716	0.039167	34.41
CDI	FMCG	6.60	0.011	0.08684	0.05837	32.78
CDI	TD	6.20	0.014	0.001966	0.012208	83.90
CDI	LPF1	6.19	0.014	77.17	87.479	11.78
CDI	MTE	5.28	0.023	37903	22863	39.68
CDI	MEDF	5.11	0.025	0.005224	0.003375	35.39
CDI	LPF2	4.85	0.029	88.24	94.22	6.35
CDI	LPF4	3.02	0.084	98.75	99.42	0.67
CDI	NCGC	2.94	0.088	58.112	41.208	29.09
CDI	LPF3	2.83	0.095	97.13	98.8	1.69
IAS	MSC	16.13	0.000	19152	23013	16.78
IAS	DSC	11.83	0.001	196009	243817	19.61
IAS	ND	10.38	0.002	0.9656	2.0493	52.88
IAS	SD	9.95	0.002	5.5239	8.1621	32.32
IAS	RMSE	9.94	0.002	5.5126	8.1448	32.32
IAS	TD	8.03	0.005	0.09647	0.20377	52.66
IAS	LPF1	7.48	0.007	94.979	93.643	1.41
IAS	DCGC	6.20	0.014	464803	589333	21.13
IAS	ACS	5.99	0.016	-0.04957	-0.07331	32.38
IAS	FDCG	4.23	0.042	0.038929	0.045462	14.37
IAS	LPF2	4.20	0.042	96.87	96.176	0.72
IAS	MEDF	4.05	0.046	0.000008	0.000077	89.61
IAS	FMCG	3.85	0.052	0.055724	0.064385	13.45
IAS	LPF4	3.24	0.074	99.606	99.516	0.09
IAS	LPF3	3.05	0.083	99.198	99.029	0.17

Segment 7 and 8: Steep turns

These segments, as has been pointed out before, warrant examination separately from the rest of the IPC flight (see Tables 9 and 10). When considering control input requirements, steep turns clearly differ from other instrument control or tracking maneuvers. Specifically, the steep turn is an accelerated maneuver, which—with precise control—impose a constant g-loading on both the pilot and aircraft not experienced during any other instrument flight task. Hence, it is reasonable to conclude that the metrics chosen for analysis, although allowing significant differences between good and poor performance to be observed, may not behave in a manner consistent with other instrument maneuvers. A more detailed task analysis of the steep turn maneuver and a controlled comparison to the normal ~ 1 g turn in instrument flight is necessary in order to draw firm conclusions about the usefulness of these metrics for accelerated maneuvering.

Table 9.
Segment 7, steep turns. Altitude, airspeed, and roll control yielded most of the significant differences between the pilot groups. Note, however, that some measures—despite their statistical significance—showed such small effect sizes that their practical utility is in doubt.

Parameter	Metric	F	p	Pass	Fail	%Diff
ALT	RMSE	62.86	0.000	52.68	104.45	49.56
ALT	TD	55.14	0.000	0.09728	0.39343	75.27
ALT	SD	51.49	0.000	33.079	61.575	46.28
ALT	ND	35.58	0.000	1.0985	2.8841	61.91
ALT	MTE	10.03	0.002	140921.00	258032.00	45.39
ALT	ACS	9.89	0.002	-0.12052	-0.10108	16.13
ALT	LPF2	5.00	0.027	89.752	90.394	0.71
IAS	TD	80.84	0.000	0.00554	0.139111	96.02
IAS	ND	71.89	0.000	0.2248	2.7611	91.86
IAS	RMSE	70.41	0.000	3.1706	6.3606	50.15
IAS	SD	70.16	0.000	36.01	76.80	53.11
IAS	LPF2	15.59	0.000	89.739	91.365	1.78
IAS	FDCG	15.41	0.000	0.13429	0.11772	12.34
IAS	FMCG	12.68	0.001	0.18043	0.16006	11.29
IAS	LPF3	11.80	0.001	96.604	97.078	0.49
IAS	LPF1	11.02	0.001	89.739	91.102	1.50
IAS	DCGC	7.78	0.006	97787	117021	16.44
IAS	LPF4	6.31	0.013	98.029	98.253	0.23
IAS	DSC	4.30	0.040	74299	83377	10.89
ROL	SD	19.23	0.000	2.6964	4.2608	36.72
ROL	LPF1	17.02	0.000	93.715	91.453	2.41
ROL	LPF2	15.33	0.000	93.715	91.566	2.29
ROL	TD	12.21	0.001	0.39553	0.66463	40.49
ROL	RMSE	6.15	0.014	5.379	10.297	47.76
ROL	MCGC	6.02	0.015	10606	8738	17.61
ROL	NCGC	4.57	0.034	9.024	10.375	13.02

Table 10.
Segment 8, steep turns. The results are similar but not identical to segment 7.

Parameter	Metric		F	p	Pass	Fail	%Diff
ALT	SD	45.77	0.000		29.825	54.898	45.67
ALT	TD	45.63	0.000		0.076	0.347	78.15
ALT	RMSE	40.45	0.000		50.044	94.215	46.88
ALT	ND	18.75	0.000		1.015	2.587	60.75
ALT	MSC	9.28	0.003		25797372	21611825	16.22
ALT	DSC	7.81	0.006		1.08E+08	89283252	17.33
ALT	DCGC	5.84	0.017		1.44E+08	1.19E+08	17.36
ALT	MCGC	5.41	0.021		46274290	39247460	15.19
ALT	LPF4	3.39	0.068		98.111	98	0.11
IAS	SD	24.15	0.000		3.1872	5.3177	40.06
IAS	RMSE	24.12	0.000		3.15	5.2536	40.04
IAS	ND	15.71	0.000		0.3431	1.5085	77.26
IAS	TD	13.60	0.000		0.011242	0.065067	82.72
ROL	RMSE	23.58	0.000		4.6204	7.9268	41.71
ROL	NCGC	20.20	0.000		8.736	11.167	21.77
ROL	MCGC	16.45	0.000		11944	8528	28.60
ROL	LPF1	14.13	0.000		93.899	92.334	1.67
ROL	LPF2	13.51	0.000		93.899	92.371	1.63
ROL	DCGC	11.95	0.001		22168	18916	14.67
ROL	LPF3	11.05	0.001		98.768	98.175	0.60
ROL	TD	9.16	0.003		0.3635	0.58311	37.66
ROL	MSC	7.87	0.006		3071.3	2640.1	14.04
ROL	LPF4	7.02	0.009		99.579	99.354	0.23
ROL	SD	6.82	0.010		2.7411	3.6656	25.22
ROL	FMCG	6.48	0.012		0.066864	0.077889	14.15
ROL	ACS	6.41	0.012		-0.10987	-0.10178	7.36
ROL	DSC	5.37	0.022		11998	10897	9.18
ROL	FDCG	3.99	0.048		0.04976	0.056167	11.41

Segment 11: ILS approach, GS tracking to DH

This is one of the most critical segments in an IPC flight an the results showed clear differences between good and poor pilots by many metrics across two primary parameters, CD and GSI, as well as airspeed (Table 11).

Table 11.
Segment 11, ILS approach. Note that the 'best' measures should be judged also by the effect size in addition to the statistical significance of the results.

Parameter	Metric	F	p	Pass	Fail	%Diff
IAS	SD	15.87	0.000	4.1508	5.9397	30.12
IAS	RMSE	15.84	0.000	4.1383	5.918	30.07
IAS	TD	15.69	0.000	0.0375	0.11357	66.98
IAS	ND	15.23	0.000	0.5793	1.4386	59.73
IAS	MSC	10.35	0.002	17925	20370	12.00
IAS	NCGC	6.51	0.012	29.714	25	15.86
IAS	LPF2	4.04	0.046	96.025	95.414	0.64
CDI	ND	26.18	0.000	0.05618	0.45157	87.56
CDI	DCGC	19.02	0.000	10103	36916	72.63
CDI	MCGC	18.81	0.000	4010	14885	73.06
CDI	DSC	18.24	0.000	5279	17409	69.68
CDI	MSC	17.33	0.000	969.8	2975.7	67.41
CDI	RMSE	15.33	0.000	15.458	28.442	45.65
CDI	SD	15.32	0.000	15.507	28.534	45.65
CDI	TD	8.51	0.004	0.006202	0.027714	77.62
GSI	MSC	73.71	0.000	1530	8295.4	81.56
GSI	DSC	62.56	0.000	8596	49342	82.58
GSI	MCGC	53.66	0.000	4122	23117	82.17
GSI	DCGC	50.05	0.000	14029	80486	82.57
GSI	SD	35.01	0.000	20.016	32.972	39.29
GSI	RMSE	34.98	0.000	19.956	32.858	39.27
GSI	ND	13.73	0.000	0.20936	0.59734	64.95
GSI	LPF1	13.07	0.000	58.327	73.285	20.41
GSI	LPF2	10.64	0.001	66.923	79.092	15.39
GSI	MEDF	9.62	0.002	0.009305	0.005026	45.99
GSI	TD	9.06	0.003	0.012238	0.035132	65.17

Summary

It appears that CDI and GS tracking measures were in general the most reliable in discriminating between pilot performance groups in a manner that was consistent with our initial hypotheses. The lack of sensitivity that other flight parameter measures showed could be due to one or both of the following reasons: our failure to take the mean value away from the time series of data so that large zero-frequency components of the power spectra would not introduce artifacts or obscure other patterns in the data, or the particular nature of the flight segment (e.g., step turns) affecting the way good and poor pilots fly the airplane. The success of the CDI measures in differentiating the pass and fail pilots is also consistent the measures that De Maio and Eddowes (1978) found were most important in differentiating skill levels of pilots. Eddowes found that a low-pass filter type aileron control measure was more important in distinguishing pilots' skill level than elevator, rudder or throttle inputs. Although we did not look at control inputs as such, the lateral control of the aircraft in tracking a VOR or Localizer is obviously closely coupled to aileron control. De Maio and Eddowes did not look at a landing task and so a comparison with the GS measures that we found to be useful is not easily made.

Comparison of Device Groups

Comparison of the device groups (airplane, Frasca, and PCATD) was done by univariate ANOVAs. The results showed significant ($p < .05$) differences between the device groups for a total of 40 metrics, from 5 segments and for 6 parameters (The pass/fail criteria was not considered in this analysis but is examined later.). The results are summarized in Table 11. It is noteworthy, however, that among the top 11 most significant (as measured by the F ratio) results are 9 measures from steep turn maneuvers (segments 7 and 8), and all of these from a single parameter, roll. As we had pointed out throughout this report, steep turns are clearly exceptional maneuvers within the IPC flight and the validity of the objective metrics developed for this project in evaluation of pilot performance in them must be researched further. Therefore, at this time, we are recommend that these segments be ignored until their validity and reliability for these maneuvers can be unequivocally established.

A second finding of interest is that 15 of the remaining 30 metrics are from segment 11 (ILS final approach), and from the CDI and GSI parameters. All of these metrics are furthermore time series based. This finding supports the notion that these metrics are particularly sensitive and well suited for examination of tracking performance.

Table 12.
Statistically significant ($p < .05$) differences between the device groups (airplane, Frasca, and PCATD). Note, however, that that among the top 11 most significant (as measured by the F ratio) results are 9 measures from steep turn maneuvers (segments 7 and 8), and all of these from a single parameter, roll. At this time, the validity of these measures cannot be unequivocally established and hence they should be ignored. The majority of the remaining results are from a single segment (11, ILS final approach) and for the CDI or GSI parameters.

					Means			
Seg.	Para.	Metric	F	p	Airplane	Frasca	PCATD	Max. Diff. (%)
7	ROL	DSC	91.95	0.000	10068	13313	11902	24.37
8	ROL	MSC	83.77	0.000	2853	3971.2	3225.3	28.16
7	ROL	MSC	80.76	0.000	2580.5	3517.4	2947.5	26.64
8	ROL	DSC	77.20	0.000	11299	14742	12671	23.36
7	ROL	DCGC	72.19	0.000	18614	24602	20630	24.34
5	HDG	LPF3	68.48	0.000	66.8	94.46	89.98	29.28
5	HDG	LPF4	54.39	0.000	85.11	98.37	95.36	13.48
8	ROL	DCGC	53.77	0.000	20811	27515	21953	24.36
7	ROL	MCGC	37.65	0.000	10015	13790	9796	28.96
8	ROL	MCGC	32.15	0.000	10868	16290	10847	33.41
8	ROL	ACS	24.75	0.000	-0.10429	-0.11696	-0.10421	10.90
11	GSI	ACS	15.50	0.000	-0.04879	-0.07917	-0.04992	38.37
11	CDI	LPF3	14.03	0.001	98.6	96.06	96.13	2.58
4	BAL	FMCG	12.48	0.000	0.24877	0.23917	0.24779	3.86
11	CDI	NCGC	12.48	0.001	32.042	56.333	57.375	44.15
11	CDI	FMCG	11.51	0.001	0.07012	0.11921	0.134	47.67
11	CDI	FDCG	10.93	0.002	0.05008	0.08204	0.09938	49.61
4	HDG	FMCG	10.69	0.000	0.19653	0.15994	0.16853	18.62
11	CDI	LPF4	10.69	0.002	99.37	98.37	98	1.38
4	BAL	ACS	10.67	0.000	-0.04896	-0.0655	-0.05525	25.25
4	HDG	FDCG	9.43	0.000	0.1267	0.1078	0.1133	14.94
11	GSI	LPF3	9.02	0.004	92.198	85.104	95.513	10.90
11	CDI	MCGC	7.59	0.008	11387	2878	321	97.18
11	GSI	LPF2	7.13	0.011	75.597	63.605	70.452	15.86
11	CDI	LPF2	7.07	0.011	74.69	57.812	68.28	22.60
4	BAL	ACR	6.59	0.002	0.64017	0.40083	0.49158	37.39
4	HDG	LPF3	6.49	0.002	75.29	85.228	84.736	11.66
11	CDI	DCGC	6.06	0.018	27135	7244	721	97.34
5	CDI	MSC	5.41	0.024	5337.1	2511.5	4139.6	52.94
4	HDG	LPF4	5.24	0.007	89.12	94.05	93.48	5.24
11	CDI	DSC	5.04	0.030	12540	3981	370	97.05
11	CDI	MSC	5.03	0.030	2159.1	755	70	96.76
11	CDI	ACS	4.68	0.036	-0.03713	-0.04975	-0.04229	25.37
5	CDI	DSC	4.64	0.036	37838	15074	19330	60.16
11	GSI	LPF1	4.62	0.037	66.302	55.56	62.208	16.20
8	ROL	LPF4	4.51	0.039	99.543	99.694	99.212	0.48
5	CDI	DCGC	4.11	0.048	70356	23388	29800	66.76
5	CDI	MCGC	4.03	0.051	18018	5993	9078	66.74
4	BAL	LPF4	3.10	0.049	54.367	65.869	58.323	17.46
4	CDI	NCGC	3.09	0.049	90.83	100.6	76.49	23.97

The observed significant differences between device groups may be explained by differences in control fidelity of the three devices. Although not empirically proven to date, anecdotal information about the handling differences of FTDs and PCATDs is well established. In fact, the authorizations for the use of these devices are specifically tied to various levels of control and visual fidelity (See FAA Advisory Circulars AC 61-126; AC 120-40; AC 120-45). In the case of the GSI parameter, pitch stability and responsiveness in both the FTD and PCATD, which differ from the aircraft, and specifically the period and amplitude of control inputs required to accomplish the glide slope task, are likely to have an impact on the time series analysis,. Also, without the sensation of yaw (only a instrument indication of yaw is available in the FTD or PCATD) a pilot may have increased difficulty noticing heading changes that subsequently require course corrections to maintain CDI alignment. This will be especially noticeable during course corrections for LOC tracking, since the sensitivity of the LOC is approximately four times greater than for other VOR course tracking. The current data support this conclusion in that for flight segment 11 we found 10 significant metric differences for LOC CDI tracking. However, we also found four more significant metrics for segment 5 VOR course tracking. It is noteworthy that segment 5 tracking is the only segment of reasonable length (approx. 6 miles of tracking) that is also within 6 miles of the VOR where the course sensitivity (similar to LOC sensitivity) is high.

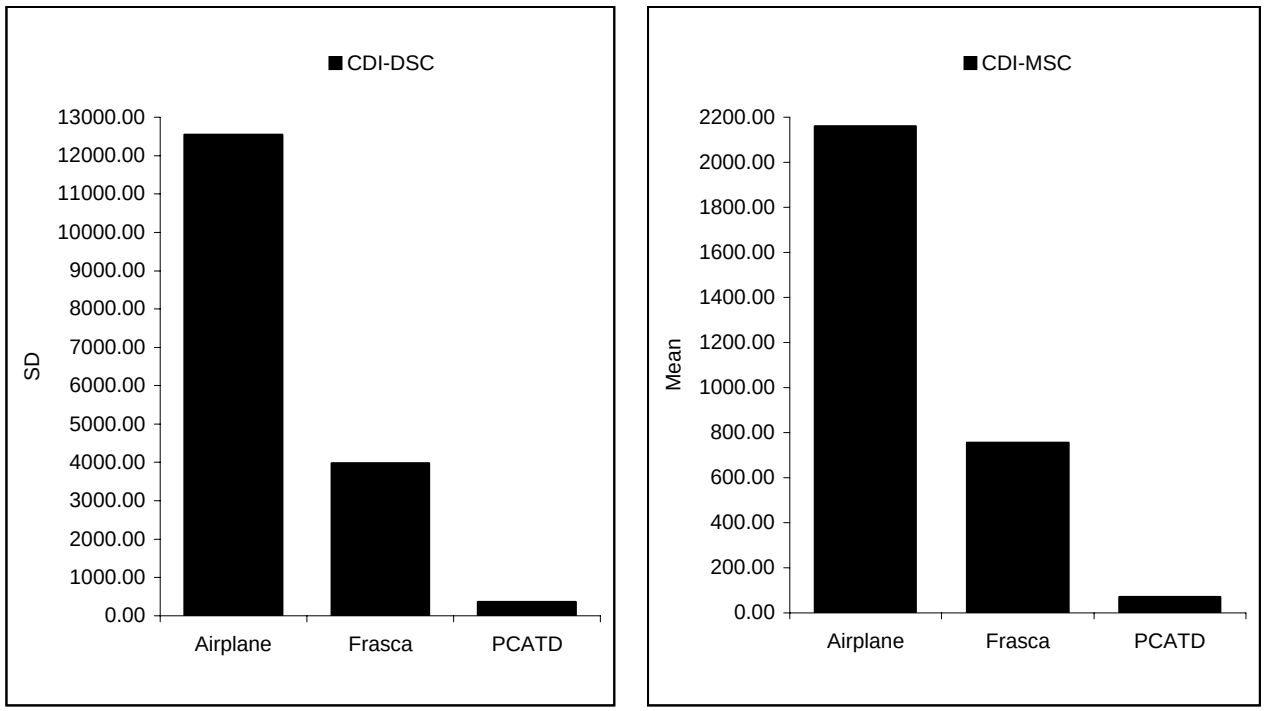

Figure 19. Both the standard deviation and mean magnitude of spectral components (DSC and MSC, respectively) show significant ($p < .05$) differences between the device groups.

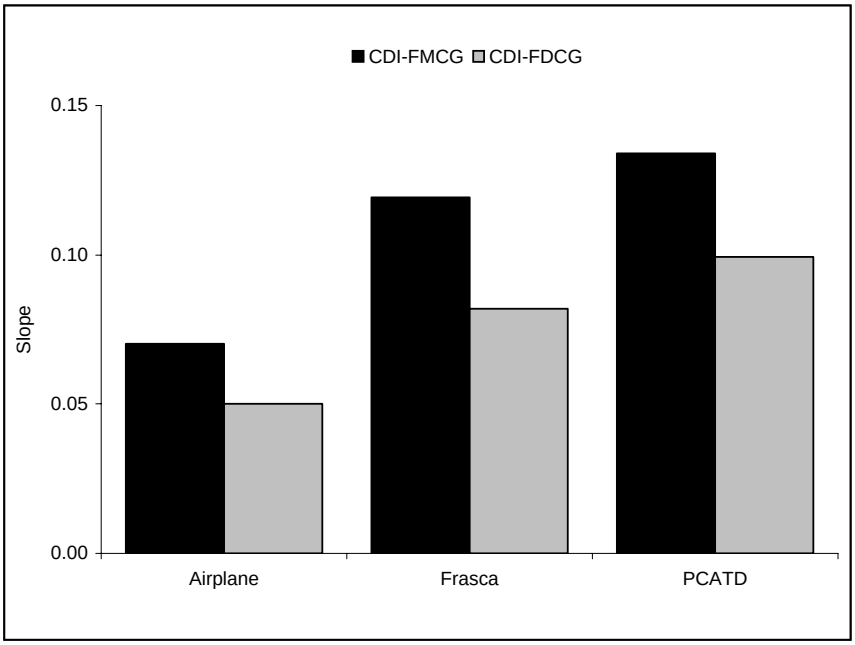

Figure 20. Significant ($p < .05$) differences between device groups can also be found by measures of spectral component frequencies (mean and SD).

Segment 4, VOR course tracking within 10 miles of the VOR, is predominantly made up of heading changes to establish a solid track of the inbound FAC for the VOR approach. The current data shows 8 significant metrics related to either HDG or BAL (yaw). Although segment 5, the continued tracking of the VOR course after crossing the VOR from segment 4, produced only 2 significant metrics related to HDG (e.g., LPF 3 & 4), headings were already well established for course tracking. As a result it is expected that fewer differences would be found on that parameters related to heading control for segment 5.

Given the many and highly significant differences between performance groups (pass/fail) discussed in the previous section, one can raise a question whether the differences in performance influenced the between device groups analyses. Rerunning the analyses by including also performance (pass/fail) as a factor in the ANOVA models—which would have addressed this point—was not possible due to the very small number of observations in the resulting groups (see Table 13 below). At the same time, however, it can be argued that the very small proportion of failed pilots could not have had undue influence on the results reported earlier.

Table 13.
The number of observations (n) in pass/fail groups by device. The small number of failed pilots rendered further analyses of between device group differences unfeasible.

		Airplane	Frasca	PCATD
S4, HDG_Eval	Pass	43	47	43
	Fail	4	1	4
S5, IAS_Eval	Pass	22	21	19
	Fail	2	3	2
S5, CDI_Eval	Pass	22	23	21
	Fail	2	1	0
S11, IAS_Eval	Pass	20	22	20
	Fail	2	2	4
S11, CDI_Eval	Pass	20	23	19
	Fail	4	1	5
S11, GSI_Eval	Pass	17	14	19
	Fail	7	10	5

Finally, although some between-device groups findings are indeed intriguing and warrant further research, the very small proportion of measures reaching significance in this comparison allows us to conclude that the device on which the IPC flights were performed did not have an impact on the performance of the participating pilots. This conclusion, based on the analysis of a multitude of different objective pilot performance metrics derived from data from different flight segments and parameters, offers substantially weighty support to our conclusions in Volume 1 of this report, that there were no differences in performance by instrument pilots on an IPC given in either a PCATD, and FTD or an airplane.

Discussion

Despite the extent of past research examined in the literature review and the amounts of data generated and analyzed within this project, the subject of objective pilot performance measures remains hard to pin down conclusively. Upon closer reflection, our experiences seem to parallel those of the Air Force researchers in the 1970s, and there may be a very good reason for this line of research to be all but abandoned in the mid 1980s. Nevertheless, this research revealed the utility of objective pilot performance metrics in discriminating between good and poor pilots and provided support to the conclusions offered in Volume 1 of this report, which were based solely on subjective performance assessment.

Our research was also unique in its systematic approach to the topic. We included many of the previously tried metrics in our analyses (in particular the static metrics, such as RMSE) but also developed a number of novel time series based measures. Unfortunately, it is difficult to compare our work to previously reported efforts, as data presented in the reviewed literature are either too sparse or they are organized in a manner (e.g., all 2436 metrics worth of raw data by Hills and Eddowes, 1974) that make meta-analyses a very time consuming process, if not altogether impossible. Our research nevertheless offers a comprehensive and systematic evaluation of all types of objective pilot performance metrics in existence and is first such effort in the general aviation domain. Furthermore, our proven inter-rater reliability (see Volume 1 of this report) ensures highest possible validity of these results (cf Knoop, 1973; Knoop & Welde, 1973).

Many objective metrics proved to be both sensitive and valid (by IP evaluations) in differentiating between good and poor performance. It appears that time series based CDI and GS tracking measures were in general the most reliable in discriminating between pilot performance groups in a manner that was consistent with our initial hypotheses as well as the measures that De Maio and Eddowes (1978) found were most important in differentiating skill levels of pilots. De Maio and Eddowes found that a low-pass filter type aileron control measure was more important in distinguishing pilots' skill level than elevator, rudder or throttle inputs. Although we did not look at control inputs as such, the lateral control of the aircraft in tracking a VOR or Localizer is obviously closely coupled to aileron control. The difference between the tasks used by De Maio and Eddowes (1978) and us makes direct comparison of the measures difficult, however.

Despite the successes of this research, there remain several obstacles for widespread use of objective pilot performance measures. First is the necessity to carefully segment the flight to be evaluated, which is a tedious and time-consuming task. Some ambiguities in our data may also have been due to the inability of even the most meticulous segmentation to sufficiently constrain the variability between individual pilots in the same maneuvers. Second, the metrics are inextricably linked to both this flight segment from which they were obtained and to the particular flight parameter measured. It is hence difficult to make general recommendations for particular metrics for general use. Rather, it should be accepted that evaluation of pilot performance by objective metrics necessitates the analysis of multiple different metrics which cannot be meaningfully combined or indexed.

Conclusions

In conclusion, we offer the following points to aid in interpretation of our results as well as to suggest areas for further research:

1. Evaluation of a multitude of objective pilot performance measures confirmed the conclusion presented in Volume 1 of this report, which was based on subjective performance evaluations, that there were no differences between the three different devices the IPC flight were administered (airplane, Frasca FTD, and PCATD). We therefore repeat the recommendation that the FAA permit the use of approved PCATD to give Instrument Proficiency Checks.
2. Objective pilot performance measures proved to be both sensitive and valid in differentiating between good and poor pilot performance. The validity of our results is highlighted by the demonstrated inter-rater reliability of the IPs providing subjective performance evaluations against which the objective results were examined.
3. Some metrics provide little additional utility over conventional measures such as SD or RMSE. Given the number of measures evaluated in this project, this was to be expected. However, time series based measures clearly provide enhanced resolution to find important differences between pilots in areas where conventional measures fail. The CDI and GSI tracking measures are the foremost examples of the utility of these metrics.
4. Some maneuvers measured warrant further investigation to determine if the reason for the inability to discriminate between good and poor pilots is related to the method of segmentation, the sampling rate, or some other peculiarity of the maneuver. Steep turns were clearly problematic to the metrics employed here. Furthermore, as no other turning maneuvers were evaluated, it is difficult to establish the stability of our metrics for such maneuvers.
5. The success of time series based metrics in providing more information than static measures (SD, RMSE, etc) especially in tracking maneuvers warrant further research to determine the effect of course width narrowing on the stability of the various metrics (e.g., when approaching a VOR, when nearing the LOC antenna, etc.). Clearly, the method of tracking as the pilot approaches the station must change to accomodate the more sensitive signal. A controlled study of tracking could easily determine how these metrics behave under such circumstances.
6. Finally, we feel confident to recommend the following subset of measures for future research on pilot performance as well as for further development and validation. These recommendations are based on the measures' ability to differentiate between pass and fail pilot groups (as determined by subjective IP ratings) by the *F* ratio and statistical significance ($p < .05$). In Table 14 below, the metrics are arranged by the frequency they appeared in the analysis at statistically significant level.

Table 14.
Frequencies of metrics exceeding statistical significance ($p < .05$) in the analyses. Recommended measures are marked with bold.

Metric	Freq.	%	Associated Parameters
SD	10	9.01	CDI, IAS
MSC	9	8.11	CDI, GSI, IAS
RMSE	9	8.11	CDI, GSI, IAS, HDG
DCGC	8	7.21	CDI, GSI, IAS
DSC	8	7.21	CDI, GSI, IAS
TD	8	7.21	CDI, GSI, IAS, ALT
LPF2	7	6.31	CDI, GSI, IAS, HDG
MCGC	7	6.31	CDI, GSI
ND	7	6.31	CDI, GSI, IAS, HDG
LPF1	6	5.41	CDI, GSI, IAS
NCGC	6	5.41	CDI, IAS, HDG
FDCG	5	4.5	CDI, IAS, HDG
LPF3	5	4.5	CDI, IAS, HDG
ACS	4	3.6	CDI, IAS
FMCG	4	3.6	CDI, IAS
MEDF	4	3.6	CDI, GSI, IAS
LPF4	3	2.7	CDI, IAS
MTE	1	0.9	CDI

Based on the data in the above table, we make the following recommendations for metrics to be used in objective pilot performance measurement:

i. Standard deviation (SD) is simple and robust measure; however, it must be kept in mind that it does not provide any information about error relative to any criteria, and hence its use should be considered against the alternative of RMSE.

ii. Root mean square error (RMSE) is widely used measure, it is simple and easy to compute, and it proved to be very robust in our analyses.

iii. Standard deviation of spectral components (DSC) is a robust metric that showed consistent differences between performance groups in accordance to our hypothesis and proved to be superior to static measures in evaluating tracking performance. It is also more sensitive than MSC and does not depend on criteria; hence, we recommend it over MSC, MCGC, and DCGC metrics.

iv. Frequency standard deviation of spectral components greater than criterion (FDCG) appears relatively low in Table 14; its inclusion in this short list is justified by the unique information it provides in addition to other metrics.

References

Benton, C. J., Corriveau, P., & Koonce, J. M. (1993). *Concept development and design of a semi-automated flight evaluation system (SAFES)* [AL/HR-TR-1993-0124]. Brooks AFB, TX: Armstrong Laboratory, Human Resources Directorate.

Carter, V. E. (1977). Development of automated performance measures for introductory air combat maneuvers. *Proceedings of the Human Factors Society 21st Annual Meeting* (pp.). Santa Monica, CA: HFES.

Childs, J. M. (1979). The development of objective inflight performance assessment procedures. *Proceedings of the Human Factors Society 23rd Annual Meeting* (pp.). Santa Monica, CA: HFES.

Connelly, E. M., Bourne, F. J., Loental, D. G., Migliaccio, J. S., Burchick, D. A. & Knoop, P. A. (1974). *Candidate T-37 pilot performance measures for five contact maneuvers* [AFHRL-TR-74-88]. Wright Patterson AFB, OH: Air Force Human Resources Laboratory.

De Maio, J., Bell, H. H., & Brunderman, J. (1985). *Pilot-oriented performance measurement* [AFHRL-TP-85-18]. Brooks AFB, TX: Air Force Human Resources Laboratory.

De Maio, J. C., & Eddowes, E. E. (1978). *Airborne performance measurement assessment: Low altitude tactical formation in two operating environments* [AFHRL-TR-79-44]. Brooks AFB, TX: Air Force Human Resources Laboratory.

Fox, J., Merwin, D. Marsh, R., McConkie, G. & Kramer, A. (1996). Information extraction during instrument flight: An evaluation of the validity of the eye-mind hypothesis. *Proceedings of the Human Factors Society 40th Annual Meeting.* Santa Monica, CA: HFES.

Gawron, V. J. (2000) *Human performance measures handbook*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gerlach, V. S. (1972). *Model and procedures for an objective maneuver analysis* [Technical Report No. 21201]. Arizona State University: Instructional Resources Laboratory.

Hennessy, R. T., Hockenberger, R. K., Barnebey, S. F. & Vreuls, D. (1979). Design requirements for an automated performance measurement and grading system for the UH-1 flight simulator [FtR-02-78]. Fort Rucker, AL: US Army Aviation Center.

Hills, J. W., & Eddowes, E. E. (1974). *Further development of automated GAT-1 performance measures* [Final Report 73-72]. Brooks AFB, TX: Air Force Human Resources Laboratory.

Hughes, M. F. & Takallu, M. A. (2002). Terrain portrayal for head-down displays experiment. *International Advanced Aviation Technology Conference*. Anchorage, AK.

Kelly, J. K., Wooldridge, A. L., Hennessy, R. T. & Reed, J. C. (1979) Air combat maneuvering performance measurement. *Proceedings of the Human Factors Society 23rd Annual Meeting* (pp.). Santa Monica, CA: HFES.

Knoop, P. A. (1973). Advanced instructional provisions and automated performance measurement. *Human Factors, 15*(6), 583-597

Knoop, P. A. & Welde, W. L. (1973). *Automated pilot performance assessment in the T-37: a feasibility study* [AFHRL-72-6]. Wright Patterson AFB, OH: Air Force Human Resources Laboratory.

Lendrum, L., Taylor, H. L., Talleur, D. A., Hulin, C. L., Bradshaw, G. L., & Emanuel, T. W. (2000). *IPC data logger operation manual* (ARL-00-8/FAA-00-5). Savoy, IL: University of Illinois, Aviation Research Lab.

Martin, E. L., & Rinalducci, E. J. (1983). Low-level flight simulation vertical cues [AFHRL-TR-83-17]. Brooks AFB, TX: Air Force Human Resources Laboratory.

McDowell, E. D. (1978). *The development and evaluation of objective frequency domain based pilot performance measures in ASUPT*. [Report No.]. Bollings AFB, DC: Air Force Office of Scientific Research.

Mixon, T. R., & Moroney, W. F. (1982). An annotated bibliography of objective pilot performance measurements [NAVTRAEQIPCEN-IH-330]. Orlando, FL: Naval Training Equipment Center.

Rantanen, E. M., & Talleur, D. A. (2001). Measurement of pilot performance during instrument flight using flight data recorders. *International Journal of Aviation Research and Development, 1*(2), 89-102.

Reising J. M., Ligget, K. K., Solz, T. J., & Hartsock, D. C. (1995). A comparison of two head up display formats used to fly curved instrument approaches. *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society.

Reynolds, M. C., Purvis, B. D., & Marshak, W P. (1990). A demonstration/evaluation of B-1B flight director computer control laws: A pilot performance study. *Proceedings of the IEEE National Aerospace and Electronics Conference* (pp. 490-494). Piscataway, NJ: IEEE

Scallen, S. F., Hancock, P. A., & Duley, J A. (1995). Pilot performance and preference for short cycles of automation in adaptive function allocation. *Applied Ergonomics, 26*(6), 397-403.

Semple, C. A., Cotton, J. C. & Sullivan, D. J. (1981). *Aircrew training devices: Instructional support features* [AFHRL-TR-80-58]. Brooks AFB, TX: Air Force Human Resources Laboratory.

Sirevaag, E. J., Kramer, A. F., Wickens, C. D., Reisweber, M., Strayer, D. L., & Grenell, J. F. (1993). Assessment of pilot performance and mental workload in rotary wing aircraft. *Ergonomics, 36*(9), 1121-1140.

Stave, A. M. (1977). The effects of cockpit environment on long term pilot performance. *Human Factors, 5*, 503-514.

Svensson, E., Angelborg-Thanderz, M., Sjoberg, L., & Olsson, S. (1997) Information complexity –mental workload and performance in combat aircraft. *Ergonomics, 40*(3), 362-380.

Swink, J. R., Butler, E. A., Lankford, H. E., Miller, R. M., Watkins, H. & Waag, W. L. (1978). *Definition of requirements for a performance measurement system for C-5 aircrew members* [AFHRL-78-54] Brooks AFB, TX: Air Force Human Resources Laboratory.

Takallu, M. A., Wong, D. T. & Uenking, M. D. (2002). Synthetic vision systems in GA cockpit—evaluation of basic maneuvers performed by low time GA pilots during transition from VMC to IMC. *International Advanced Aviation Technology Conference*. Anchorage, AK.

Ververs, P. M., & Wickens, C. D. (1996). The effect of clutter and lowlighting symbology on pilot performance with head up displays. *Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society.

Vreuls, D., & Obermayer, R. W. (1985). Human-system performance measurement in training simulators. *Human Factors, 27*(3), 241-250.

Vreuls, D., Wooldridge, A. L., Obermayer, R. W., Johnson, R. M., Norman, D. A., & Goldstein, I. (1975). *Development and evaluation of trainee performance measures in an automated instrument flight maneuvers trainer*. [Report NAVTRAEQUIPCEN 74-C-0063-1]. Orlando, FL: Human Factors Laboratory, Naval Training Equipment Center.

Wickens, C. D., & Holland, J. G. (2000). *Engineering psychology and human performance* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

Wooldridge, L., Obermayer, R. W., Nelson, W. H., Kelly, M. J., Vreuls, D. and Norman, D. A. (1982). Air combat maneuvering performance measurement state space analysis [AFHRL-TR-82-15]. Brooks AFB, TX: Air Force Human Resources Laboratory.

Appendix

Publications Emanating From This Research

Johnson, N. R., Rantanen, E. M., & Talleur, D. A. (2004). Time series based objective pilot performance measures. *International Journal of Applied Aviation Studies (IJAAS), 4*(1), 13-29.

Johnson, N. R., Rantanen, E. M., & Talleur, D. A. (2004). Criterion setting for objective, fourier analysis based pilot performance metrics. *Proceedings of the 48th Annual Meeting of the Human Factors and Ergonomics Society* (157-160). Santa Monica, CA: HFES.

Johnson, N. R., & Rantanen, E. M. (In review). Objective pilot performance measurement: A literature review and taxonomy of metrics. *Proceedings of the 13th International Symposium on Aviation Psychology.*

Rantanen, E. M., Johnson, N. R., & Talleur, D. A. (In preparation). Evaluation of objective pilot performance measures across training platforms. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans.*