

OBJECTIVE PILOT PERFORMANCE MEASUREMENT:  
A LITERATURE REVIEW AND TAXONOMY OF METRICS

Nicholas R. Johnson and Esa M. Rantanen  
University of Illinois at Urbana-Champaign, Aviation Human Factors Division  
Savoy, Illinois, USA

This paper will present a review of development and evaluation of objective pilot performance measures. A basic taxonomy of measure types is also presented where objective measures are classified distinguishing between direct measures, derivative metrics based on these, and indirect measures of unobservable constructs and phenomena. The recent development of novel objective measures and their ability to predict and classify pilot performance will also be reported.

### Introduction

Traditional means of evaluating pilot performance have typically involved the subjective evaluation of an instructor or highly experienced pilot. Performance assessment based on subjective evaluations has the advantages of relative ease of implementation, high face validity, and simplicity in providing specific feedback to the examinee pilot. However, subjective evaluations of performance are prone to the problems of inter and intra-rater reliability. Additional factors such as group performance levels and raters' temporal variance of performance standards may also influence the reliability of subjective evaluations of pilot performance. Objective measures based on flight data recordings have the potential to alleviate these concerns and their use can provide an alternative or complimentary approach to pilot performance measurement in training and research environments. The utility of objective performance measures includes use in transfer of training studies, validation of training methods, development of automated or adaptive training methods, logging of solo flights for subsequent evaluation, and standardized check rides. In addition, by alleviating the time constraints and information overload often associated with direct observation, automated data collection can enhance and expand traditional proficiency evaluation methods by an instructor pilot. Furthermore, quantitative performance data can be utilized in research and subjected to various statistical analyses to reveal underlying, covert patterns in pilots' performance.

### Review of Metrics

Although our literature review was aimed to be exhaustive, much of the past military 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. Gawron's (2000) handbook of performance measures includes objective measures of however, the major focus of the review is on workload measures.

### Basic Measures

Ideally, objective measures should be temporally invariant (repeatable), criteria based, insensitive to group performance, transferable between simulator and aircraft, interpretable, and for training purposes, be immediately available for feedback or monitoring. Basic measures such as Root Mean Square Error (RMSE) and Standard Deviation (SD) satisfy most of these requirements. However they do not contain information about the direction of the deviations or the frequency of these deviations from the mean. Consequently, identical numeric values for RMSE (or SD) can result from quite distinct performance. Hubbard (1987) noted that by using the mean and the SD together instead of the RMSE, a more complete picture of performance could be obtained, albeit at a cost of having to simultaneously interpret two measures instead of just one.

Measures based on amplitudes of flight parameter data can provide additional information on pilot performance when combined with tolerance values. Number of deviations (ND) and time spent outside parameter tolerances (TD) are an example of such measures. The 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. The TD measure provides an indication of tracking performance beyond the RMSE and ND measures. TD 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, Kramer, Wickens, Reisweber, Strayer and Grenell (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.

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. In subsequent analysis, the MTE from tracking the localizer on an instrument landing system (ILS) approach showed a significant difference between pilots who passed an instrument proficiency check flight and those who failed, by flight instructor evaluation (Rantanen & Talleur, 2001).

#### Novel measures

The above measures are relatively simple, have a high degree of face validity and are easily interpreted. Attempts have also been made to develop more novel measures, based on somewhat more complicated constructs or data analysis.

De Maio, Bell, & Brunderman (1985) defined a critical control input as a pilot control input that changed or led to a change from positive vertical acceleration to negative vertical acceleration (or other flight parameter) or vice versa. Conversely, a non-critical control input did not cause the vertical acceleration to change from positive to negative or vice versa. The authors 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. A measure of "smoothness," was subsequently 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 et al. (1985) found that that smoothness and critical error rate were affected by flight task difficulty (straight vs. turning flight, both at constant altitude).

The  $n^{\text{th}}$  moment of a series of data is the summation of individual series values raised to the  $n^{\text{th}}$  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 McDowell (1978), where the aileron second moment showed differences between pilot experience groups in the study.

Gaidai and Mel'nikov (1985) developed a measure based on an integral equation to evaluate pilot performance in a landing task. The measure took the weighted sum of normalized deviations from criterion values over several flight parameters. The explicit form of the equation can also be found in Gawron (2000).

Frequency analysis (Bloomfield, 1976; Gottman, 1981) has also been identified as a useful tool to aid in performance measurement (Semple, Cotton & Sullivan, 1981; Benton, Cooriveau & Koonce, 1993). However, actual implementations of such frequency-based measures have been limited. Hills and Eddowes (1974) and Vreuls, Wooldridge, Obermayer, Johnson, Normal & Goldstein (1976) 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 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 that attempted to classify pilot experience groups. Vreuls et al. (1976) performed a similar analysis, however the exact nature of the frequency-based measures that

were included is not clear. By contrast, McDowell (1978) did not use a manual tracking approach 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 0.125Hz 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.

Following a similar approach to McDowell (1978), Johnson, Rantanen and Talleur (2004) developed a number of frequency analysis based measures. Using data collected from instrument proficiency check flights in an aircraft and two types of flight simulators, Johnson et al. derived seven distinct measures of performance across nine different flight parameters. The authors found that many of the measures were sensitive to differences in pilot performance as judged by an instructor pilot (IP). Specifically, measures of the mean and standard deviation of the magnitude of components in the frequency distribution of a flight parameter's time series data were found to be sensitive to pilot performance. Using the same data, Rantanen, Johnson and Talleur (2004) found such differences in pilot performance were most clearly seen during localizer and glide-slope tracking on an ILS approach. In addition, low-pass filter measures were implemented in a similar manner to McDowell (1978). While these measures were not, in general, as effective at separating pilot performance groups (as judged by an IP), their sensitivity may have been limited by not setting the cut-off frequency optimally.

#### Combining Metrics

Further objective measure development has included techniques based on combining individual flight parameter measures into an index of pilot performance. Knoop and Welde (1973) used a summation of absolute values of flight parameter deviations from criterion values at four chosen points in a lazy-8 flying maneuver performed in a T-37 military training aircraft. The flight parameters used in this index included airspeed, altitude, heading, pitch, roll and pitch, roll and yaw rates. The index was compared to subjective evaluations of an instructor pilot (IP) where it was found that the index accounted for 67% of the variance of the IP's ratings. In a parallel study, Knoop (1973) also introduced the idea of a linear combination of Boolean measures based on pilot performance within flight parameter

tolerance ranges. That is, a 1 or 0 would be scored by the pilot depending on whether they flew the aircraft within the acceptable range of flight parameter value. Problems of intra- and inter-rater reliability made meaningful comparison of this index with subjective evaluations difficult. Childs (1979) also used an index based on categorical values in assessment of helicopter pilot performance. Altitude, airspeed and heading performance were evaluated and a score from 1 to 6 was assigned in each flight segment based on whether performance was within one of three tolerance bands. These segment scores were then combined to form an overall flight score from 1 to 6. Results indicated this measure was sensitive to training time, but no other validation was attempted.

Connelly, Bourne, Loental, Migliaccio, Burchick and Knoop (1974) performed a study that was concerned with developing candidate measures for pilot performance evaluation in a T-37 aircraft. They studied lazy-8, approach and landing, barrel roll, split-S, and cloverleaf maneuvers, specified the flight parameters and control inputs to be recorded and developed measures for performance evaluation. These measures were formulated in terms of continuous differences from a reference trajectory, where this trajectory could be empirically derived, and tolerance values were based on either external criterion or SDs from empirical data. Linear combinations of weighted errors (c.f., Knoop & Welde, 1973) and vector combination of error terms (allowing simultaneous comparison of all error terms) were discussed but because no data was collected, these combinations could not be evaluated.

Bortolussi and Vidulich (1991) developed a figure of merit (FOM) of pilot performance from six primary flight variables (control inputs, altitude, airspeed and heading). The authors studied both a total FOM (derived from standard deviations of the six variables and the altitude, airspeed and heading means) and specific flight parameter FOMs. For example, an altitude FOM was derived from altitude mean and SD and from the SD of elevator inputs. The FOMs were produced by a weighted linear combination of their component primary variables. The authors found the weighting coefficients by using an analytical hierarchy process. The FOMs were evaluated by comparison between two flight scenarios of differing difficulty. While the total FOM was not sensitive to scenario difficulty, the altitude and airspeed FOMs did differ between conditions.

It is evident that complex tasks—such as flying— involving multiple dimensions (exemplified by flight parameters in our discussion) can yield a vast number

of measures. In addition to the efforts described above to combine measures, several attempts have been made to statistically reduce the number of measures using discriminant analysis (Hills & Eddowes, 1974; Vreuls et al, 1975; Kelly, Wooldridge, Hennessy and Reed (1979). For example, Hills and Eddowes' (1974) study yielded a total of 2436 measures per subject. The authors attempted to distinguish three pilot experience groups based on the objective measures derived from the flight tasks. One-way ANOVAs were used to determine the ability of each measure to independently predict pilot group membership. Only a little over 17 % (420) of the variables were found to be statistically significant in separating groups. Standard deviation measures produced the highest proportion of significant variables (32%), followed by frequency analysis measures (20%), means (18%) and correlations (11%). A linear discriminant function that was derived from the results of the first experiment was used to classify performance in the second experiment with new subjects. The classification process was statistically successful. However, the discriminant function misclassified 33% of subjects, leading the authors to question the practicality of using discriminant functions to diagnose performance.

Vreuls et al. (1976) also sought to limit the number of performance measures and utilize those that could discriminate between early and late training in an automated IFR training simulator. Basic measures and measures derived from frequency analysis of standard flight parameters were used to generate a discriminant function. The discriminant function contained 9 derived measures on average, including several control input variables. Using these discriminant functions in automated feedback training scenarios reduced training time to set criteria by 34–40% compared to the original method that was not based on a discriminant function. The mixed results of Hills and Eddowes (1974) and Vreuls et al. (1976) highlight the difficulty in reducing a large number of performance measures into a manageable set that can be used to reliably predict skill level or measure performance.

#### Taxonomies of Measures

In addition to statistical techniques, some kind of classification system should be considered to help manage the large number of metrics. A thorough review of past and current research efforts and organization of the findings in a manner that facilitates the use of existing knowledge is critical for future evolution of pilot performance

measurement. On one hand, this helps to avoid 'reinventing the wheel.' On the other hand, periodic literature reviews provide for a foundation for future research efforts by defining a 'toolbox' of measures that would predict pilots' success in their task and the impact of changing training protocols, procedures, and new technology on the system as a whole.

The basic structure of the taxonomy of measures proposed here is classification by flight parameters and distinguishing between *direct* measures, *derivative* metrics based on these, and *indirect* measures. Direct measures are momentary values of flight parameters, for example, altitude or heading. Derivative metrics are based on these, for example, mean and standard deviation of altitude values. Indirect measures are those that cannot be measured directly but must be inferred from derived measures, for example, *pilot performance* based on standard deviation of altitude in level flight.

Table 1 depicts the directly measured variables found in the literature, the frequency of their encounters, the percentage of all parameters, and a cumulative percentage. Altitude, airspeed, roll, control inputs, heading and pitch were the most frequently measured variables, together accounting for over 65% of all parameters measured.

Table 1.  
*Frequencies of flight parameters.*

Parameter	Freq.	%	Cum. %
Altitude	21	12.88	12.88
Airspeed	19	11.66	24.54
Roll	17	10.43	34.97
Control Inputs	17	10.43	45.40
Heading	16	9.82	55.21
Pitch	16	9.82	65.03
Vertical Speed	11	6.75	71.78
VOR Tracking*	8	4.91	76.69
Yaw	5	3.07	79.75
Turn Rate	5	3.07	82.82
Glide Slope Tracking	5	3.07	85.89
Flaps	4	2.45	88.34
Trim	4	2.45	90.80
Speed Brakes	3	1.84	92.64
Sideslip	3	1.84	94.48
Landing Gear	3	1.84	96.32
Acceleration	3	1.84	98.16
Position	2	1.23	99.39
NDB tracking**	1	0.61	100.00

\*VOR = Very High Frequency Omnidirectional Range

\*\* NDB = Non-directional Beacon

There are two main issues to consider when

interpreting these results. First, the ease and practicality of making measurements of any particular parameter clearly plays a role in their ranking in Table 1. Second, the relevance of the parameters depends heavily on the particular flight maneuver to be evaluated. For example, altitude measurements may yield little useful information if the pilot is climbing or descending (c.f., Rantanen, Johnson, & Talleur, 2004).

The second major class, derivative measures, may be further divided into several subclasses according to the particular (mainly statistical) techniques used to reduce the often massive amounts of data into something manageable and interpretable. These derivative metrics are depicted in Table 2, again ranked by the frequency they were encountered in the literature. Not surprisingly, RMSE ranked first, followed by SD, maximum and minimum values and mean.

Table 2.  
*Derivative measures used in the literature.*

Derivative Metric	Freq.	%	Cum. %
RMSE	16	21.92	21.92
Std. Dev.	8	10.96	32.88
Max/min	8	10.96	43.84
Mean	6	8.22	52.05
Frequency Analyses	5	6.85	58.90
Range	5	6.85	65.75
Deviation from criterion	4	5.48	71.23
Time on target	4	5.48	76.71
Mean absolute error	3	4.11	80.82
Autocorrelation	3	4.11	84.93
Time outside tolerance	3	4.11	89.04
Median	2	2.74	91.78
ND	2	2.74	94.52
Boolean	1	1.37	95.89
Correlation	1	1.37	97.26
Moments	1	1.37	98.63
MTE	1	1.37	100.00

Finally, the third main class of measures, indirect measures, only had one subcategory: pilot performance. What is noteworthy, however, is that very little was found in our literature review that would link direct measures to the measures of real interest, that is, performance, via a valid or even plausible theoretical construct. As availability of data from flight data recorders and data outputs from ground-based trainers is not a problem, and as there exists many established techniques to process and reduce these data to metrics (c.f., Table 2), the lack of theoretical foundation for measurement of pilot performance is conspicuous.

Another measure classification scheme to help in data

reduction and interpretation is based on task analysis of piloting an airplane. The navigational goals of a pilot (e.g., a given heading, altitude, or track over ground) are hierarchical (Wickens, 2003). Furthermore, the control order changes across the hierarchy, and, given that humans have increasing difficulty controlling higher order systems, it is important to recognize what is the appropriate parameter to control in a given task or situation. For example, aircraft altitude control depends on the zero-order control of elevator angle, which results in the first order control of pitch angle, which in turn affects the second order control of vertical speed of the aircraft, which finally determines the aircraft's altitude, which can be seen as a third order control task. Obviously, other controls are coupled with this task, for example, engine thrust (zero-order) and airspeed (first order), further complicating the pilots' task. However, such hierarchy offers a promising framework for the choice, analysis, and interpretation of objective metrics available from different maneuvers.

## Discussion

This review has highlighted both standard objective measures and attempts to develop novel diagnostic measures of pilot performance. Despite a relatively long history and numerous and varied approaches to development of objective pilot performance measures, successes of measure validation for all but the most basic metrics have been limited. Efforts to corroborate the effectiveness of objective measures in describing pilot performance have focused on the measures' sensitivity to training, correlations with subjective evaluations, or performance in cross-validation studies.

While some of the mathematical techniques described in this paper offer the potential to uncover detail or patterns in pilot performance that may not be perceptible or quantifiable to a human observer, the task of flying an airplane is such a multi-faceted one that simply looking at a single flight parameter may not yield much diagnostic information. Instead, it appears that the greatest potential for diagnostic objective indices lies in the formation of measures combined from various related direct measures. Such combinations should be based on a detailed analysis of the flying task involved and utilize the natural linking of flight parameters through the hierarchical structure of pilot goals and control order. Even if combinations of objective measures fail to produce the performance sensitivity and diagnosticity required for research and training purposes, they can still be used to assist pilot performance evaluation.

## 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 Lab., Human Resources Directorate.
- Bloomfield, P. (1976). *Fourier analysis of time series: An introduction*. New York: Wiley.
- Bortolussi, M.R. & Vidulich, M.A. (1991). An evaluation of strategic behaviors in a high fidelity simulated flight task. Comparing primary performance to a figure of merit. *Proceedings of the 6<sup>th</sup> ISAP*, 2, 1101-1106.
- Childs, J.M. (1979). The development of objective inflight performance assessment procedures. *Proc. 23rd HFES Mtg.* 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]. WPAFB, OH: AFHRL
- De Maio, J., Bell, H. H., & Brunderman, J. (1985). *Pilot-oriented performance measurement*. [AFHRL-TP-85-18]. Brooks AFB, TX: AFHRL.
- Gaidai, B.V. & Mel'nikov, E.V. (1985). Choosing an objective criterion for piloting performance in research on pilot training on aircraft and simulators. *Cybernetics and Computing Technology*, 3, 162-169.
- Gawron, V. J. (2000). *Human performance measures handbook*. Mahwah, NJ: Erlbaum.
- Gottman, J. M. (1981). *Time series analysis: A comprehensive introduction to social scientists*. Cambridge, UK: Cambridge University Press.
- Hills, J. W., & Eddowes, E. E. (1974). *Further development of automated GAT-1 performance measures*. [Rep 73-72]. Brooks AFB, TX: AFHRL.
- Hubbard, D.C. (1987). Inadequacy of root mean square error as a performance measure. *Proceedings of the 4<sup>th</sup> ISAP*. Columbus, OH: OSU.
- 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.
- 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*. 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: AFHRL.
- McDowell, E. D. (1978). *The development and evaluation of objective frequency domain based pilot performance measures in ASUPT*. Bollings AFB, DC: Air Force Office of Scientific Research.
- Mixon, T. R., & Moroney, W. F. (1982). An annotated bibliography of objective pilot performance measurements [NAVTRAEQPCEN-IH-330]. Orlando, FL: Naval Training Equipment Center.
- Rantanen, E.M., Johnson, N.R., & Talleur, D.A. (2004). *The effectiveness of a personal computer aviation training device, a flight training device, and an airplane in conducting instrument proficiency checks, Vol. 2: Objective pilot performance measures* (AHFD-04-16/FAA-04-6). Savoy, IL: AHFD
- 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.
- 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
- Simple, 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.
- 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. [NAVTRAEQUIPCEN 74-C-0063-1]. Orlando, FL: Human Factors Laboratory, Naval Training Equipment Center.
- Wickens, C. D. (2003). Pilot actions and tasks: Selections, execution and control. In P. S. Tsang and M. A. Vidulich (Eds.), *Principles and Practice of Aviation Psychology* (pp. 239-263). Mahwah, NJ: Lawrence Erlbaum.
- 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: AFHRL