

## TAXONOMIES OF MEASURES IN AIR TRAFFIC CONTROL RESEARCH

Esa M. Rantanen and Ashley Nunes  
Institute of Aviation, Aviation Human Factors Division  
University of Illinois at Urbana-Champaign  
Savoy, IL

Demands by airlines in recent years to move away from the constraints of the current air traffic control (ATC) system have prompted much research in two related avenues of human factors. The first is concerned with assessing the effects implementing advanced ATC concepts, such as Free Flight, on controller performance, whereas the second is concerned with determining how automated tools can help controllers ensure that this performance is not negatively affected by the implementation of such concepts themselves. In both instances, a systematic evaluation process is required to make such assessments. Such evaluations however, have traditionally proven to be an arduous task given the relative difficulty of selecting the appropriate performance measure to effectively answer the research question being addressed. This paper tackles this problem by providing an evolving taxonomy of ATC performance measures. The goal of providing this taxonomy is to help researchers more easily identify and apply the relevant measurement techniques during empirical investigation

### Introduction

All scientific research and subsequent engineering applications are dependent on methods of measurement (Chapanis, 1959). The need for reliable and valid measures is particularly compelling in the research and development (R&D) activities on a variety of aspects of air traffic control (ATC). Ever-increasing traffic volume and the collateral demand to improve aviation safety necessitate introduction of new technologies and automation applications in ATC. While potentially allowing the national Airspace System (NAS) to accommodate the increasing demand, these technologies will conceivably also have a fundamental impact on the system's functionality as well as air traffic controllers' working methods, workload, strategies, and performance (Hopkin, 1995; Wickens, Mavor, & McGee, 1997; Wickens, Mavor, Parasuraman, & McGee, 1998). Thorough assessment and evaluation of the consequences of new technologies on the system capacity and safety on one hand, and on the working conditions and performance of individual controllers on the other, is hence of utmost importance. Success of these evaluation efforts, however, is subject to the availability of valid and reliable measures.

Until recently, observation and subjective evaluations have been the primary sources of controller performance data from both operational and simulated ATC. Direct observation, or the Over-the-Shoulder (OTS) method, can be a valid and reliable method for performance measurement if a number of prerequisites are met (e.g., standardized checklists evaluator training, and verified inter-rater and intra-

rater validity). The OTS method, however, is labor intensive, time-consuming, and expensive. Moreover, a human evaluator may not be able to provide sufficiently accurate quantitative data for research purposes, due to the limitations of human observation capabilities. The latter is the case particularly in observation of simultaneous events. Another widely used class of ATC measures is subjective measures where controllers themselves rate their performance and workload. This method is relatively easy to use and inexpensive, and expert subject-evaluators can also readily detect and process task-related information that would otherwise require vast amounts of objective data to be recorded, stored, coded, reduced, and analyzed to yield useful measures (Wickens, Gordon, & Liu, 1998). However, evaluations by subject-raters suffer from many of the same problems as those by an outside observer, that is, they depend on the expertise and experience of the subject-raters and their ability to make absolute and comparative judgments. Furthermore, concurrent measurement is often very intrusive and interferes with the task at hand, and measures taken after the task are subject to changing perceptions and decay of memory.

Sound and valid arguments have been made both for (e.g., Hennessy, 1990) and against (e.g., Kosso, 1989; Scheffler, 1967) the use of subjective measurements. In the ATC domain, two particular arguments can be made in favor of objective measures. First, valid and reliable objective evaluation methods are particularly desirable both in conjunction with high-fidelity, realistic ATC simulation and in operational settings, where they can be collected and analyzed concurrently and unintrusively during the task, and

subjected to data mining techniques to detect trends in the system's performance, before any possible problems are manifested as incidents or operational errors. Furthermore, recent advances in area of digital technology and the ATC modernization efforts potentially make available new sources for data as well as data collection and storage methods. An example of access to data from which ATC measures can be derived is the System Activity Recordings (SAR) that stores all flight and radar information in Air Route Traffic Control Centers (ARTCCs). These data can be further processed by two specific computer programs, the Data Analysis and Reduction Tool (DART) (Federal Aviation Administration [FAA], 1993) and the National Track Analysis Program (NTAP) (FAA, 1991), which produce a number of text-based output files. These files can be further analyzed by specialized computer programs, such as the Performance and Objective Workload Evaluation Research (POWER) (Mills, Manning, & Pfeleiderer, 1999; Manning, Mills, Fox, & Pfeleiderer, 2000). The current POWER "suite" includes such measures as traffic count, control duration, and variability in aircraft headings, altitudes, and speeds, as well as latencies of handoff initiation and acceptance. A number of different controller activities are also recorded.

There is, however, a substantial gap between measures such as described above and measures of real interest, that is, measures of sector complexity and controller workload, situation awareness, and performance. It is therefore important to distinguish between what can be termed direct and indirect measures. The latter are based on primary measures but make inferences on variables that were not directly measurable (e.g., workload or performance). The problem is therefore not in availability of data, but in derivation of valid, reliable, and meaningful measures from the abundance of data. For example, Hadley, Guttman, and Stringer (1999) listed no less than 170 separate measures or measurement techniques in their air traffic control specialist (ATCS) performance measurement database. Many of these measures have substantial overlap, are derivatives of each other, and measure diverse aspects of ATC functions, all of which are not relevant to performance of an individual controller. The purpose of this paper is to propose a classification system for ATC measures that will allow for a cross reference between different types of measures, their purposes, and the required data, facilitating development of comprehensive models of ATC performance and additional measures as new sources of data become available.

## Taxonomy of Measures

There are several possible taxonomies for ATC measures. One is the dichotomy of measurement of system performance and the measurement of an individual controller or a team of controllers (Buckley, DeBaryshe, Hitchner, and Kohn, 1983; Hopkin, 1980). System measures are defined in system terms (i.e., capacity, throughput, delays, and channel occupancy times). Although they are greatly influenced by human performance, they are usually insufficient in the measurement of the performance of an individual controller. Identified task performance, human activity, errors, omissions, physiological and biochemical indices, and subjective assessment are possible measures of individual controller (Hopkin, 1995). Of these, task performance, activity, and possibly error measures could be derived from DART and NTAP data. Task performance measures compare the controller's output to that what is required in the task and encompass broad measures of errors and omissions. Human activity measures passively record what occurs in the task, such as radio transmissions, equipment usage, and communication and coordination with other sectors in terms of times, frequencies, and sequences of the activities. It is crucial to also distinguish between direct and indirect measures.

The main division of measures in the proposed taxonomy (see Figure 1) is hence between direct and indirect measures. Direct measures are defined here as those that can explicitly measured. Examples of such measures include a direct observation of a controller's action, measurement of a response latency, or count of aircraft in a sector at a given time. Indirect measures are those that cannot be measured directly but must be inferred from directly measurable variables. For example, certain actions of a controller may be indicative of his or her performance, response latency can be used to make inferences on some covert cognitive processes, and a number of aircraft in a sector can be used to signify sector complexity. We will return to indirect measures later; however, as such measures are derived from direct measures, it is important to lay a foundation for a systematic examination of these, which is the explicit goal of out taxonomy.

### Criteria

Within the class of direct measures, the next sublevel is created by differentiating between subjective and objective measures. The method for making this distinction is identification of objective criteria, a prerequisite for meaningful measurement (Meister,

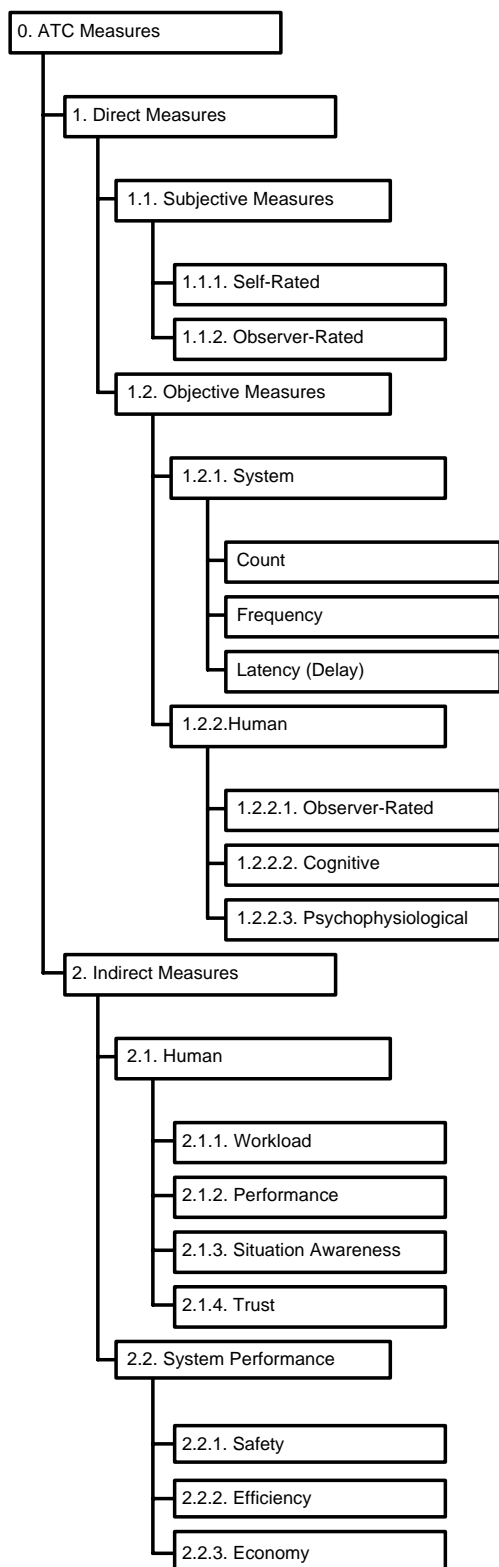


Figure 1. A proposed taxonomy of ATC measures; the top three levels.

1989). An example can be seen in Figure 1: Note that observer rating appears under both subjective and objective categories. The classification is here based on an objective criterion against which the observer bases his or her judgment. For example, an FAA OTS evaluation sheet (Form 3430) contains items such as “standard phraseology is not adhered to” and “awareness is not maintained.” The former has an objective criterion, the standard phraseology as published in the FAA Air Traffic Control Handbook (7110.65) and hence any deviations from it would warrant a check in the aforementioned box. However, the latter, while providing several guidelines for making this judgment, provides much latitude for a subjective assessment.

Another example of the importance of criteria in classification of measures is the differentiation between what is termed here primary and secondary measures. Primary measures are those that are measured directly, for example, count of aircraft in a sector, or number of heading changes per aircraft. Secondary measures are those derived from primary measures, for example an average number of traffic in a sector, its variance, or range. In the case of the average, the criteria are implicit (the time duration or interval during which the aircraft were counted and the number of samples) but nevertheless have an impact on the eventual measure. The role of a criterion is explicit in the case of error measures. The term “error” clearly presupposes some threshold value, or outcome of an action, that separates correct action from an incorrect or erroneous one. In other words, without explicit criteria no actions could be classified as errors (see Figure 2).

#### Measurement Scales

Measurement involves the assignment of a number system to represent the values of the variables of interest. There exist four distinct measurement scales—the nominal, ordinal, interval, and ratio scales—and the measured variables must be explicitly associated with the appropriate scale and its corresponding mathematical properties (Krantz, Luce, Suppes, & Tversky, 1971; Ghiselli, Campbell, & Zedeck, 1981).. The measurement scale is also a possible taxonomic criterion. Many of the direct measures are ratio measures. However, the measures derived from these are only ordinal at best. Hence, it is only possible to state that one sector is more complex than another, or that one controller experienced higher workload than another. The identification of measurement scales is also inexorably linked to development of criteria, as described above. This task is crucial in order to

ensure correct interpretation and proper statistical treatment of the data, and the success of subsequent modeling efforts.

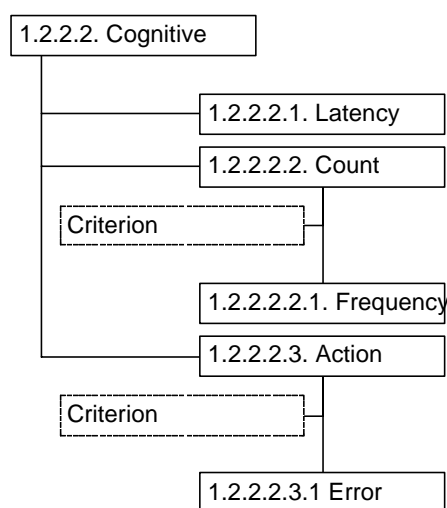


Figure 2. An example of the role of criteria in deriving secondary measures.

### Summary and Conclusion

The availability of data is likely to be enhanced with the implementation of new technologies in ATC (e.g., digital voice switch system) in TRACON and ARTCC facilities. This makes possible to derive new measures to complement and support the currently available and feasible measures. The multidimensional nature of ATC is evident from the numerous attempts to measure the various aspects of the system (e.g., system performance, controller workload, controller performance) and the difficulties encountered in these efforts. Furthermore, individual measures typically allow only a very narrow view to the behavior of the system or the human operator as a whole. Therefore, it is important to develop indices that capture the majority of the relevant variables and combine them in a meaningful and informative manner. This proposed taxonomy represents an emphatically systematic and comprehensive approach to the measurement problem in ATC. This approach serves a dual purpose: On one hand, it is essential to perform a thorough review of past and current research efforts and to organize the findings in a manner that facilitates the use of existing knowledge for a basis of future evolvement of ATC measurement. On the other hand, it is prudent to proceed cautiously on an issue as complex as ATC measurement and consider carefully all the

constraints, assumptions, and threats to validity that may emerge. In the face of the unprecedented challenges to the nation's air transportation system it is imperative to secure a "toolbox" of measures that would predict controller success in his or her task and the impact of changing procedures and advancing technology on the system as a whole. Eventually, it is hoped that this work would lead to development of measures that allow detection and prediction of performance decrement to levels that might lead to accidents or incidents before being manifested as such. The contrast here is between errors as a very coarse—and unwanted—measure of performance and more fine-grained and sensitive performance measures.

### References

- Buckley, E. P., DeBaryshe, B. D., Hitchner, N., & Kohn, P. (1983). *Methods and measurements in real-time air traffic control system simulation*. (DOT/FAA/CT- 83/26). Atlantic City, NJ: FAA Technical Center
- Chapanis, A. (1959). *Research techniques in human factors*. Baltimore, MD: Johns Hopkins University Press.
- Federal Aviation Administration (1991). *Multiple Virtual Storage (MVS); Subprogram Design Document; National Track Analysis Program (NTAP)*. NASP-9247-PO2. Washington, DC: Author.
- Federal Aviation Administration (1993). *Multiple Virtual Storage (MVS); User's Manual; Data Analysis and Reduction Tool (DART)*. NASP-9114-H04). Washington, DC: Author.
- Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco, CA: W. H. Freeman & Co.
- Hennessy, R. T. (1990). Practical human performance testing and evaluation. In H. R. Booyer (Ed.), *MANPRINT: An approach to systems integration* (pp. 433-479). New York: Van Nostrand Reinhold.
- Hadley, G. A., Guttman, J. A., & Stringer, P. G. (1999). *Air traffic control specialist performance measurement database*. (DOT/FAA/CT-TN99/77). Atlantic City, NJ: The Federal Aviation Administration.
- Hopkin, V. D. (1980). The measurement of the air traffic controller. *Human Factors*, 22(5), 547-560.
- Hopkin, V. D. (1995). *Human factor in air traffic control*. Bristol, PA: Taylor & Francis.
- Kosso, P. (1989). Science and objectivity. *Journal of Philosophy*, 86, 245-257.
- Krantz, D. H., Luce, R. D., Suppes, P., &

Tversky, A. (1971). *Foundations of measurement, volume 1: Additive and polynomial representations*. New York: Academic Press.

Manning, C. A., Mills, S. H., Fox, C. M., & Pfleiderer, E. (2000). Investigating the validity of performance and objective workload evaluation research (POWER). *Paper presented in the 3<sup>rd</sup> USA/Europe Air Traffic Management R & D Seminar*, Naples, Italy, June 13-16, 2000.

Meister, D. (1989). *Conceptual aspects of human factors*. Baltimore, MD: The Johns Hopkins University Press.

Mills, S. H., Manning, C., & Pfleiderer, E. M. (1999). Computing en route baseline measures with POWER. *Poster presented at the 10<sup>th</sup> International*

*Symposium on Aviation Psychology*, Columbus, OH, May 3-6, 1999.

Scheffler, I. (1967). *Science and subjectivity*. Indianapolis, IN: Bobbs-Merrill.

Wickens, C., Gordon, S., & Liu, Y. (1997). *An Introduction to Human Factors Engineering*. New York: Addison-Wesley Longman

Wickens, C. D., Mavor, M., & McGee, J. (1997). *Flight to the future: Human factors of air traffic control*. Washington, DC: National Academy of Science.

Wickens, C. D., Mavor, M., Parasuraman, R., & McGee, J. (1998). *The future of air traffic control*. Washington, DC: National Academy of Science.