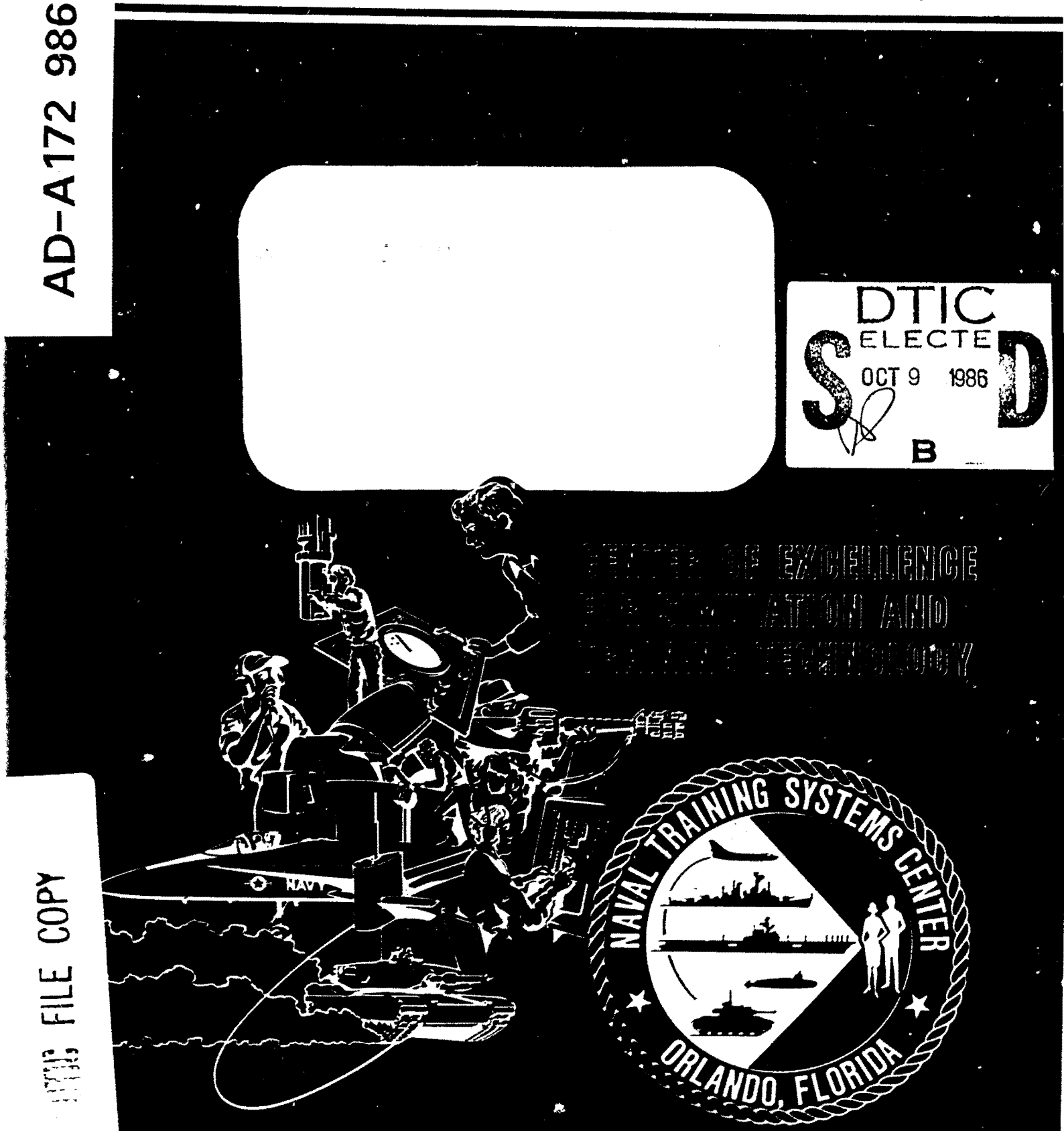
Naval Training Systems Center

(1)

AD-A172 986



DTIC
ELECTE
OCT 9 1986
S D
B

FILE COPY

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

86 10 8 152

NTSC TR-86-008

ISSUES IN PERFORMANCE MEASUREMENT FOR MILITARY AVIATION WITH APPLICATIONS TO AIR COMBAT MANEUVERING

NORMAN E. LANE

APRIL 1986

Prepared under
Contract DAAG29-81-D-0100, D.O. 1443
with
Battelle - Research Triangle Park Office
Research Triangle Park, NC 27709

Prepared for:

Naval Training Systems Center
Orlando, FL 32813
and
U. S. Army Research Office
Research Triangle Park, NC 27709

Prepared By:

Essex Corporation
Orlando, FL 32803

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) EOTR 86-3	5. MONITORING ORGANIZATION REPORT NUMBER(S) NTSC TR-86-008
6a. NAME OF PERFORMING ORGANIZATION Essex Corporation 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Naval Training Systems Center
6c. ADDRESS (City, State, and ZIP Code) 1040 Woodcock Road, Suite 227 Orlando, FL 32803	7b. ADDRESS (City, State, and ZIP Code) Orlando, FL 32813-7100 Attn: Code 711
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Naval Air Systems Command 8b. OFFICE SYMBOL (If applicable) 330J	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER DAAG29-81-D-0100, D. O. 1443
8c. ADDRESS (City, State, and ZIP Code) Washington, DC 20361	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 63733N; PROJECT NO. 4796; TASK NO. 2P1; WORK UNIT ACCESSION NO

11. TITLE (Include Security Classification)
Issues in Performance Measurement for Military Aviation with Applications to Air Combat Maneuvering

12. PERSONAL AUTHOR(S)
Norman E. Lane

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM 4/15/85 TO 4/4/86	1986 April 04	VIII + 143

16. SUPPLEMENTARY NOTATION This Task was performed under a Scientific Services Agreement with Battelle Columbus Laboratories, 200 Park Drive, P.O.Box 12297, Research Triangle Park, NC 27709. The task was requested and funded by the Monitoring Agency.

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	
05	08		Performance measurement, proficiency measurement, air combat maneuvering, construct validity, measures of effectiveness, reliability, system evaluation
05	09		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This report describes the history, development and current practice of measuring operator performance in systems, in particular military aviation systems, with additional emphasis on measurement in air combat maneuvering (ACM). The principal themes are that:

a) "Performance" is used interchangeably with "proficiency," and as such has acquired evaluative meanings about "goodness" or "badness" of individual capabilities. Performance "measures" thus require the same attention to measurement properties and validation as any other measures on individuals.

b) There are tendencies to substitute physical measures, which scale physical quantities or events, for behavioral measures, which are representative of how well ...

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL LCDR Michael G. Lilienthal	22b. TELEPHONE (Include Area Code) (305) 646-5130 22c. OFFICE SYMBOL Code 711

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

19. Abstract (cont'd)

an individual can perform a given task. Performance measures are behavioral measures, and acquire meaning through validation operations beyond those required for physical measurement. Not all the measures obtainable on individuals or systems are appropriately called performance measures.

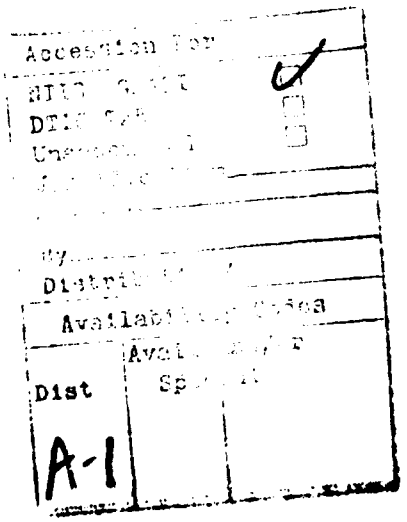
c) Task performance must be viewed as a "construct." Tasks are complex and multidimensional; individual proficiency must be inferred from limited observables. For proposed measures, it is necessary to show that measures are reliable, that they tap the most important components of successful performance, and that they are credible as representatives of individual task proficiency. Evidence supporting these demonstrations is acquired through the process of "construct validation."

Supporting these themes, the report synthesizes work in performance measurement since 1940, with detailed examination of lessons learned from the performance measurement programs during World War II and Korea. Trends in measurement interests since that time are outlined; changes in approaches resulting from improved data collection/reduction capabilities are described, in particular the more frequent reliance on automated algorithms. The logical and philosophical bases of measurement are reviewed, and characteristics of the skilled performance to be measured are defined and related to recent findings on skill acquisition.

Seven "criteria" for evaluating performance measures are defined and discussed, including Reliability, Validity, Sensitivity, Diagnosticity, and Utility and Value. Reliability is posited as the most basic concern; distinctions are made between accuracy and precision as properties of physical measures and stability/consistency of the behavior being measured as separate aspects of reliability. While accuracy of measurement has improved over the last 40 years, unreliability of behavior is inherent to the phenomenon; it is not reduced by better instrumentation. The rationale of construct validation is presented; it is argued that measure sets must possess certain minimum properties to justify use of the performance measure label. Discussions examine the importance of the process of task performance as well as task outcome or product, and the inadequacies of outcome variables as measures are noted. Procedures for decomposing a task into component processes are outlined. The use of process variables is described, both for validation of measures and for improving measure diagnosticity. Two variants of process measures, proxies and surrogates, are introduced along with their applications in special measurement situations.

Requirements from previous sections are focused on measurement in the special context of air combat maneuvering. ACM on simulators and ranges is briefly reviewed and related to previous ACM measurement literature. Problems of reliability and validity particularly critical in ACM are presented. There is a tendency in ACM developments to confound in a single measurement structure two distinct "constructs" of performance, one assessing overall proficiency of individuals, and one diagnostic of deficiencies on processes underlying the skill. These confusions of intended purpose complicate measure validation, and it is suggested that future developments reflect awareness of these distinctly different end uses.

TABLE OF CONTENTS



ABSTRACT

This report describes the history, development and current practice of measuring operator performance in systems, in particular military aviation systems, with an additional emphasis on measurement in air combat maneuvering. The principal themes are that:

a) The term "performance" is used interchangeably with the term "proficiency," and as such has acquired evaluative meanings about the "goodness" or "badness" of the capabilities or skills of individuals. Performance "measures" thus require the same attention to metric issues, particularly validation, as any other measures taken on individuals. Most measure development efforts have dealt only superficially with the properties of the numbers which they provide.

b) There are apparent confusions between physical measures, which represent the scaling of physical quantities or events, and behavioral measures, which are numbers representative of how well an individual can perform a given task. Performance measures are behavioral measures, and acquire their meaning through a set of validation operations that go well beyond those required for physical measurement. There are tendencies in military aviation to substitute physical measures for performance measures, with a serious potential for error in inferences about pilots or systems.

c) Because the tasks typical of aviation are so complex and multidimensional, and because the skills of individuals on those tasks must be inferred from a limited set of observables, task "performance" must be considered as a "construct." For a proposed measure set, it is necessary to demonstrate that the measures are dependable (reliable), that they tap most of the

important components of successful performance, and that they are credible as representatives of individual task proficiency. The evidence supporting these demonstrations is acquired through the process of "construct validation."

In support of these themes, the report synthesizes work in performance measurement and related areas since 1940, with a detailed examination of lessons learned from the large scale programs in performance measurement during World War II and Korea. Trends in measurement interest since that time are outlined. The changes in approaches resulting from improved data collection and reduction capabilities are described, in particular the more frequent reliance on automated algorithms for capturing physical measures. The logical and philosophical bases of measurement are reviewed; and the characteristics of "skilled" performance on complex tasks, the phenomenon to be measured, are defined and related to recent findings on skill acquisition.

Seven "criteria" or issues that must be considered in evaluating performance measures are defined and discussed -- Reliability, Validity, Sensitivity, Comprehensiveness, Separability of Operator/System Components, Diagnosticity, and Utility and Value. Reliability is posited as the most basic concern; distinctions are made between accuracy and precision as properties of physical measures and the stability or consistency of the behavior being measured as separate aspects of reliability. It is noted that while the accuracy of measurement has steadily improved over the last 40 years, the unreliability of behavior is inherent to the phenomenon; it is not reduced by better instrumentation but by an understanding of its ubiquity in the task being measured. The rationale of construct validation is presented, and it is argued that measure sets, to be labelled as "measures of performance," must possess certain minimum properties to justify use of the label.

Discussions illustrate the importance of examining the process of task performance as well as the outcome or product of the task, and the inadequacies of outcome variables as measures are defined. Procedures for decomposing a task into its component processes are outlined, and the use of process variables is described, both as a means of measure validation and as a mechanism for improving measure diagnosticity. Two variants of process measures, proxies and surrogates, are introduced, and their applications in special measurement situations are indicated. Performance measures as assessments of individual proficiency are contrasted to "measures of effectiveness" as system or unit level measures, and the appropriate uses of each are described.

The report concludes by addressing the problem of measurement in air combat maneuvering (ACM). The requirements for systematic measure development established by previous sections are focused on the special context of ACM tasks. The process of ACM performance on simulators and instrumented ranges is briefly reviewed and related to previous literature on ACM measurement systems. Some problems of reliability and validity particularly critical in ACM are presented. It is noted that there is a tendency in ACM measure developments to attempt to combine in a single measurement structure two distinct "constructs" of ACM performance, one oriented toward assessment of overall proficiency of individuals on the ACM task, and one intended for diagnosing deficiencies on component processes underlying the skill in order to improve overall performance. These confusions of intended purpose complicate measure validation, and it is suggested that future developments reflect an awareness of these distinctly different end uses.

INTRODUCTION AND BACKGROUND

PROBLEM

The lack of capability to measure job and training performance of military personnel is a chronic constraining factor on progress in a wide range of research and development areas. The performance measurement problem is remarkable for its longevity and persistence. Over 30 years ago, Ericksen (1952) noted that "...large numbers of research problems directed at improving methods of...training are becoming more and more severely bottlenecked by fundamental needs for improved methods of measuring pilot proficiency." Since that time, writers and reviewers have commented recurrently on the limitations imposed by lack of measurement systems. Gardlin and Sitterley (1972) describe the inhibiting influence on the systematic study of retention of skilled performance of the inability to measure that performance. Martin (1984), in a discussion of practice in aviation training, notes an inability to move students through training on the basis of measured proficiency in a stage rather than by fixed practice time. In one of the most recent summaries, Vreuls and Obermayer (1985) discuss technical issues of performance measurement that echo many of the same concerns expressed by Ericksen more than three decades previously and, before Ericksen, by the architects of the massive World War II aviation psychology research program (Crawford, et al., 1947; Flanagan, 1948; Miller, 1947a). While the sophistication of data collection instruments and the power to crunch numbers have improved by many orders of magnitude over the last 40 years, the basic core problems of assessing an operator's or maintainer's job proficiency or training progress still appear as crucial constraints.

REQUIREMENT

Good measures are required for a variety of important purposes in training and operational evaluations. Rusis, Spring and Atkinson (1971) present a list of major applications of measures. Similar lists with much the same content are given by a number of other authors, most recently Vreuls and Obermayer (1985). The Rusis et al. list (slightly paraphrased) suggests the following measurement applications in the training context:

- Determining the present proficiency or capability of an individual.
- Predicting the future performance of an individual.
- Diagnosing individual strengths and weaknesses.
- Qualification for advancement or movement to a later stage of training (minimum standards/quality control).
- Feedback to student or instructor about progress.
- Evaluating alternative training methods.

Most efforts at measurement address several of these requirements at once, although specific emphasis will vary as the assessment situation moves from one environment (such as training) to another (field or operational). The importance of basic "criteria" for goodness of a measurement system will likewise vary between training and operational settings. Measurement in operational and advanced training environments presents some unique (and challenging) problems in user acceptance and in the practicality, utility and cost of measures that go beyond the usual concern for those issues in earlier training stages. Some factors for evaluating measurement systems will be presented in a later section.

SCOPE AND EMPHASIS

By far the bulk of literature in performance measurement, particularly that dealing with "objective" or "automated" measuring systems, has been concerned with military aviation. There are equally compelling problems in industrial appraisal (a considerable literature, and most of the theory, is due to that research) and in non-aviation settings in the military. The heavier emphasis on aviation has come about because of the nature of aviation training and operations. High cost and safety hazards require a quality control on output not experienced in most other settings, and the leverage from improvement can justify a considerable research and investment cost. Discussions in the following sections, while drawing from all applicable sources, will emphasize military aviation as the primary arena of application. In particular, the focus will eventually be narrowed onto the more specific problems of air combat maneuvering (ACM) measurement as a special case of the principles developed from a broader view of assessment of job proficiency in the military.

OBJECTIVE

There is an extensive body of literature dealing with performance measurement issues or applying performance measures in a variety of contexts. Much of this literature is naive with respect to the operational measurement domain, and sheds little light on either the measurement problem in that setting or possible resolutions. Other studies have documented the problems of reliability and validity that have plagued the last 45 years of operational studies without grappling with how such problems might be prevented or reduced in impact. Only a limited set of studies have specifically addressed the methodological concerns involved in achieving satisfactory

measurement at the operational or near operational level. The primary objective of the present effort is to draw together what is known about the strengths and weaknesses of measurement methodologies from a pragmatic "lessons learned" viewpoint, integrated into a specific consideration of measurement needs in air combat and of the properties of measures required to fill those needs.

While a broad range of concepts are addressed in the following discussions, and the general background of major issues will be summarized as required, it is not the intent of this report to provide a primer on basic measurement concepts or on the nature of ACM. A number of such documents are available, more comprehensive and in greater depth than is possible in the present effort. The discussions which follow presume a general familiarity with test theory and measurement issues and with the activities involved in military air operations. Where necessary, citations to background sources for additional detail will be provided. Sources such as Allen and Yen (1979), Ebel (1979), Lord and Novick (1968) and Thorndike (1971) give an overall view of measurement theory. Smode, Hall and Meyer (1966) provide a comprehensive summary of work up to the time of their review and describe its relevance to aviation training. Rusis, Spring and Atkinson (1971) summarize over 1400 articles and reports prior to 1970. Youngling, Levine, Mocharnuk and Weston (1977) provide a large bibliography oriented to combat effectiveness measurement. Annotated bibliographies by Mixon (1982), Mixon and Moroney (1982), and Rehmann (1982) introduce the broader literature in applications of measurement to aviation, while the bibliography by Edwards et al. (1985) expands coverage to infantry and other military settings.

SOME HISTORY OF MEASUREMENT IN MILITARY AVIATION

WORLD WAR II AND KOREA

Early in World War II, the Army Air Force instituted a large-scale pilot selection program to improve the efficiency of a massively expanded training effort. Every imaginable variety of paper-and-pencil and personality testing and an extensive collection of apparatus tests were evaluated as predictors of pilot success. Selection approaches attempted and the outcomes of research are summarized in Flanagan's (1948) overview volume, as well as several others in the series. It became apparent early in these efforts that the pass/fail criterion against which tests were validated left much to be desired, both in metric properties and in sensitivity, and major programs were mounted to develop both improved subjective and new "objective" measures of proficiency in training and in operational flying.

This work was conducted over more than four years and involved hundreds of highly skilled professionals (see Youngling et al. [1977, p. 3-23] for a list of research group leaders). It still remains the largest and most systematic study of measurement and measurement problems ever carried out. Extremely large samples of students were available, and the rapid pace of training provided rapid data maturity, which in turn allowed a relatively quick cycle of development, testing and refinement of measurement ideas and approaches. Heavy emphasis was placed on the analysis and improvement of the reliability and predictive power of measures in both training and in combat environments. Findings of these efforts served as the basic structure for most later aviation measurement approaches, and, some 40 years later, are still a rich source of information on the problems of performance measurement in operational and near operational environments.

Research in the World War II Aviation Psychology Program (APP) was thoroughly documented. A series of published volumes and articles was issued in the years immediately following the war. Among those most germane to measurement issues are Army Air Force APP reports by Ben-Avi (1947), Carter (1947), Carter and Dudek (1947), Crawford et al. (1947), Cook (1947), Ericksen (1947), Flanagan (1948), Gleason (1947), Henneman (1946), Kemp and Johnson (1947), Lepley (1947), Miller (1947a, b, c), Thorndike (1947) and Youtz (1947), and descriptions of Navy work by Fiske (1947) and Jenkins, Ewart and Carroll (1950). These reports provide a number of findings highly relevant to the present effort that will be introduced as appropriate in later sections, particularly in discussions of reliability and validity.

Research of a similar nature to the World War II programs on measurement for pilot selection and training was re-initiated during the Korean War, although on a much smaller scale. Again, the primary focus was on development of "objective" methods, covering both the use of readily available system output variables (bomb drops, gunnery records, etc.) and improved ways of collecting and recording subjective evaluations of proficiency. Typical of work during this period were efforts by Wilcoxon, Johnson and Golan (1952), Dannieskold and Johnson (1954), and Smith, Flexman and Houston (1952), all dealing with structured checklists and forms for recording in-flight summaries of student activities, and by Hemphill and Sechrest (1952), who compared assessments from bombing records to peer and supervisor ratings of proficiency.

POST-KOREA TO LATE 1960'S

The middle to late 1950's and early 1960's saw an increased emphasis on systems engineering, a greater interest in human engineering as a part of system design (Van Cott & Altman,

1956), and the rise of the "Personnel Subsystem" and "Qualitative and Quantitative Personnel Requirements Information (QQPRI)" (Demaree, Marks, Smith & Snyder, 1962). In this context, the need for performance measures as system design and evaluation tools became even more apparent, and the systematization of available methodologies was emphasized. A number of guidance documents were prepared, describing systems-oriented approaches to measurement in weapon systems (Buckhout & Cotterman, 1963; Keenan, Parker & Lenzycki, 1965; Marks, 1961). Technology remained essentially that of earlier periods. Measurement was constrained to the use of available data, primarily observation and ratings of performance and the collection and manipulation of system output variables produced as a natural by-product of system operation.

During the same period, active interests developed in the dimensions and components of flight performance. Systematic factor analyses of flight variables (Fleishman & Ornstein, 1960; Wherry, Jr. & Waters, 1960) led to later proposals for "taxonomic" or dimension-based measurement approaches (Fleishman, 1967; Parker, 1967; Zavala, Locke, Van Cott & Fleishman, 1965). A full development of taxonomy anchors for understanding the nature of measurement variables was presented by Farina and Wheaton (1971). An outgrowth of these interests in multi-dimensionality assessment was the development of "synthetic tasks" (Alluisi, 1967; Morgan & Alluisi, 1972), constructed to contain sub-tasks or components representative of each of the major factors believed to be important in the class of task for which assessment was desired.

THE 1970'S

The computer technology of the 1960's had, by the early 1970's, brought about the availability of powerful, compact and affordable data recording and reduction. This capability

allowed the extraction from the system of data of higher potential relevance and greater refinement than that readily observable as system outputs, and led to the growth of several new areas of interest, the study of which had not been possible prior to advanced computer support for measurement methodology.

Automated Measurement Algorithms

The first of these new technologies was the development of "automated" machine-scoreable algorithms for capturing an operator's activities in a system and comparing them to standards or templates of desired performance. By the beginning of the 1970's, a number of efforts were underway to produce automated systems for machine recording and scoring of operator-system data. These typically involved one or both of two general approaches. The first approach was (and is) essentially an automated analog of criterion-referenced measurement. System status variables and actions taken are recorded at multiple points throughout a maneuver and compared algorithmically to "tolerance bands" or other predefined standards of "correctness" at each point. Work representative of this approach is described by Connelly, Schuler, Bourne and Knoop (1971), Knoop (1973), Knoop and Welde (1973), Baum, Smith and Goebel (1973), and Connelly, et al. (1974) for various aspects of a large-scale Air Force in-flight "grading" system, by Burgin and Fogel (1972) for a state-transition model of ACM based on payoff matrices, by Moore, et al. (1979) for a "Good Stick Index," and by Waag, Eddowes, Fuller and Fuller (1975), Waag and Knoop (1977), Fuller, Waag and Martin (1980) and De Maio, Bell and Brunderman (1983), for a performance measurement system development and evaluation on the Air Force Advanced Simulator for Pilot Training (ASPT).

With few exceptions, these efforts involve comparison of system variables to some pre-established objectives or

analytically-determined standards to derive performance measures. As in any criterion-referenced system, "good" performance is equated to doing the job in a prescribed way, according to some specified procedures or doctrine, and to demonstrating the capability to meet defined goals or objectives in self-contained segments of a task.

The second broad class of approach to automated algorithm development places less reliance on prior determination of standards and conformity to "doctrine," and is oriented more toward empirical derivations of combinations of variables and weighting systems to make up a performance score. Emphasis in this approach is heavy on relevance of measures to concepts of validity external to the operator-system variables themselves, such as sensitivity to experience and to changes under variation of task difficulty. In some efforts, it can be seen that performance is treated implicitly as a "construct" for which the physical measures that can be extracted from the system are "candidate" variables for use in assessment; "meaning" is attached to a measure by empirical linkages to other variables or factors. This idea of "performance as a construct" is a crucial one in measurement systems and will be treated in depth in a later section.

Most of the work using the "empirical" approach employs initial analyses similar to that of the criterion-referenced approach -- problem or task analysis, selection of maneuvers, maneuver segmentation, etc., but reduces data on some "criterion-relevant" basis. Representative efforts under this orientation include Vreuls and Obermayer (1971a), Hill and Goebel (1971), Vreuls, Obermayer, Goldstein and Lauber (1973), Obermayer et al. (1974), Kelly, et al. (1979), and Wooldridge et al. (1982). General discussions of steps in the approach and procedures used are given in Vreuls and Goldstein (1976). Semple, Cotton and Sullivan (1981, Chapter IV and Appendix B)

summarize the state-of-the-art in automated performance measurement systems as of the end of the 1970's, and give specific guidance for development of such systems on simulators.

Examining the Processes Underlying Performance

As an outgrowth of the same technology that allowed automated algorithm approaches to flourish, it also became technically feasible to examine the "process" by which an operator arrives at an end outcome or "product." One of the most intractable problems in early measurement efforts was an inability to deal with individual differences in the way a job or operational task was performed. Identical terminal outcomes on a task can be produced by quite different orderings of procedural activities which represent widely divergent skill levels and energy investments. This has been, and remains, a critical weakness in the use of "outcome" measures of proficiency. The capability to record the total activity of an operator allowed for much more "fine-grained" analysis and created the potential for real-time use of measurements both to provide feedback to the operator or trainee and to manipulate conditions and parameters of the task in a dynamic way.

The need to examine the processes underlying task performance was identified and addressed long before technology made possible their detailed study in complex tasks. Much of the field of skill acquisition and learning is concerned specifically with the procedures used and mental processes invoked by a given operator in performing a given task. This work has been directed almost exclusively toward the description and understanding of acquisition behavior, and has been only tangentially concerned with the measurement of a performance as an end goal. The analysis of acquisition processes in a military training context and some key literature bearing on its impact are summarized by Lane (1986). There were, in addition

to studies of skill acquisition, several systematic attempts to examine process variables for measurement purposes that predate the "coming of age" of detailed breakdowns of operator activities in the 1970's.

The Natural-Pilot Model -- It has long been recognized that the classes of highly skilled behaviors characteristic of military aviation are marked by certain common features. The tasks involved are multicomponent and heterogeneous in nature, requiring a combination of cognitive, motor and perceptual abilities. They are characterized by distinct differences between expert and less-expert operators in the manner in which a job is performed (see Lane [1986] and Schneider [1985] for more detailed expansions on these differences). Krendel and Bloom (1963) proposed an approach to measurement based on these task properties which they called the "natural pilot model." They defined as potential measurement factors three characteristics of the extremely proficient pilot: 1) Economy of effort (less energy and attention is required to achieve a given quality of performance); 2) Consistency (goal-related output is constant for many different conditions of input); and 3) Adaptability (automatic compensation for varying task conditions or reduced feedback).

In Krendel and Bloom's development, these factors are present to some extent in all skilled tasks, and can be identified from control activities and from system variables, particularly when the task conditions are systematically perturbed. A key point of their discussion is that observation of outcome variables or even intermediate criterion measures are insensitive to all these factors. Maintaining control of an aircraft within "tolerance bands" on a maneuver, for example, could result from hundreds of control inputs by a novice or only a few from a highly skilled pilot. The less-expert performer will "dither" with the controls, expending greater energy,

showing lower consistency, and focusing a significant portion of his attention on the task, creating less "reserve" for handling unanticipated events. Rather than focusing on the surface manifestations of "how well the job was performed," measurement systems should look for indicators of the three factors to obtain generalizable measures of true proficiency.

This early interest in "process" variables was well ahead of the technology required to exploit the model. Although Ryack and Krendel (1963) proposed and demonstrated some applications of the method on experimental tasks, the requirement of the method for detailed data far outstripped the capability during that period to record information for operational tasks. Parenthetically, there are some interesting tie-ins of the natural pilot model to a number of later technical developments. The authors' descriptions of skill acquisition are identical to those presently referred to as "automaticity." Their emphasis on "reserve" and spare capacity emerges later as a major concern of "workload" theorists.

Development of carrier landing measurements -- Examination of the processes by which the outcome of a complex task is produced is a time-consuming endeavor. In the middle 1960's, Brictson and his colleagues (e. g., Brictson, Hagen & Wulfeck, 1967; Brictson, Ciavarelli & Wulfeck, 1969; Brictson, Burger & Kennedy, 1971; Brictson, Burger and Wulfeck, 1973) undertook the development of measures for carrier landing performance. In their analysis, approach radar data for hundreds of successful and unsuccessful carrier "passes" were systematically decomposed into sequential "waypoints;" the status of each pass at each point was examined in relation to eventual outcome and compared to other performances within the samples. Data were obtained for both "novice" carrier aviators and highly experienced pilots. The significance of position at each waypoint in determining eventual success was calculated empirically. Among

other findings, it was determined that early waypoints were relatively unimportant in affecting success and that, consistent with most other studies, pilots tended not to "follow the book" while performing the task quite satisfactorily. The initial intent of the investigators to examine deviations from standard procedures as indicators of performance was of necessity replaced by a more empirical philosophy.

The importance of the Brictson et al. work to the present historical overview is twofold. First, the careful analysis of "process" underlying outcomes which they carried out required a large amount of data to provide stable analyses of a relatively simple outcome variable. Decomposition and analysis of the task involved a series of refinements and validations that took over 5 years to complete. Second, the effort was the first large-scale search for measures of a complex performance which departed from a "criterion-referenced" framework into a determination of what processes actually mattered in achieving a desired task outcome.

Adaptive measurement -- A natural by-product of interest in the variables and processes underlying skilled performance was the emergence of the concept of adaptive training (Kelley, 1969). If some conditions, factors or parameters of the task environment are more important than others in affecting successful performance, a greater focus in training on those factors should produce greater skill for a given training investment. Dynamic variation of task parameters presupposes the capability to measure those parameters on an ongoing, real-time basis. Further, the values of parameters at which a trainee or operator could produce consistently successful performance could serve as good performance measures in and of themselves (Kelley & Prosin, 1968, 1969; Kelley & Kelley, 1970; Matheny, 1969). Adaptive training and adaptive measurement require an advanced capability to record and reduce system

status variables and, like other "process-oriented" approaches, needed the technology of the 1970's for study and implementation.

Adaptive training/measurement, at its outset, appeared to offer a powerful alternative to conventional training in its potential to challenge the highly skilled while not overwhelming the less able trainee, particularly in simulator training situations for which careful control over conditions could be maintained. Vreuls and Obermayer (1971b) and a series of studies by Charles and his colleagues (Charles & Johnson [1971] is representative) showed good results for adaptive approaches to training and to skill measurement. Conway and Norman (1974) described the advantages of adaptive training and suggested future areas of application. For reasons which may be due more to the inertia of the military training system than to the merits of the adaptive technique, approaches using an adaptive framework have been little used since that time.

The ability to study in depth the processes involved in producing an outcome or terminal behavior (and the interest of investigators in doing so) represent, in the author's judgment, the most important development in performance measurement over the last two decades. As later discussions will show, outcome measures tend to be insensitive (and frequently inappropriate) indicators of the true capability of an operator or trainee. Measurement approaches should a) address the manner in which outcomes are arrived at and b) quantify performance or ability on the task components which account for variance in those outcomes. Without such a structure, measures will be neither diagnostic of performance difficulties nor useful in estimating the robustness of an individual's performance under the "other than standard" conditions which are invariably encountered in the operational environment. Later sections will expand on these product/process distinctions and their implications for measurement.

THE 1980'S

Measurement efforts in the present decade have in general followed the trends of the late 1970's. There is a distinct absence of research on measurement methodology, particularly on development and evaluation of performance criteria, as compared to earlier decades (Zedeck & Cascio, 1984). This is particularly evident in defense-supported research and development efforts. In the most recent relevant review, that of Vreuls and Obermayer (1985), only one citation is post-1980 (that citation is from 1982, and likely represents work initiated in the late 1970's). In the search of Defense Technical Information Center (DTIC) records conducted for the present effort, publications from the 1980's almost without exception described work from the previous decade, ACM measurement developments, or measurement efforts that were predominantly oriented toward the assessment of operator workload.

While there is continuing interest in defense-related measurement in the 1980's, work tends to address highly specific and well-defined subject matter areas, such as ACM (this literature will be discussed in a later section). There are indications of trends toward the use of models for identification and assessment of key parameters in complex systems (Hawley, Howard & Martellaro, 1982), but these tendencies as measurement efforts distinct from general test and evaluation techniques are not yet well established.

THE NATURE OF MEASUREMENT

The "measurement of performance" involves the simultaneous consideration of philosophical, physical and behavioral issues. There has been an overwhelming tendency in the history of work referred to as "performance measurement" to bypass some of these considerations for reasons of expediency or from an unawareness of their nature. Measures from the physical domain are all too frequently assigned meaning in the behavioral domain by a leap of faith, without establishment of the logical links necessary to make such a mapping appropriate. As Kelley and Prosin (1969) pointed out, "...The more knowledgeable the investigator, the more formidable the measurement problem appears. Those unsophisticated ... gather measures by some available means, assuming that the variance in the scores they gather represents a meaningful variation in task performance. The experienced investigator knows that such...variance need not and frequently does not represent significant parameters...in a complex task."

The use of the term "performance measure" is inconsistently applied in description of measurement systems. It is the author's belief that most people assign to the word "performance" definite evaluative connotation, that is, it acquires meanings of "goodness" or "badness" as an indicator of skill that equates it with what might be more precisely considered as "proficiency." Because of these tendencies, discussions in the present effort will treat "performance," not in its less common and simpler meaning as a description of task-related behavior, but as equivalent to proficiency, with a requirement for numbers which arrange individuals from low to high on the basis of how well they perform a task.

The development of measures that can be legitimately considered as representative of some aspect of a human

operator's performance or proficiency should follow the same sorts of "rules of evidence" that characterize any scientific or quasi-scientific endeavor. This section will address some of these rules, examine what it means to "measure" performance, and discuss the requirements that should be met in order to ascribe meaning to measures. Discussions will focus on the nature of measurement process logic rather than the psychometric and statistical properties required of measures. The section immediately following will address those criteria for measures.

NUMBERS, DATA, INFORMATION AND MEASURES

Any man-machine system in the course of its operation can be described in terms of one or more sets of events, ongoing sequences of operator activity and system responses. If a series of "snapshots" or samples of these events are taken at selected moments in time, a variety of numbers can be obtained from the physical measurement of inputs, outputs, and other system status variables. These physical measures are precisely that; they have "meaning" only as they are defined by the scale on which they are measured (how much, how far, how fast). As Kelley and Prosin (1969) note, scales in the usual performance measuring operation are in physical terms (like magnitude), rather than in terms of any abilities that may be underlying the scale. The "numbers" are legitimate. They are also, if the physical operations are defined, legitimate "data" (future discussions will equate for convenience the terms data and physical measures). They can not yet, however, even be considered as "information," and are not in any sense measures of "performance," either of the system or of the operator.

If there is associated with each of the physical measures at each moment in time some indication of the "desired" or "correct" value of the measure, and/or the acceptable deviation from this "target value," an additional level of meaning is

attained. The discrepancies between actual and desired values represent the "goodness" or "badness" of the combined human-machine system in meeting the criteria defined by its objectives or goals. These allow for measures such as absolute or mean square error, time-in-tolerance, bomb miss distance, and so forth. By the addition of "correctness" information to the measure set, the deviations from criteria move beyond descriptions of physical events to provide potentially useful but incomplete "evaluative" information about the system/operator combination.

It is not uncommon in the literature to find measure sets in "deviation from criteria" or criterion-referenced form used as indicators of system or even operator performance. While these deviations are legitimate "information" about how well the system meets each criteria, they are not yet dependable measures of "performance" in any important sense. A considerable amount of additional "information" must be added to the measure set to convert measures of deviation to estimates of system performance, and even more information to legitimately represent the measures as representative of operator performance.

At such a point in performance measure development, there are still two major sources of uncertainty about measure properties.

a) The system variables on which measures and deviations are obtained may or may not be ones which are influential in bringing about the ultimate outcome of the system process. The variables selected on some a priori or face-relevance basis may not account for any major part of variation in system performance. While a set of variables may appear to represent important aspects of task content, it may in reality not matter at all for the purposes of the measurement effort what the values of those variables are. There are a number of ways in

which this apparent anomaly can come about. There may be, for the population involved, no non-error variation on the measures; the task may be too simple, the tolerance bands too wide or too narrow, or the measurement system too insensitive to detect small differences. This would occur, for instance, in the use of highly experienced pilots as subjects on a light-plane experiment, or in advanced training situations in which aircrewmen already perform as accurately as the system itself or the environment will allow. In such a case, there are essentially no "individual differences," and there is no component in the measures due to operator contribution to overall system performance. Many writers have noted that there must be real differences in skill within the population for skill measures to be logically meaningful. It is also possible that measures are "irrelevant" because operators do not perform the task "by the book," and the criteria from which deviations are taken are thus inappropriate. This issue will be discussed in some detail later.

It is important to note that the question of relevance or representativeness of the selected variable set and associated metrics cannot be resolved on the basis of information available within the measure set itself. The variable set must be mapped into some measure space using one or more of several other kinds of external information which can increase or decrease the "believability" of the measures.

b) A second important (and related) uncertainty concerns the confounding in measure development of operator vs. system contributions to overall system performance. As noted above, there are many ways in which selection of variables, metrics and tolerance bands can result in measures for which variation is totally attributable to variability in the system without any operator component; the converse (no system contribution to variance) is, of course, also possible, but is much less likely for a system of any complexity.

Depending on the intended usage of measures, the confounding of the two sources may or may not be inappropriate. If measures are to be used exclusively to evaluate a system's capability to meet specific design goals, and if the system meets those goals, it makes little difference if components of performance are separable. Kelley and Prosin (1969) describe an evaluation of tracking displays in which there were no differences among subjects (all were highly skilled), but large differences among displays, sufficient to clearly determine the most effective display. In such a case, all the variance is due to system effects. The purpose of the evaluation was to evaluate systems, not to measure proficiency, and the lack of operator contribution was unimportant. If, however, there had been discrepancies between the system's capabilities and what it was designed to do, the inability to isolate the separate contributions to performance shortfalls would be a key deficiency in the evaluation.

In the same vein, if the objective of measurement is to determine an operator's ability to use the system to accomplish specified outcomes, i. e., some form of "minimum standards" of proficiency, confounding is a less crucial problem, at least so long as each operator meets the standards. If not, it becomes important to know why, and the need for diagnosis imposes additional constraints both on the isolation of the operator's contribution to system outputs and on the demonstrated "relevance" of measures to ultimate outcomes. Measures used to make inferences about the capabilities of individuals, either for diagnosis of specific skill deficiencies or for any form of personnel decision, require the application of much more stringent "rules" that establish measure "validity."

Systematic progression in performance measure development from physical measures or "data" to "information" to "measures

of performance" is the basic structure underlying any properly formed measure set. It is difficult from the literature on aviation performance measurement to tell if the bulk of workers in the field a) have not understood the distinctions involved in attaching labels (meaning) to numbers, or b) have felt it necessary to ignore distinctions for reasons of convenience or expediency. In only a few cases has the "philosophy" of measurement been recognized and discussed in the context of human-machine systems. Although terminologies vary, each discussion shares certain key elements -- an emphasis on extracting meaning from the measurement set by systematic addition of information about the system "objectives" and by the mapping of data descriptive of physical events into one or more additional domains.

Leuba (1964) described the logical distinctions required to translate numbers from physical measures into measure sets useful for description of performance. Leuba identified three "universes" or sets involved in the measurement of performance. The Universe of Real Events contains phenomena or events that occur in a operator-machine system. These events generate the Universe of Numbers; numbers are symbols only, not measures, and have no meaning except as defined by the processes used to obtain them. The Universe of Theory provides the laws, logic systems, assumptions and external definitions required to use symbols for quantification. There is no direct path between "Numbers" and "Events;" they can only be linked by the separate set of information external to both. Quantification (measurement in our terminology) thus requires the mapping of the number space into the event space on the basis of theory. There are, as noted above, two aspects to theory. One would convert numbers to physical measures; a second set of theoretical linkages converts physical measures to behaviorally-based or performance measures.

Distinctions similar to those above, but with slight variations in terminology, are made by Vreuls, Obermayer, Wooldridge and Kelly (1985). In their paradigm, numbers move from information to measurement by the attachment of an objective -- measurement is information for a specific purpose, and corresponds closely to Leuba's "quantification." They add also the concept of "assessment," similar in meaning to previous usages of operational performance measurement or "figures of merit." In their terms, "...Assessment requires the use of many sources of information to determine the quality of performance for a particular purpose, such as goodness or badness of performance relative to criteria for training or operations."(p. 4). It is important to note that their definition specifically excludes from measurement, and includes under assessment, the "evaluative" component of the measure set; measurement would include operations which (for example) take deviations from desired parameter values, but would not incorporate the "validation" operations required to verify if the parameters employed were important components of performance.

The complexity of arriving at "assessments" defined as above has apparently discouraged such a final step in performance measurement. Semple, Cotton and Sullivan (1981), in their review of Automated Performance Measures (APM's) on simulators, note the need to distinguish between true APM's and systems for automated data collection and recording. They conclude that essentially all automated measurement capabilities in existing or near future simulators and devices "...are best described as performance monitoring or data collection systems," and that "outputs from parameter monitoring capabilities are frequently not used." (p. 76).

PERFORMANCE AS CONSTRUCT

Much of the lack of "good" measurement practice in the human performance field is likely due to both the practical

difficulties and the logical complexities involved in carrying out the necessary operations. There has been only limited treatment of measurement "theory" in the man-machine literature. Leuba's (1964) development of measures as numbers converted to meaningful form by the application of theory-anchored rules was the first visible attention in human engineering to the concept of measures of performance as what are referred to in other contexts as "constructs." "Construct validity" emerged in the testing literature in the middle 1950's (American Psychological Association, 1954; Cronbach & Meehl, 1955). That approach holds that labelling a measure is insufficient to justify its use in practical applications, that the nature of the behavior assessed by a labelled measure must be determined by reference to the largest possible variety of external information. The more rigorous basis for measure development which "construct validation" demands was slow to take root in the measurement of human performance in systems, and remains so to the present time.

Within military aviation, descriptions by Lane (1975) and by Waag and Knoop (1977) made explicit the need for conceptualizing operator performance as a construct for purposes of validation. Procedures for evaluating measures suggested by Waag and Knoop and by Breidenbach, Ciavarelli, Sievers and Lilienthal (1985) come closest in the aircrew performance literature to the spirit of "performance as construct," presenting a sequence of steps involved in determining the credibility of candidate "performance" measures. Waag and Knoop suggested that first the content of measure sets should be established as directly relevant to the objective of measurement, that is, the measures should make sense on their face given the use that will be made of measurement outcomes. They then enumerate a series of tests which determine empirically the relation of data in the measure set to other external indicants of performance. Among these empirical requirements are that measures intended to reflect

differences in proficiency among individuals should vary as a function of overall task experience, should improve with increased practice, and should be consistent with other independently obtained estimates of performance such as ratings. The Breidenbach et al. procedures are quite similar, and will be elaborated in a later section.

The ways of examining measure "validity" suggested the above authors are conceptually identical to the construct validation process described in the general testing and measurement literature (Guion, 1974; Nunnally, 1967; Messick, 1975, 1980, 1981). It is not surprising that these specifications for measurement are encountered in the operational measurement literature, but it is discouraging that they are encountered so seldom. With the exception of the authors cited in this section, there is little apparent interest in either theory or application of the "rules" for measurement. As Semple et al. (1981) pointed out, virtually the entire field of APM in simulators treats data recording and reduction as equivalent to performance measurement without any consideration of measure relevance. Dickman's (1982) overview of APM, for example, provides an excellent discussion of the state-of-the-art in parameter recording, reduction and playback in APM's, but completely omits any mention of how and why the parameter values might be useful or important.

PHYSICAL AND BEHAVIORAL MEASUREMENT

To a significant extent, the seeming inability or unwillingness of aviation performance "measurement" efforts (particularly those described as APM systems) to deal with measurement issues and requirements may be due to the inherent duality of the human-machine discipline. Both training equipment design and human factors engineering are hybrids of engineering and behavioral technologies. It was noted

previously that the instrumentation and computer capability to record and process detailed data about system events triggered a number of such recording and processing efforts. Many of these efforts, uncertain about what mattered, measured far more than was needed or interpretable. As Vreuls and Obermayer (1971a) pointed out early in these developments, "...everything that moves should not be measured."

It is important in understanding the work of the last 15 years to reinforce the distinctions noted above between physical measurement and behavioral measurement. Measurement in an engineering or physical sciences sense consists of assignment of numbers to physical events or phenomena at a moment in time; the principal concern is whether or not numbers are sufficiently "precise" (the yardstick has enough gradations) and "accurate" (the yardstick is a "true" one and properly scaled throughout its length). Numbers so obtained are, in the context of their own definitional rules, proper and "valid" measures. They have the meaning defined by the processes used to derive them, and this is logically sufficient for good physical measurement.

Measurement is viewed quite differently from a behavioral standpoint. The assignment of numbers to attributes describing characteristics of people or people-system combinations involves issues which exist in a domain quite distinct from that of physical measures. The questions asked about yardsticks are different. Is the measure "consistent" (will two users of the yardstick get the same results)? Is the measure "reliable" (will the yardstick give the same outcome the next time it is used)? Is the phenomenon stable across time (is a different yardstick needed every day)? Is the measure "valid" (does the yardstick measure the attribute suggested by its label)? What is the "dimensionality" of the measure (does the yardstick measure many attributes combined into a single scale)? In order to claim proper and "valid" measurement in the behavioral sense, these and similar questions must be addressed.

While the complexities of these issues have been of critical concern to the personnel-oriented behavioral disciplines for more than 30 years, the measurement rules which they engender have not yet come into common use in the performance measurement arena. The term "measurement" has become so varied in meaning in its different performance-related applications as to be nearly useless for description of the objective of a development effort. Much of this confusion is due to the differential requirements for the properties of measures imposed by the physical and behavioral domains. The representation of physical quantifications as measures says in effect, "Here are some dependable numbers which describe properties of the system performing its task. You may do with them as you wish." The behavioral rules would respond, "These numbers have no meaning for assessing people. I don't know if any of them represent important, stable or useful properties of the system or the operator. There are many questions that must be answered before you can legitimately label those numbers as performance measures." It is important to keep in mind these two viewpoints on measure development in later discussions on performance issues and on the use of measurement systems (however defined) for making decisions in training and operational settings.

THE NATURE OF PROFICIENT PERFORMANCE

It is obviously the ultimate objective of all performance measurement systems to quantify or assess in some way how "good" or "proficient" an individual is at performing a given task. The job of the military aviator includes tasks quite different from the routine and well-ordered activities performed in the bulk of human work efforts. To measure effectively the performance on such tasks, it is necessary to understand how high and low levels of proficiency are manifested in performance variables. The preceding section described in general terms the events and rules of the measurement process; this section addresses the nature of the phenomena being measured.

Over the last five to ten years, there has been a rapid growth of interest in the nature of highly skilled behavior and how proficiency on complex tasks is acquired through practice. There is an explicit recognition that tasks that require certain classes of "skilled performance" are qualitatively distinct from other tasks, and need to be studied and addressed in special ways. The "ground rules" for recognition of highly skilled tasks are being more clearly defined. Anderson (1982) describes characteristics of tasks which require extended practice for proficiency. Schneider (1985) lists a series of criteria useful in identifying what he calls "high performance skills." Training for such skills requires certain specific approaches that are not necessarily useful for training skills of lesser complexity. Likewise, the measurement of performance on "high performance skills" is likely to require special attention to both qualitative and quantitative cues that might aid in recognizing good performance.

CHARACTERISTICS OF PROFICIENT BEHAVIOR

An earlier discussion described the Natural Pilot Model presented by Krendel and Bloom (1963). They noted that performances characteristic of highly proficient aviators show three basic properties: Economy of effort, consistency, and adaptability. These characteristics have been expanded and supplemented by other writers. Spears (1983) proposed the concept of "robustness" of performance, the resistance of performance to disruption by changing conditions. Fuller, Waag and Martin (1980) presented four attributes of proficiency that resemble those above. They suggest that superior flying performance (of maneuvers in a simulator) is reflected by:

a) Keeping certain critical aircraft state parameters (airspeed, altitude, etc.) close to defined criterion values.

b) Executing maneuvers smoothly by avoiding excessive rate and acceleration changes.

c) Accomplishing the above objectives with the least amount of effort (minimum number of control inputs).

d) Not exceeding procedural or safety limitations while performing maneuvers.

Conformance to Doctrine

There is in the Fuller et al. development, and in a similar description by Waag et al. (1975), an implication that "doing it by the book" is a key component of good performance. While it is true that performances which correspond closely to prescribed doctrine are likely to be at least acceptable if not superior, proficiency in the broad sense is not the same as conformance to doctrine. While trainees or students in the initial phases of

learning a task may attempt to approximate the recommended procedures, this becomes progressively less true as experience increases. A number of authors (Connelly, 1982; Knoop & Welde, 1973; Krendel & Bloom, 1963; Spears, 1983; Wilcoxon, et al., 1952) have noted the tendencies of experienced aviators to deviate from "the book" and to use widely variant strategies while showing apparently equivalent and satisfactory task performance.

Deviation from doctrine is also insufficient for a broad class of important aviation tasks that involve responding to changing task conditions which are initially unspecifiable or to an adversary which has degrees of freedom in its actions. For such tasks, there is no "book." There are many possible combinations of actions and reactions, and judgment is a key variable in achieving successful outcomes. Air combat maneuvering is a classic example of such a reactive task. An engagement may involve parts of many different standard maneuvers (as well as a few non-standard ones), with the choice of "next action" conditional on the behavior of the adversary aircraft.

Stability of the Phenomenon

Attempting to measure proficiency on a task or skill implies that the nature of the performance is relatively fixed, i. e., that it is an enduring characteristic of the individual being measured. This is almost never the case with students, and likely not true even for moderately experienced aviators. One of the characteristics of skilled performance is the relatively long periods of practice required for initial competence and the continuing gradual improvement over many hundreds or even thousands of task exposures (Lane, 1986; Newell & Rosenbloom, 1981; Schneider, 1985). This creates complications for assessment of performance, particularly in the less experienced

aviator group for which measurement is likely to be most important. Not only are there large day-to-day variations in a student's performance, but the "skill" being measured is itself changing over time.

Along with extended practice curves, skilled-performance tasks are characterized by the instability of initial performance and the presence of large individual differences in rate and shape of the acquisition curve. Some students learn faster than others and maintain superiority; some start more slowly and attain competence only after additional practice; others start rapidly, level off and are eventually overtaken by the slower learner. Bittner et al. (1984) show the wide range of individual differences encountered in more than 100 tests, and further indicate that differences in individual acquisition rates are more prevalent in some measures than others. Skill, as Bilodeau and Bilodeau (1961) point out, is a within-subjects phenomenon. Only after performance has stabilized and is in the neighborhood of an asymptote can proficiency of an individual be dependably measured. Measures prior to that point are more properly measures of progress, and are not necessarily good predictors of ultimate proficiency for a given individual.

These tendencies for day-to-day variation in individual performance are thoroughly documented in the World War II studies as well as later efforts. In a series of independent studies of proficiency on student navigators, bombardiers, fighter pilots, and multi-engine pilots, assessing a variety of skills (weapons delivery, gunnery, navigation), there was one consistent finding -- the day-to-day performance of students was too unstable to be measured reliably by any single measure or any one method. As Smode, Hall and Meyer (1966) and Dannieskold and Johnson (1954) concluded, the low reliability of measures was not the result of deficiencies in the measurement system but of the instability of the phenomenon being measured. The

implications of this problem for measurement during training are obvious and wide-ranging in considering ways to assess an individual trainee's progress. As Miller (1947a) noted, measures of transient phenomena are likely to be trivial unless there is a use for data on how an individual is performing at a specific given moment in time, and is only of value in comparison to normative standards based on previous "successful" students at equivalent points in training.

Changes in the nature of skill

A further difficulty in quantifying the performance of student or novice aviators is that the components of the skill being measured are likely shifting across practice. It is well established that correlations between successive measurements during learning tend to decrease systematically as the number of intervening measurements increase; the same decreasing relationships are seen with ability measurements taken immediately prior to training (Woodrow, 1938; Jones, 1959; and many others). There was for a time an active controversy as to whether this outcome was due a) to changes in the task structure that resulted in different abilities being required for successful performance in early vs. late practice, b) to abilities improving at differential rates with practice while the task structure remained fixed, or c) to changes in both tasks and abilities. Lane (1986) describes the history of this controversy. Alvares and Hulin (1973) demonstrated that it was logically difficult and empirically impossible to distinguish between (a) and (b), and that the truth was likely somewhere between the two extreme positions.

From the standpoint of performance measurement on students or novice aviators entering a new phase of training, the implications of changes in task components or in student abilities are considerable. Either position results in

instability of measurement. If tasks are changing, the phenomenon being tapped is not the same for successive measurements and cannot yield measures that are related across time. For changing abilities, it is those abilities themselves which contribute to the performance, creating an equivalent outcome -- measures that cannot be dependably generalized forward in time. Such restrictions on measures obtained during early phases of learning a new skill must be considered in interpretations of all proficiency measures on other than individuals who are well practiced in performing the task being assessed.

The Dimensionality of Performance

Another characteristic of proficient performance is that it tends to be multidimensional, while measures of performance are frequently unidimensional. There are many different components involved in performance of a complex task, involving perceptual skill, motor skill, planning and, particularly in aviation, ability to make rapid and accurate decisions. No one of these components constitutes the full spectrum of proficient performance.

Ghiselli (1956) identified three aspects of measure dimensionality. Static dimensionality denotes that, at any one point in time, performance can be evaluated on several task dimensions; dynamic dimensionality, that the dimensions important to success change across time; individual dimensionality, that individuals judged equally "effective" at doing the task differ in the components of the task emphasized to achieve results. These separate measures of component skills do not always combine readily into a global assessment. Crawford et al. (1947), in an analysis of multiple performance estimates for fighter pilots, concluded that there was no single measure that served adequately as an index of proficiency, but

also reported that combinations of separate measures yielded little improvement in relationships to other variables of interest. Human beings are extremely adept at compensating for a relative weakness in one skill or ability by devising strategies which rely more heavily on a skill in which they are relatively strong (Ghiselli's individual dimensionality). Certain minimum levels of proficiency on individual skill components are likely required for a given task (analogous to cutoff scores); above those minimums, skill mixtures that vary widely will result in apparently identical successful outcomes.

Whether and how to combine separate components of a multidimensional performance has been the subject of debate in the measurement community. Toops (1944) strongly supported the importance of an overall composite measure of job performance. Nagle (1953) discusses the advantages and disadvantages of each and describes a number of ways of constructing a single measure. Ronan and Prien (1966; 1971) also discuss the controversy and the arguments for each approach. Guion (1961) advocates the use of multiple performance criteria, but not their combination into a single composite. Dunnette (1963) maintains that a) the world of performance is inherently multidimensional and should be accepted as such; b) there is no such thing as the criterion, and c) composite measures, despite their convenience and appeal, are unwarranted. Among the principal objections of these authors and others summarized by Dunnette and Borman (1979) is the essential lack of interpretability attached to any composite measure and the arbitrariness of scale for some potential composites (how many speed units is one accuracy unit worth). In their judgment, composite measures are prime candidates for the criticism by Wherry (1957) of many measurement efforts, "....We don't know what we are doing, but we are doing it very carefully, and hope you are pleased with our unintelligent diligence" (p. 1).

Problems of skill instability and measure dimensionality have, as noted previously, major implications for selecting and interpreting the results of measurement approaches. These implications and associated data are discussed in an expanded context in the next section, dealing with the fundamental properties required of numbers in order for them to be evaluated as "good" or "bad" measures.

CRITICAL ISSUES IN EVALUATING PERFORMANCE MEASURES

Discussions in several previous sections have described the logical and "philosophical" issues involved in translating physical measures or "data" into behavioral measures or "information" about performance. Along with those requirements which deal with measure logic, there is an additional set of requirements for measures to meet which are concerned with both the metric properties and the value of the measures in use. These latter issues serve in a sense as "criteria" for measures, questions which must be asked of any set of numbers in order to support or refute the case that those numbers are "proper," "valid" and "useful" measures of performance.

There have been more several dozen different "criteria" suggested by various writers on measurement issues. The present effort adapts, combines, borrows from and adds to those lists, in most part without credit for origin. The seven resultant criteria are identified briefly below, followed by a separate discussion for each. The "short titles" for criteria are in general the traditional ones used in reference to measurement properties, although some require considerable qualification and expansion in the context of aviation measurement, particularly that dealing with operational or near-operational conditions. It is also apparent from the titles that there is considerable overlap in the topics; this redundancy will also be encountered in discussions. Some factors affect nearly all of the criteria, and evidence for one will frequently reinforce (or weaken) another.

- Reliability.
- Validity.
- Sensitivity.

- Completeness (Dimensionality, Comprehensiveness).
- Separability of Operator from Measurement Context Contributions.
- Diagnosticity (Specificity).
- Utility and Cost Benefit (Value against Alternatives).

RELIABILITY

Reliability of measures is considered first because it is in the metric sense the most basic issue. If measures are not dependably replicable over the required time period, other criteria are of little or no importance. Validity, the other crucial cornerstone of measurements, cannot even be examined in the absence of acceptable measure reliability, since unreliable measurement precludes the empirical investigation and demonstration of measure validity, however defined.

Sources of Unreliability

Like many labelled concepts, particularly those of behavioral disciplines, the term "reliability" has too great a burden of different connotations for a single word to carry. The term is used here in a general sense because it is one that people are accustomed to seeing, only sometimes recognizing that it is employed with so many varying meanings.

Ronan and Prien (1971) distinguish between two main sources of "unreliability": a) the reliability of the performance itself (the phenomenon being measured), and b) the reliability of the observation of that performance (including numbers obtained from both subjective and objective methods). This is a

crucial clarifying distinction. The first class is sometimes referenced with terms such as stability, repeatability, consistency and generalizability of the measures. The second includes concerns for factors such as accuracy and precision. From the standpoint of measure use, the "reliability" of the behavior is most important; measures of an unstable phenomenon have no value beyond the moment in time at which they are obtained, regardless of the accuracy with which they are measured.

Physical measures -- Earlier discussions contrasted physical and behavioral measures as requiring basically different kinds of verification. For physical quantities, the essence of the test for reliable measurement is the extent to which the phenomenon is captured precisely enough (enough yardstick gradations) and with an acceptably small absolute error of measurement (satisfactory yardstick calibration) at a given moment in time. The same is true of observer ratings or subjective assignment of numbers. To what extent are "observers" using the same yardstick, and are they reading it without imposing constant errors or changes in scale?

Behavioral measures -- For those measures considered to be quantifications of behavior, verification of reliability requires additional consideration of whether the behavior being measured really "exists." It should not be transient, subject to either systematic or random fluctuation over time. Skills which are not yet "learned" tend to be highly unstable because the behaviors which are being measured are not yet well defined within the individual, and unreliability is an inherent and unavoidable consequence of that characteristic. Behaviors which are not stable cannot be measured reliably in a single limited time frame. Fluctuations due to learning of the skill, although apparently random, are basically systematic; over time, the general trend of performance for a learned behavior is toward

improvement. For complex skills, this trend will not necessarily be revealed in any short sequence of repeated measurement; performance of a given individual may rise or fall between successive measurements, with only an upward shift of group averages to indicate the presence of learning. Thorndike (1949) refers to the effect of such individual inconsistency in behavior as intrinsic unreliability.

An effect indistinguishable from that due to learning fluctuation occurs, for very different reasons, in measures which show random fluctuation because of uncontrolled changes in task performance conditions (weather, equipment differences, different observers, etc.). If the environment in which measures are taken changes randomly across successive measurements, the measures will be unreliable regardless of the stability of individual performances. If the changes in environment are systematic, that is, they cause individual performances to vary in the same way, measures will appear to be reliable on statistical grounds, but values will be artifactually high. They will not be measures of behavior, but of system effects independent of individual contribution, and will be inappropriate for use on individuals. Thorndike (1949) refers to such influences external to individual behaviors as extrinsic unreliability.

There is an extensive literature on the analysis of reliability in aviation performance measures, dealing with both the stability of skilled behavior and the effects of fluctuating task environments.

Evidence on "Reliability" of Measures

Reliability of behavior -- Miller (1947c) summarized the findings of the extensive World War II research into performance reliability as follows: "It has been possible to make measures

objective enough,...only to find that the chief source of difficulty was not in errors of measurement but in erratic day-to-day fluctuations of performance."

There are two important points that should be noted about Miller's summary comment. First, he is clearly distinguishing between the operations of the measurement process (the properties of the yardstick) and the nature of the performance being measured. Secondly, additional examination of the literature which he summarized indicates that the finding of day-to-day variations in the performance itself was present in every study. In virtually every situation in which reliability was analyzed, the obtained coefficients for single performance measures were so low as to make their use for any purpose problematic. This finding of skill inconsistency applied to both "subjective" and "objective" measures alike, and in every variety of aviation-related skill for which measurement was attempted. Consistently, reliability values obtained on a single day (within-mission) were described as "moderate" or "satisfactory"; those obtained across missions or days were "low" or "near-zero."

"Objective" measures --Carter and Dudek (1947) reported within-mission reliabilities for objective measures of navigation performance of .48, but between-mission reliability was essentially zero. It should be noted for this and other studies using "objective" measures (deviation from desired value) that both the two effects on unreliability identified above -- behavior instability and changing task conditions -- were affecting between-mission outcomes. In most reports, efforts were made to isolate and treat the two sources separately. Carter and Dudek, for example, found that while initial within-mission reliability was .77, much of this was spurious, attributable to system effects on performance. While odd-even scores in the same seat for the same individual

correlated .77, scores in the same seat for different individuals correlated .30, and scores between successive missions for the same individuals in the same seat were essentially uncorrelated. In addition to the two sources of variation already discussed, the authors suggested also that the vast majority of navigators were sufficiently proficient that the small differences in performance among individual navigators were unimportant compared to other sources of error. This is a critical point, indicating a lack of sensitivity of the measures used (perhaps due to insufficiently difficult tasks or high group homogeneity of skill), and is one that will be revisited frequently in discussions of other measurement properties. The authors concluded that measures were so influenced by sources of variation that were not of interest in measurement as to be essentially useless as criteria for validation of training (the intended use).

Ericksen (1947) reports findings for multi-engine pilots identical to that of Carter and Dudek, reliabilities "fairly good" within-mission, "very low" across missions. Findings by Gleason (1947) on fixed gunnery performance, by Cook (1947) on gunnery and bombing performance, by Crawford et al. (1947) on fighter pilot gunnery, and by Kemp and Johnson (1947) on bombardier performance are all consistent with that pattern. Crawford et al., for example, report within-mission reliabilities ranging from .34 to .68 (the lower value is more representative), and between-missions reliabilities of about .27-.32. It is significant that the .32 value was one of the highest between-days reliability reported in any of the World War II studies, and required 1200-round samples of behavior to obtain. As with Carter and Dudek (1947), all of the above-cited writers also found it necessary to deal with the problems of isolating variations due to individual instability from those due to system or environment-induced fluctuations.

Beyond that of the World War II/Korea period, there is only limited information available about the reliabilities of operational or near-operational measures. In none of the more recent measure development efforts are the reliabilities of measure sets addressed. Data from a series of studies conducted over the last 8 years on the Naval Training System Center's Visual Technology Research Simulator (VTRS) indicate that the problem of measure instability is essentially unchanged over the last 40 years. In a number of simulator transfer studies with a range of aviation tasks (e. g., Lintern, Nelson, Sheppard, Westra & Kennedy, 1981), reliabilities of the performance measures obtained on the simulator were surprisingly and consistently low. Reliability of <u>averaged</u> approach scores in simulated carrier landings was approximately .38. This rather low value was <u>despite</u> extremely precise data recording and reduction systems (measurement error was virtually zero), the computation of each measure from a large number of data points, and the precise control of task conditions only possible in simulator measurement.

Westra (1981) reported similar but somewhat smaller correlations. In his study, glideslope deviation (the <u>average of four trial means)</u> had retest reliabilities of .23 and .32 in two different samples, and landing performance (touchdown accuracy score, also the average of four trial means) had reliability values of .20 and .29. Note that the values being correlated are <u>averages of averages</u> of single performances; the reliability of a single approach or landing would be very much smaller (no greater than .05).

Reliabilities for comparable landing measures obtained under field conditions are (somewhat surprisingly) in the same range as those from the simulator. Analysis by the present author of field carrier landing practice data taken from Westra et al. (1986) showed considerable variance in intertrial correlations

(likely as a function of changes in ability with practice). Several different but related measures of deviation from glideslope were obtained, each based on the mean scores for blocks of four trials (i. e., each score represented four approaches). Correlations between blocks ranged from .09 to .71 with a typical value of .30 to .35. Interestingly, scores were much more reliable for pilots having prior simulator training, with reliabilities for late practice between .62 and .70. These latter reliabilities are unusually high values, and illustrate the importance of skill stability in reliability of measurement. After extensive simulator training, individual pilots had become much more consistent in their performances, that is, the phenomenon being measured had stabilized. Comparable reliabilities for pilots without simulator training were between .18 and .30 and were actually somewhat lower in late practice than in early trials. Reliabilities from both simulator and field studies also reinforce the necessity of averaging single measures to obtain values less subject to trial-to-trial fluctuations, since many of the obtained measures were of marginal reliability even when scores were based on a large number of separate data points.

A similar analysis of air-to-ground weapons delivery scores taken from Lintern et al. (1985) showed much lower reliabilities than those for approach and landing, with values more typical of other field data findings. Measures in this study were based on circular error of impact in bomb and rocket delivery. Because of reliability problems in previous studies, each measure (a trial) was the average of impact error from eight drops. Despite this averaging, intertrial correlations were low, averaging only slightly more than .20. While this was sufficient, with sizable N, to draw conclusions about simulator-to-field transfer, measures at that level of reliability, even averaged ones, are clearly not appropriate for use in assessment of individual proficiency. Further, the reliability of a single

impact error measure (working backward from the Spearman-Brown equation) would be estimated at .03! This is consistent with World War II findings on bombing error of exactly .00 (Bingham, 1949) and Korean War findings (Hemphill & Sechrest, 1952) of ~.01. These values suggest that the use of such "relevant" outcome measures as impact error are extremely perilous. This issue will be examined in more depth in later discussions of process vs. product variables.

"Subjective" measures -- The tendencies for individual performances to be erratic across successive measurements, while supported by evidence from studies using "objective" measures, is even more clearly seen in efforts examining the reliability of subjective assessments. In many of the studies cited above, which report virtually no between-day reliability for objective measures, there are comparisons with subjective measures (usually ratings) of the same performances. In almost every case, the reliabilities of between-mission subjective measures were *larger than or* equivalent to the objective ones.

Crawford et al. (1947), for example, noted that the reliabilities of subjective criteria were "somewhat" higher. Carter (1947) reported similar findings. Later, Hemphill and Sechrest (1952) found the reliability of circular bombing error to be almost exactly zero, while peer ratings of bombing proficiency for the same exercises had a reliability of .91, and superior's ratings (based in part on bombing data) had reliabilities ranging from .61 to .95. Youtz (1947), using "objective/subjective" scoring (standardized check lists) for pilot trainees, found inter-observer reliabilities as high as .80-.90 for a given performance, but correlations between successive performances in the range of .30. (Note that this between-mission reliability exceeds nearly all those reported for objective measurements). Danneskiold and Johnson (1954) and Wilcoxon, et al. (1952), using a similar "objective check list"

technique, found day-to-day reliabilities in the same range as those reported by Youtz, with the "objective" measures (based on recording of specific aspects of performance) slightly inferior to those obtained from conventional "subjective" global judgments by instructors.

The consistent patterns in all the above studies are that: a) The day-to-day performance of an individual varies dramatically as a result of factors such as fatigue or changes in the way a task is performed, and b) the conditions under which the task is performed (differences in weather, equipment, etc.) account for far more variation in performance than do individual differences. The relative superiority of subjective measures across successive performances is readily explained. Observers providing summary judgments are to some extent taking into account the effects of varying task conditions, and are making judgments relative to conditions -- "That's pretty good, considering the weather." In Hemphill and Sechrest (1952), ratings by supervisors based in part on bombing scores were more reliable than the scores. (For numerous other reasons, this is not a recommended technique). Despite this compensation, none of the techniques, subjective or otherwise, produced satisfactory reliabilities for a single performance of a task. Another consistent finding was that, in order to achieve useful measures, it was necessary to combine measures across methods and across time. Youtz (1947), for example, reported that combining "instructor as recorder" measures with subjective judgments raised reliability from .30 to about .60 or higher. The sum of multiple checkrides based on the checklists had reliabilities near .80.

Reliability of observation -- Compared to the instabilities introduced in measures by the effects noted above, the contributions to unreliability attributable to "measurement error" are relatively minor. The sources of unreliability in

behavior described above are due to instabilities in the phenomenon being measured; measurement or observational error has to do with deviations of numbers assigned away from their "true" values at that specific moment in time. For instrumented recordings, unreliability is introduced by factors such as differences in calibration between successive measurements which add a random component to values, or by a precision insufficient to capture the required number of "significant digits" of performance. Note that absolute accuracy may not matter. Consistent calibration errors (a shift in scale) introduces "bias" but will not affect reliability unless output is being compared to fixed standards or desired values in a "deviation from criteria" model.

Although measurement error introduced by recording systems was a legitimate matter of concern in earlier measurement efforts, technologies presently available are sufficiently accurate and precise that error so defined need not be seriously considered in examining sources of unreliability. In most cases, inputs and outputs can be recorded with a level of precision that materially exceeds requirements, given the essential instability of the phenomenon being recorded.

Sources of "observation" error include both those arising from the instrumented recording system and those due to errors introduced by human observers. These operate on reliability in exactly the same way. Humans as "recorders," however, are likely to add considerably more error to the measure. Humans, for a variety of reasons, are probably incapable of recording events without some error resulting from inattention or physical limitations and some bias from preconceptions about the correctness of performance. When rendering summary judgments, each observer brings to the task a different template of correctness and a different scale of "goodness."

It is because of these perceived limitations of human recorders/observers that so much effort has been invested in the deriving of "objective" measurement systems. Some of the intensive efforts to "objectify" measures during World War II and Korea are described above. Smith, Flexman and Houston (1952) compared in-flight observation by instructors with derivation of comparable information from camera recording, and found camera data more subject to error, probably due to the relatively primitive equipment available. Horner, Radinsky and Fitzpatrick (1970) also used video recordings of student flights. Instructors graded proficiency based on recordings using a special scale. Reliabilities were described as "high." Prophet (1972) described 15 years of work in developing measures of helicopter pilot performance. He concluded that the reliability of measurement was materially improved by careful and systematic attention to the content of scales on which performance was rated, describing reliabilities of checkrides improved from values of .08-.09 into the .40-.50 range. He suggested that further improvement was attainable from automated recording, but (along with many other writers) noted that key components of performance (complex decision making and other cognitive skills) would be sacrificed by a total reliance on hardware for assessment.

Subjective vs. objective measurement -- There are several important caveats in the tendencies toward total reliance on "objective" techniques. There is among developers of measurement systems an equating of "objective" with "good," and "subjective" as "bad." Much of this perception results from a presumption that reliability must of necessity be increased by removing observers' contributions to "measurement error" (a bad thing), and replacing it with extraction of objective quantities which can be accurately recorded (a good thing). As many of the studies above show, however, this is not always so. Further, human observers and instrumentation are not necessarily

measuring the same performance attributes. The literature is clear that when humans assign values to whatever they measure, and instruments record the physical quantities they are designed to record, the resultant behavioral measure sets tend to be equally reliable. There is, however, an additional step in "objective" measuring which introduces a subjective element. As Muckler (1977) pointed out, all objective measurement involves subjective judgments. Decisions must be made about which physical measures will be recorded and, further, how they will be reduced and translated into behavioral or performance measures. This subjectivity cannot be avoided, and constitutes a distinct problem for automated measurement systems, since it brings into play issues of "relevance" and "validity" of the measure set. An abundance of literature suggests that observers' ratings of proficiency, while they suffer from problems of bias and scale, are in general keyed to detection of the appropriate aspects of performance, i. e., they tend to be reasonably relevant and valid. Their deficiencies (different templates of good performance, etc.) can ordinarily be overcome by pooling of judgments across observers and across time. It is far more difficult to "fix" the problems of objective measure sets. If the measures selected for inclusion are not the "right ones," there is a problem which Nagle (1953) called "criterion deficiency." There is insufficient overlap between the measure set and the theoretical "ultimate" performance, and no combination of irrelevant measures will yield relevant scores. It is thus possible, as Miller (1947a) observed, to make measures more reliable (as objective systems may do), only to find that they are less valid.

There are also, as many writers have noted (e. g., Prophet, 1972), aspects of performance that are extremely difficult to tap with objective approaches. Factors of planning, air-discipline and what is sometimes called "headwork" are not well reflected in the simple observation of inputs and outputs.

To the extent that these are significant components of a given task, their omission will affect the completeness of measurement and hence both reliability and relevance of measures on that task. Danneskiold and Johnson (1954) found that checklist-based measures were more reliable and had higher correlations with other measures when subjective judgments of such factors were combined with scores extracted from recorded observations.

Reliability summary I -- The above studies taken together present a picture of chronic low reliabilities for all classes of performance measures in aviation settings, both training and operational. The reliability problem appears to be equally present, although there is limited recent data, for measures obtained with both subjective and objective techniques. There is evidence that the more complex the skill for which measures are desired, the less likely it is that any single measure of performance will be sufficiently reliable to be useful. The problem is particularly acute for trainees or novices since the skills being measured have not yet fully developed, and proficiency is likely to show considerable variation between successive performances.

Evidence about measurement problems from work during World War II and Korea constitute the largest and most directly applicable source of information about the reliability characteristics of performance measures. Conclusions from that work are still valid. The problems documented during that period were not caused by a lack of precision/accuracy or other factors likely to be improved by modern methods, but rather were due to instabilities inherent in the phenomenon and the task environment. Obtaining reliable measures is likely to require the pooling of multiple estimates of the performance. As Miller (1947c) summarized the issue, to acquire stable and non-trivial measures of proficiency, it is necessary to repeat measures on different days in different airplanes (with different observers).

Measuring Reliability

Discussions about reliability have thus far treated coefficients as if they were comparable, regardless of the basis of their computation. This is clearly not so, since different ways of obtaining reliability coefficients make different assumptions about which sources of overall variation will be considered as reliable (non-error) and which will be treated as error, and will thus yield different values. Correlations of odd-trial with even-trial scores on a single mission, for example, excludes from reliable variance any momentary effects on performance unique to a trial, but considers as reliable variance a) any task condition effects that are mission-unique (visibility, weather, equipment characteristics) and b) all performance effects due to characteristics of the individual that remain in force throughout that specific mission (fatigue, variation in strategy, practice). Correlations between performances on successive days treats all mission- and day-specific components as error, and would yield lower reliability values.

Reliability coefficients -- There are three major types of reliability -- internal consistency, test-retest, and equivalent (alternate) forms -- derived from "classical" test theory. These conventional measures of reliability are not as clearly defined for aviation measurement operations as they are in conventional testing situations, and are sometimes of limited direct generalizability. All three coefficients estimate the same quantitity, the reliability defined as the ratio of "true score" variance to the total variance. As noted above, they differ primarily in terms of assumptions about the components contained in "true score" vs. error. Detailed descriptions of the theory underlying reliability definition and computation are provided in a number of previously referenced measurement texts, e. g., Lord and Novick (1968) and Allen and Yen (1979).

Succinct descriptions of reliability in the context of aviation measurement are given by Breidenbach, Ciavarelli and Sievers (1985).

There are not always direct analogs of each type of reliability in practical measurement situations. The operations involved in obtaining test-retest and equivalent forms on aircrew tasks, for example, are identical, since the only available "equivalent form" of an integral task is the task itself, performed again in a subsequent time frame. The usual distinctions between these two coefficients are thus unnecessary and of uncertain meaning in performance measurement development. In a similar vein, internal consistency reliabilities (odd-even, split-half) have a somewhat different meaning for most aircrew tasks than they do for tests. Internal consistency coefficients require the assumption that "items" being compared (the activities on odd and even trials, for example) are homogeneous. The dynamic nature of aircrew functions makes this assumption much less tenable in aviation performance measurement than in formal testing situations in which there are "items" whose properties can be examined and controlled. In many such real-world job settings, there may be no two periods of time in which the work performed involves identical inputs and operations.

In addition to some difficulties with the tenability of assumptions, reliability estimation methods based on classical test theory have other characteristics which can reduce their generalizability to operational or near-operational measurement conditions. All such methods are variance-based, i. e., they define reliability in terms of true score and total score variability. They are thus extremely sensitive to factors which affect the deviation of a sample variance from its "actual" or population value. If the range of ability represented in a measurement sample is low, or if there are floor or ceiling

effects which artificially constrain variance, reliability can be seriously underestimated. It was noted previously that high homogeneity of proficiency is characteristic of even moderately experienced operator groups. Everyone is capable of performing the task "satisfactorily," even though some are better than others, and the real individual differences are small compared to error of measurement, causing the ratio of true to total variance to be an incorrect estimate of the actual reliability value. Chiles (1967) specifically notes the problem with reliability when the group is either highly selected or has had extensive training, and suggests that it may be necessary to use variance estimates based on the full range of ability to obtain accurate reliability values.

Reliability summary II -- Previous discussions have identified two "villains" of unreliable measuring -- the instability of the phenomenon being measured and the changes in conditions under which the task is performed. A third problem is that the available techniques for quantifying reliability are not always appropriate for the forms of measurement required in operational or near-operational settings.

Reliability as Generalizability

Cronbach, Gleser, Nanda and Rajaratnam (1972) present an interesting extension of reliability theory into the realm of measure generalization. Their concept of "dependability" of a measurement involves the degree to which the components that make up a measure are known and are quantified to an extent that the measure can be confidently generalized to other situations in which those same components may be present. The process of determining dependability combines elements of reliability or repeatability of measures with aspects usually associated with "validity," such as the nature of content and the consistency of the "construct" being measured. Although a summary of their

complex book-length effort is well beyond the scope of this survey, their theoretical development provides a capability to identify separate measure components and to break out the amount of variation accounted for by these separate sources. Such a breakout would allow those parts of the measure which are of interest to be isolated and removed from those which are considered, for a specific measurement purpose, as error, contamination, or bias, and would simplify considerably the process of establishing measure validity or relevance. So far as the author is aware, the Cronbach et al. (1972) approach has not been used or considered for use in operational measurements situations, but the application in those contexts would appear to be of considerable merit in improving the usability of measures otherwise too confounded with unwanted variance to be satisfactory.

Criterion-Referenced Measures

Distinctions are sometimes drawn between a) "norm-referenced" measures, for which indications of "goodness and badness" of a value are derived by comparison to the measured performance of other individuals, and b) "criterion-referenced" measures, which are defined in terms of how closely a performance corresponds to some established standard or "target" value. There are definite differences between the two "kinds" of measurement with respect to the meaning of reliability and the methods of computing coefficients.

Although it is normally discussed as if it were a unitary concept, criterion-referenced measurement has several different meanings in common use. In a typical educational setting, it involves a requirement for an individual to continue instruction and testing on a body of material until he/she can show a "mastery" of the material by achieving a test grade of (for example) 95% correct. In other contexts, particularly aviation

measurement, criterion referencing implies that a measure (usually a physical one) is compared to a "standard" or desired value, with the criterion-referenced measure taken as a deviation from the standard.

Determining criteria -- These standards or "criteria" can be derived in at least three different ways. Each of these ways involve sources of "error" which are introduced into the measure by different processes of defining a desired performance value, and which differentially affect reliability as a result of that variance in definition.

The most direct form of criterion referencing is the deviation from some "natural point," such as a target. Bomb drop miss distance is an example of such a measure. Such deviations, because of their obvious "relevance," are not always recognized as being of criterion-referenced form. Measures of this type are notoriously unreliable, due in part to their being "outcome" measures, which are heavily influenced in operational settings by the uncontrolled factors previously discussed (weather, etc.). As later discussions will elaborate, they consistently include, as a component of individual proficiency, sources of variation which are uncontrollables entirely unrelated to individual skill.

A second method of deriving criteria involves the a priori definition of some "book" value, the attainment of which is considered "good" performance. In a loop, for example, there may be designated values for airspeed, nose attitude and other parameters at a given point in the maneuver which are considered "correct." Deviations from these values are the measures. The most critical problem with such measures is that the book values are likely to be inappropriate to the way the task is actually performed. In previously discussed work by Westra et al. (1986) using deviation from glideslope, student pilots consistently

flew slightly below the "correct" glideslope in both simulator and field studies. Data from the carrier landing studies of Brictson and his colleagues (e. g., Brictson et al., 1973) show that experienced pilots making successful approaches were consistently above glideslope, particularly during night recoveries. Using the book values introduces a significant error component into deviation measures. Similarly, tolerance bands may be established around the desired value, and a correct performance is one which stays within those bounds. This latter method involves a critical step of determining in some way the size of the tolerance bounds and can introduce a large error component if they are too loose or too tight, in addition to any error resulting from an incorrect choice of the value around which tolerances are defined.

Establishing a priori or non-empirical definitions of criteria or "mastery level" to be attained and the creation of desired bounds around the criteria has been recognized as a potential source of error in measurement for some time (Glaser & Klaus, 1962; 1971). Bahrick, Fitts and Briggs (1957) described the "subtle discrepancies" introduced into skill measures by tolerance bands that make the task too easy or difficult. Hayes and Pereboom (1959) discuss the instability of criterion-referenced measures resulting from chance fluctuations in performance around the criterion. Lane (1986) summarizes the effects of inappropriate criteria on inferences about the patterns of changes in skilled performances which occur with practice.

A third method of established desired values is to determine empirically the behavior of "experts" in performing the task or maneuver and to create a "template" or profile, correspondence to which is defined as good performance. Tolerance bands around the desired value can also be determined empirically, and are thus not subject to the symmetry requirements of bands

established on an "arbitrary" basis. While this method avoids most of the difficulties described above, it requires considerable amounts of data collection and analysis to establish templates. It also introduces another class of potential discrepancies. One could, for example, have instructors fly the maneuvers on which students will be measured, record their performances, and establish criteria on that basis. There is ample literature which suggests, however, that persons learning a skill and those already proficient do not perform the task the same way, i. e., there are distinct qualitative differences in the performances of novices and experts. Trainees are likely to be primarily concerned with control of the process, the steps or procedures which bring about the performance. Experienced pilots are likely to be "goal" oriented, that is, they know what outcomes are required and dynamically adjust the process to existing conditions to achieve the goal (what Krendel and Bloom (1963) described as "adaptability"). To the extent that expert templates are applied to novices who perform the task differently, both measurement and conceptual errors are introduced, and measures become both unreliable and potentially irrelevant.

Special reliability problems -- It was noted previously that establishing the reliability of conventional or "norm-referenced" measures is primarily a concern for the consistency or repeatability of the measure, as it is affected by phenomenon instability, uncontrolled task conditions, and to a lesser extent by the accuracy of recording or observation. Reliability of criterion-referenced measurement, in addition to all these effects, is also a function of the "accuracy of criteria" (Glaser & Klaus, 1962).

It is well established that methods of defining reliability derived from traditional test theory are inappropriate for criterion-referenced scores without considerable conceptual and

computational modification. Popham and Husek (1969) and Livingston (1972) described the bases for this lack of applicability of conventional reliability techniques. Swezey (1978) summarized several different positions on the issue of criterion-referenced reliability.

The principal point of controversy is that traditional ways of computing reliability assume a true-score model, i. e., error variance is defined as deviation around true score, and total variance is defined as deviation around the group mean. For the criterion-referenced approaches used in educational settings, in which performance is based on "percent correct" values, these assumptions are clearly invalid. Scores at achievement of mastery level are always at or near the ceiling value, and variance and midpoints are correspondingly artifactually restricted. Livingston (1972), however, argues that classical reliability equations are appropriate for use in such situations <u>if</u> deviations in deriving variance are taken, not from the mean, but from the "criterion" score itself. This would bring the meaning of variance more into agreement with that required by classical theory.

With few exceptions, the usual (educational) meaning of the term "criterion-referenced" is not encountered in operational measurement situations. It involves the demonstration of "mastery" of some well-defined content for a "unit of training." Few situations in aviation training and probably none in operational flying are well enough understood for the required specification of content. The other meanings of the term, those involving deviations from an externally-derived standard, relate more closely to Livingston's contention that conventional test-retest and internal consistency reliability coefficients are appropriate given that deviations are taken around the criterion value. There are, however, some remaining interpretive difficulties in use of such estimates of

reliability, having to do with assumptions about the nature of error.

Definitions of error -- Discussions above identified three ways in which criterion values could be established. For each of these, the definition of error embedded in the reliability computation needs to be closely examined in order to interpret the coefficient.

a) It was noted previously that taking deviations from the "book" treated as error any contribution due to the book being "off" by a constant factor. An individual slightly above glideslope when the remainder of the group is consistently below will have a small deviation score, when in reality his deviation from the typical successful approach is quite large.

b) When criteria are based on "relevant outcomes" (hitting targets, kills, etc.), performances which may be almost "perfect" can be judged as unsuccessful because of factors in the outcome that are not controlled by the individual being measured, e. g., any inaccuracy in a weapon or its delivery mechanism would be inappropriately treated as error in the performance of the individual.

c) Similarly, given criteria based on "expert templates," a novice could be performing extremely well relative to his stage of practice and receive poor scores if he/she can not yet use the procedures of the expert group. This can create an insensitivity of measures that tends to "wash out" true individual differences.

Minimum standards -- It is also important not to confuse criterion values and tolerance bands as aspects of measurement techniques with their use within a "minimum standards" framework. Through the use of historical or analytical data on

the inherent accuracy of operations, fairly accurate boundaries can be determined within which it is reasonable to expect a proficient performance to fall. This is not the same operation as proficiency measurement. It is well established, for example, that the numbers obtained by measuring the distance from target to the impact point of weapons has virtually no value as an indicator of individual proficiency. This is so because the differences between numbers result almost entirely from sources that are not of interest for performance measurement purposes, i. e., 50 feet and 100 feet are not discriminably different performances, although one is "50 feet" more than the other. One could, however, determine that a) the weapon and delivery system are sufficiently accurate to deliver weapons consistently within kill range (say 75 feet) and b) over a long series of drops by proficient aviators, that 90 percent fall within that range. It could then be established on judgmental grounds that in order to be considered "proficient" in that skill, an individual should be able to deliver, across repeated trials, some acceptable percentage of ordnance within that distance. Such use as minimum standards is not subject to the relatively stringent requirements associated with the representation of numbers as measurement of individual proficiencies.

Reliability summary III -- It is obvious from the foregoing that "criterion-referenced" approaches are by no means simple alternatives to conventional ones. Error control and interpretive difficulties have a pernicious effect on reliability that is easily overlooked in the reliance on correspondence to standards as a way of assigning some "absolute" meaning to measures. It is the author's belief that criterion referencing as a measurement philosophy has not received the critical scrutiny that it requires during measure development and application.

The Impact of (Un)reliability -- An Example

One common use of performance measurements is in evaluation of effectiveness of some alternative method of training. Training Effectiveness Evaluations (TEEs) typically examine the performance of trainees (variously measured) at the end of a course or segment of training, or sometimes on the job after training is complete, with the intent of determining which of two methods, existing or alternative, results in better performance. As Pfeiffer and Browning (1984) point out, there are far more obstacles to obtaining the requisite "performance measures" than are commonly supposed. Previous discussions (and those which follow) identify a host of complications that can be encountered when developing measures in field settings. If these are not addressed in the measure development effort, it is probable that the obtained measures will show the extremely low reliabilities cited above (.10 or less is typical), with 95 to 99 percent of the variance attributable entirely to error. Such unreliability has an insidious, potentially disastrous, and usually unrecognized impact on the inferences from a TEE.

The literature reviewed for the present effort, along with the author's experience, suggest that there is one overwhelmingly common finding characteristic of TEE's: When a new method of training is compared to an existing method, the two are found to be "equally effective." Rarely is Method A significantly better or worse than Method B. This leads to a trap in interpretation of results:

a) If A is a current method that is less expensive than B, the finding is that the new method isn't cost effective and should not be implemented.

b) If B is cheaper (shorter duration, etc.) than A, the finding is that the new method is "just as effective" but uses fewer resources and should be implemented.

Both of the above statements can be made with the same data; the distinction is entirely on the basis of cost, not training effectiveness. The trap is that, with the unreliable measures used, no statement about effectiveness is possible, nor is one warranted without information about measure reliability. In those rare cases where statistically significant differences are found (about 5 percent at the .05 level), the magnitude of differences is seriously understated because of the large error component, and cost-effectiveness judgments using quantitative tradeoffs are likely to be substantively incorrect.

VALIDITY

It was noted previously that satisfactory reliability of a type appropriate to intended measure use is <u>the</u> most basic property of a potential measurement. Reliability, however, only demonstrates that whatever it is that is being quantified is likely to be sufficiently dependable to be useful; it provides no evidence one way or the other about the <u>meaning</u> of the measure(s). As a number of previous sections have discussed, the "legitimate" ascription of meaning or interpretability to numbers requires the joint consideration of the numbers and a variety of external referents in a series of operations that are traditionally subsumed under the single term, "validation."

All too often in aviation performance measurement efforts there is a tendency to fall into the "naming fallacy" (Guion, 1974), to rely on what Guion (1965) called "faith" validity, the process of giving a number a label and then proceeding as if the number somehow represented a measure of what it was named. Far too frequently, one encounters literature on "measuring pilot performance" which describes exclusively the capabilities of hardware and software for data recording and reduction. The operations involved in verifying or "validating" the meaning of

a measure are complex, tedious, time-consuming and costly. They are, however, ethically essential for any measure set to be used for judgments or decisions about individuals, and are critical on economic grounds even for measures which will be used for such purposes as evaluating "system performance." As Wallace (1965) pointed out, it is possible to develop extremely plausible measure sets, with high apparent relevance, which are in reality mostly irrelevant and provide no evidence of any sort germane to the purpose of the evaluation.

"Types" of Validity

The word "validity," like "reliability," carries the burden of many different connotations, too many for any precise use of the term. In 1954, the American Psychological Association (APA) promulgated recommended terminologies for "validities," describing in detail the operations of Content, Concurrent and Predictive validities, and introducing the concept of "Construct" validation. Although the original intent of the 1954 recommendations was to standardize the terms for varying aspects of validation evidence, the net effect was that the labels for "validities" became, as Anastasi (1986) noted, "...reified and endowed with an existence of their own." Over subsequent revisions of test standards, the distinctions were gradually removed, and in the most recent (APA, 1985) the discreteness of terminology has been virtually eliminated.

In most of the literature on applied measurement, the "reification" to which Anastasi objected is apparent. Distinctions are drawn in practice among a variety of "kinds of validities" -- Face, Content, Concurrent, Predictive, Domain and so forth. These are treated in most usages as separate entities. Face validity is equivalent to what has been described previously as apparent relevance, and is a factor of user acceptance, not a proper validity in any formal sense.

Content and Domain validities are the usual validity operations associated with criterion-referenced measurement; they represent the fidelity of a measure to some domain of task operations, the degree to which samplings from the total content may be used to represent the whole domain. Predictive and Concurrent validities are the mechanism for demonstrating what is sometimes called "empirical" validation. Types differ with respect to the formalism of the evidence on validation which they provide, i. e., "empirical" validation approaches produce coefficients of relationship, the others typically do not.

Construct Validity

For the 30 years since the initial descriptions by APA (1954) and the landmark article by Cronbach and Meehl (1955), there have been increasingly stronger tendencies to reduce the formal distinctions between traditional types of validity, and to consider all such validity information as evidence toward a more global conceptualization of the validity issue, subsumed under the term "construct validity." (See Anastasi [1986], Messick [1981], and Tenopyr & Oeltjen [1982], for discussions of these shifts).

There have also been increasing criticisms of the use of the term "validity" to refer to the evidence obtained from the various kinds of "validity operations." Messick (1980) argues that validity properly concerns the meaning of a measure, and the the word "validity" should be reserved exclusively for the demonstration of meaning derived from the operations of construct validation. Content and Domain "validities" should be "content relevance" and "content coverage." Predictive and concurrent "validities" should be considered as predictive and diagnostic "utilities." Trends in the present usage of the validity term are more and more supportive of Messick's (1980) position.

The "philosophy" of construct validation -- In simple terms, construct validation can be viewed as the process by which credibility is attached to the label or name of a performance measure, and to the representativeness of the operations which generate the measures. It was noted previously that the notion of "performance" on a task is itself a construct, in that it is intended to carry evaluative connotations (good, bad, high, low). As such, any set of "performance measures" should have three important characteristics: 1) Scales of measurement must be representative of, and capable of being directly mapped into, the "universe of events" that are ultimately important in successful outcomes for the task; 2) the scores assigned to individuals must be at least monotonic with respect to degrees of goodness/badness of the measured skill(s); and 3) differences among scores should be due primarily to differences in occurrences of "successful" events or processes rather than to other factors. The contents of the performance "universe" are never directly accessible; they can only be inferred from "observables." The degree to which the observables contained in a measure set faithfully represent the universe of performance is assessed through the process of construct validation.

Another avenue to understanding the "philosophy" of construct validity as it applies to performance measurement is through Thorndike's (1949) concept of the "ultimate criterion." The ultimate criterion is an abstract embodying everything that is needed for successful performance. Behaviors which define the ultimate criterion are complex and multidimensional. Thorndike's concept is logically equivalent to the later notions of the "performance" construct which measure sets are intended to represent. In Thorndike's development, the validity of a measure is defined by its "relevance" to, and is measured by its theoretical relationship with, the ultimate criterion. The "construct validity coefficient" (if such were attainable) would

thus be the correlation between the measure and the construct. Hakel (1986) has objected to the continued use of the "ultimate criterion" terminology, not because the concept is inappropriate, but because it, like many validity-related terms, has been reified, and its continued misinterpretation has hindered solutions to problems of the real-world multidimensional measurement space.

The demonstration of construct validity in real-world measurement situations is, unfortunately, never reducible to the simple terms of validity coefficients. It is, as Cronbach and Meehl (1955) described in one of the earliest presentations of the concept, a judgment process based on "...the entire body of available information that leads to or away from the construct." Evidence on the construct should be obtained from the widest possible variety of sources. Messick's (1975) discussion of "philosophy" underlying the demonstration of construct validity suggests "content" as a key form of evidence. The degree to which a measure adequately samples the domain of skills and knowledge required for performance is obviously crucial to the "completeness" with which that measure represents the underlying construct. As Messick notes, however, content "validity" alone is "never sufficient" for measurement. Scores defined only by the measurement operations used to acquire them (the analog of "content" for physical measures) cannot be legitimately used outside the specific context of their definition and thus have no meaning as generalized measures of the phenomenon described by their labels. Even if, for example, deviation from glideslope is considered a sufficient "sampling" of the content of approaches to landing, it is not a satisfactory measure of "approach performance" until it is related to other sources of evidence external to its own definitional operations.

The broader the range of evidence brought to bear on a measure of a construct, the greater the understanding of the

behaviors subsumed and the higher the confidence that the measure is (or is not) a satisfactory assessment of that construct. The process of determining construct validity, as Nunnally (1967) points out, involves the obtaining of as many different kinds of information in as many contexts as possible. This would likely include, particularly for the classes of performance measurements with which the present effort is concerned, a series of studies to explore the behavior of the measure in varying contexts and to relate obtained measures to a variety of external measures which should (conceptually) be selectively related or unrelated to the measures of interest. As Guion (1974) notes, the process is iterative. It may be necessary to build a theory to understand the phenomenon and to guide generation of ideas about the nature of evidence that would support or negate its existence. Specific "hypotheses" about the construct should be generated and evaluated through experiments designed to be as definitive as possible in yielding positive or negative instances about the content of the phenomenon being measured.

Construct validity in the measurement literature -- An awareness of the need for systematic exploration of the properties of performance measures is seen in a limited but increasing literature oriented to training and operational applications. Lane (1975) used a construct model to examine relationships between flight grades over successive stages of undergraduate and readiness training. The previously described sequence of empirical tests for measures presented by Waag and Knoop (1977) reflects the systematic approach required by construct validation. Breidenbach, et al. (1985) present an explicit series of steps for determining the "construct" validity of air combat measures that generalize well to other measurement situations. Their three-step strategy consists of:

a) Tests of the measurement framework -- Determining the relevance of intermediate or part-task metrics to final outcomes

through analysis of the subprocesses involved in task performance and through formal path analysis to determine causal relationships of intermediate to final measures.

b) Tests of skill discrimination -- Demonstrating that measures are sensitive to individual differences in proficiency based on other external data and on presumptive differences in skill among varying experience levels.

c) Tests of user acceptance and training utility -- Is the measure set diagnostic of deficiencies, easily presentable to users, and feasible for practical use in its setting.

In addition to the specific paradigms above, the measurement guidelines document by Vreuls et al. (1985) and the brief review by Vreuls and Obermayer (1985) allude to the need for anchoring measure sets in terms of relationships to external factors. They draw specific distinctions, however, between "measurement," described as "information about performance," and "assessment" which reflects quality of performance (connotations of good or bad). The differences in operations involved in one concept versus another are not entirely clear. "Assessment" would appear to demand the full demonstration of construct validation, whereas "measurement" exists at a somewhat lower level of demonstrated "meaning," requiring some subset of the full validation sequence, most probably sensitivity (differences in scores between experience levels) and content (correspondence to prescribed procedures).

Within the training literature, Goldstein (1978) raises the important question of the domains of settings and people to which training program effectiveness can be generalized. Although his concern is with generalized use of training programs, the validity issues addressed are the same as those involved with measurement. He suggests that the training

context in which "validations" have been carried out determines the extent of their generalizability to other contexts. Four kinds of "validity" are identified, representing progressively more reliance on factors external to the specific context. "Training" validity is based on performance of a given group in a specific training environment. "Performance" validity involves generalization of results with that group to performance in another environment (similar to demonstration of transfer). "Intra-organizational" validity is based on extensions of validation to a new group in the original training environment, while "inter-organizational" validity entails generalizations to other trainees in a different "organization" (possibly with different training requirements). Goldstein's distinctions are related to what might be called the "robustness" of the construct, the extent to which the phenomenon being measured can be said to exist independent of the specific environment in which measures are acquired and the particular operations used to define the measures.

The issues underlying measurement have been explored in far greater depth in the industrial and personnel communities than in those concerned with military training and operational measurement. There is an extensive literature on "criterion" problems in measurement dating back to the World War II studies and prior. Almost without exception, the "criterion" literature can be directly mapped into the measurement issues addressed in the present effort by substituting "performance measures" wherever the word "criterion" appears. Zedeck and Cascio's (1984) Annual Review of Psychology article, for example, reviews all the previous Annual Review summaries of personnel-related literature. They note that every review has identified one or more deficiencies in criterion theory as the major problem in performance measurement. Ash and Kroeker's (1975) review describes the limited progress made in criterion development over the previous two decades, and suggests that, because of the

complexity of the issues and the difficulties of experimentation in applied settings, further breakthroughs are "unlikely" for the next "50 or more years." Hakel (1986) notes that "...we continue to pour most of our resources into...schemes for relating predictor scores to contaminated and deficient but convenient criterion measures" (p. 532).

It was noted previously that the early literature on reliability issues was still directly generalizable to present concerns in operational and near operational measurement. A similar statement can be made with respect to the less current work on criterion and validation issues. It is in fact necessary to go back several decades to find the systematic discussion and guidance on measurement development that is required to satisfy the more "modern" concerns for demonstrating construct validity. There is remarkable agreement between Thorndike's (1947, 1949) use of the concept of "ultimate criteria" as an organizing theme in describing validity operations and the later invoking of the principle of "construct validation" for the same purpose. Delineation of procedures for "criterion development" by Flanagan (1948) and Crawford, et al. (1947) suggest, for instance, that criterion "relevance" should be determined on a "rational basis" (content validity), carefully supplemented by "relationships to partial and intermediate criteria" (Flanagan, 1948, pp. 276 ff.). The correspondence between these procedural requirements and those alluded to in the above discussions of construct validation is obvious. Other discussions of "criterion development" issues (e. g., Nagle, 1953; Guion, 1965; Wallace, 1965; Ronan & Prien, 1971; Guion, 1974) show a similar concern for systematic exploration of the "meaning" of measures. Wherry, Sr.'s (1957) description of "unintelligent diligence" in the "careful" obtaining of measures with unknown properties is the most succinct statement imaginable of why the operations of construct validation are necessary in the evaluation of proposed measurement systems.

A Contention about Validity

The foregoing was intended to establish the author's contention that proposed measure sets, if they are to be described as "measures of performance" or used in ways that suggest possession of those characteristics, must "pass" certain tests. They must be critically examined in a series of systematic operations that provide evidence pro or con about their "believability" as representatives of the "construct" implied by their labels. Buckhout and Cotterman (1963), in describing procedures for obtaining measures for weapon system evaluation, reminded the reader that the recommended process seemed "difficult" because it was difficult. As previous discussions have suggested, the process of attaching meaning to numbers is invariably complex and time consuming, but it is mandatory for certain generalized uses of those numbers.

It is obvious that there will be many occasions in the real-world of measure development when neither time nor resources are realistically available to do a proper job of exploring the "measurement space." In such situations, the objection to some current developments implied by the author's contention is not to doing the best that can be done under the circumstances, even if it departs from "good" measurement practice. Such constraints are frequently (perhaps usually) unavoidable. The objection is rather to the process of attaching labels which suggest, without qualification, measure properties not in fact demonstrated by the operations. Later discussions of measurement in air-combat settings will describe "performance measurement systems" which have been developed and installed entirely on the basis of "apparently relevant" content, without even the most rudimentary attempts to determine measure validity. As Messick (1975, 1981) points out, content "validation" is only the first (insufficient) step in

demonstrating validity. It is important to recognize that the user of "performance measurement systems" will invariably equate "performance measures" with "proficiency measures," and will presume that the numbers associated with an individual carry a meaning of "goodness" and "badness" about that individual's capability. In such a context, the burden of proof is on the developer of the "measurement system" to either ascertain the properties of measures in an appropriate way or to qualify the interpretation of numbers which the system yields.

It may be that the spectrum of meanings attached to the term "performance measure" is too broad for a single label to carry, subsuming as it does both the physical and behavioral aspects of measurement. Distinctions may be necessary in terminology between physical and behavioral measures, between performance and proficiency, between measures which capture system inputs and outputs and those which attach meaning to those quantities. Such a terminology is ambitious, and beyond the scope of the present discussions. The purpose of the above arguments is to suggest (strongly) the need for such distinctions in the interpretation of what a given "measurement system" can realistically be expected to provide to the user.

Operations in Demonstrating Measure Validity

Given the need for dealing with the full aspects of validation implied by the construct validity argument, how does one go about developing a "valid" measure set? The approaches described by Waag and Knoop (1977) and by Breidenbach et al. (1985) are illustrative of the operations required. Any validation approach must contain at a minimum the following steps:

Identifying candidate measures --It is, of course, first necessary to decide on a "candidate" measure set to develop

what previous discussions have referred to as the physical measures or numbers to be obtained on the individual or system. Excellent descriptions of procedures for development of the initial measure set are contained in Vreuls and Goldstein (1976) and in the more recent guidelines document by Vreuls et al. (1985), and will not be repeated here. Steps involve the systematic analysis of the task(s) for which measurement is desired, the isolation of critical behaviors, and the determination of candidate parameters, inputs and outputs that are believed to reflect those behaviors. Inputs and evaluations may be acquired both subjectively and analytically. Procedures are essentially those of demonstrating content validity, and, if done with proper care and comprehensiveness, can yield a sound initial set of candidate measures. The set is usually larger than is feasible or practical to use without further reduction.

Reduction of the measure set -- Eliminating from the candidate set measures which are irrelevant or unnecessary is materially more complex and error-inducing than usually recognized, and a separate section will be devoted to some inferential problems which must considered in those operations. These discussions will point out the importance of remembering that there are two distinct and logically different aspects of measure set reduction. The first involves analytic selection of measures to include or eliminate without reference to any data collected. This makes use of external general information, either subjective or based on previous studies. One could, for example, elect to omit all measures of a certain type because that class of measure has been found to be too unreliable or insensitive in other measurement efforts, or because such measures are too difficult to obtain in an operational setting.

The second type of reduction occurs after data is available on some sample for the candidate measure set. Two distinct kinds of operations can then be performed, those that deal only

with information internal to the sample of observations, and those that relate sample information to other external measures in some statistical way. For those reduction operations that rely only on data within the measure set, measures could be eliminated because they are redundant (correlationally) to measures that might be easier to obtain or because they are found to be unstable or unreliable within the sample. The chance capitalization involved in this operation (as with any use of sample values) is likely to affect reliability but is materially less productive of error than the chance effects which can influence inferences obtained in a second class of operation based on sample data, that which includes or eliminates measures based on their statistical relationship to external variables.

Selecting "valid" measures -- Operations which link candidate measures to external variables (experience, subjective assessments, outcome measures, etc.) which should have variance in common with the measure set are a key aspect of demonstrating validity of the measure set (and its associated constructs). They are thus essential in translating physical measures into performance measures. There are a number of classes of external variables which can be utilized to further define the nature of the candidate set. It is usual to show that experienced operators differ from inexperienced ones on the candidate variables, and those variables on which there are no sample differences are typically eliminated from the set. Candidate measures are often compared to ratings or other subjective assessments to determine shared variance, and their validity evaluated on that basis. Likewise, performance measures during the learning of a skill would normally be expected to increase with practice or time. Special experiments may be performed in which individuals believed to highly proficient on the task are compared to those expected to be less proficient. Task conditions may be systematically varied to increase the

"sensitivity" of the measures in order to determine the measure properties, i. e., operators with higher capability as measured by the candidate set would be expected to maintain performance under more demanding conditions than those of lower measured skill.

One of the critical considerations in determining the meaning and "content" of candidate measures is that measures should relate to some external variables and should not relate to others. Campbell and Fiske (1959) introduce the concept of the "multitrait-multimethod" matrix as a means of distinguishing between convergent and discriminant validation operations. Briefly stated, these two forms of validation evidence hold that measures should correlate more strongly with other measures of the same construct obtained via different methods than they do with measures of different things obtained by the same method, i. e., measures with the same label should converge across varying measurement operations and be discriminable from measures with different labels obtained via the same operations. Thus a measure of "instrument flying proficiency" obtained in a simulator should correlate more highly (presuming equivalent reliabilities) with measures of the same skill in operational flying and with instrument flight grades from training than with "weapons delivery" scores obtained on the same simulator at the same time. (Reliabilities are crucial determinants in these comparisons. A correlation of .20 may be "bigger" than one of .40 if reliabilities of two measures vary widely).

The combination of findings from all the above relationships to externals answers the question of what is being measured by the candidate set, and evidence serves both to reduce the set size to manageable proportions and to shed light on the "validity" of the measure set. (It should be noted that relationship to externals as a means of reducing or validating

the measure set logically demands that those external variables are known to be at least as reliable as the candidate measures themselves). There are no validity coefficients associated with the evidence; deciding whether a measure set is "valid enough" is a judgment call dependent on both its intended use and the "validity" of alternatives available for that use. Measures of relatively "low" validity may be of great value if no other comparable means exist; those of "high" validity may not be useful if they do not improve the quality of measurement enough to justify any additional cost. This latter consideration is that of utility of the measure set, and will be discussed further later in this section.

The operations which examine the meaning of measures by systematically relating candidates to available or specially developed external variables which should "covary" with the measures are the definitional steps in establishing construct validity. They are also the principal source of inferential error about measure validity, in that full advantage is taken in measure selection of the chance effects that will be present in any statistical comparisons. Paragraphs below address these effects and some ways of reducing their impact during measure set selection and reduction.

Determining the size of the measure set -- Physical measures collected during operation of an aviation system and the various transforms of those measures can produce number sets with a large number of variables representing many different families of operator input and system output metrics. There are, in the typical recording of input/output data, far more variables than are required and more than are statistically manageable. The ability to record and assign numbers to virtually all system parameters, and to add to the set without effort unlimited combinations and transforms of basic measures, can create a data explosion.

In a helicopter simulation measurement study by Vreuls et al. (1973), the candidate measure set contained 749 parameters across 12 maneuvers, times the number of repetitions for each subject, further subdivided by experience (2 instructors vs. 2 trainees). They performed over 3000 t-tests. Somewhat more were significant than expected from chance (268 at .05 level vs. 156 expected). There was clearly something in the total data set which differentiated trainee performance from instructor performance, but which of the 268 differences were real and which were part of the chance effect was indeterminate without replication. The authors also performed discriminant analyses on the measure set, reporting a different subset selected than that produced by univariate analyses, likely due to the reduction of redundant variance weighting brought about by multivariate techniques. They further indicated that the sample size of 4 was too small to properly reduce a candidate set of such magnitude. In later studies, Vreuls and Goldstein (1976) expanded further on the problem of measure set reduction, suggesting a series of systematic steps. Vreuls et al. (1985, Section 7 and Appendix C) discuss some multivariate methods specifically modified to reduce the impact of chance capitalization in measure reduction.

Regardless of the sophistication of the technique employed, however, large candidate sets which must be reduced by information from within the set itself pose serious risks of error. There are only a certain number of independent inferences that can be made from data on a given number of subjects. Unless sample sizes are very large, much bigger than is usual in training or in operational measurement situations, extreme care must be exercised to avoid results which appear "significant" in the statistical sense but can be explained even more readily by chance effects.

Obtaining additional replications on the same subjects, while it can assist with reducing capitalization on chance, does

not increase the number of allowable inferences about the discriminating power of measures in the same way as would adding more subjects. Replicated data sets on the same subjects are inherently correlated, and share sources of variation (daily fluctuations, interactions of people with tasks and environments) that would emerge as statistical error on cross-validation. Repeated measurements, although they increase power to detect certain kinds of effects, also introduce the problems associated with repeated measures designs -- homogeneity of correlations between measurements, statistical management of between and within subject variance, and so forth.

Repeated measures problems are to some extent controlled for purposes of significance testing by the use of multivariate analyses, but when the objective of the effort is statistical prediction or weighting of variables into a single measure of performance, there are two distinct disadvantages to multivariate methods.

First, the additional data points from replication are not useful in determining statistical weights for prediction of some criterion since there are as yet no satisfactory techniques for combining between and within subjects effects for such purposes. Wooldridge, Breaux and Weinman (1976) present a method which successfully adjusts individual scores for between/within variance components, but produces indeterminancies in the degrees of freedom associated with the residual scores, create interpretive uncertainties. Second, there are further interpretive difficulties associated with composite scores in multivariate space. Discriminant factors, which are the derived variables on which two or more groups are compared, are not always easy to explain to users as measures of an individual's proficiency. In addition, while the discriminant-factor measures may be appropriate for some applications (such as APM's), it may be hard to use in others, such as assisting instructors or students in problem diagnosis.

There are obviously a multitude of potential measures that can be extracted from even the simplest system. The reduction in some way of that multitude to a manageable set is the most basic methodological problem in any performance measurement effort. Chance has a major influence in measure set reduction because of two interacting effects.

a) Although only a relatively small subset of measures may ultimately be selected for inclusion, the opportunity for capitalization on chance occurs when a variable is initially examined or screened for inclusion; whether or not it is in the final set is relatively unimportant for the stability of the set. As Lane (1971) demonstrated, the major determiner of how stable a subset is likely to be in subsequent samples is the ratio of total set size (M) to the sample size (N). The M/N ratio had a strong linear relationship to the amount of subset variance attributable to chance effects, while the actual size of the ultimately selected subset and its relationship to N had no systematic relationship. In simpler terms, for any given sample size, it is the size of the initial, not the final measure set that determines how much error gets into the measures. In situations where subset selection is based on statistical testing within a single sample, Lane's (1971) analysis recommended that there be at least ten subjects for every variable in the initial set, and preferably twenty or more subjects per variable, in order to get stability across samples.

This attention to initial vs. final set size is crucial for avoiding potential statistical disasters. In an examination of possible automated measures for a light plane simulator, Hill and Goebel (1971) used 3 groups (10 subjects per group), with each group representing a different level of flight experience. Their initial measure set consisted of 266 variables including a number which were derivations or transforms of one another.

They reduced the 266 variables to 57 by one-way ANOVAs (all variables significant at .10 level were included). They then performed stepwise multiple discriminant analysis to find the weighted subset of the remaining 57 variables which best differentiated among the 3 groups.

Of the remaining 57, the discriminant function "selected" 27 variables, producing perfect group separation (analogous to a multiple correlation of 1.00). The number of variables selected (27) is exactly equal to the total number of degrees of freedom available (3 groups x [10 -1 = 9] per group). It is well understood that two pairs of observations will always produce a correlation coefficient of 1.00; they "must" do so since they have no "degrees of freedom" left. For precisely the same reasons, any composite derived from weights based on sample data and containing the same number of variables as there are degrees of freedom available will produce perfect prediction. For Hill and Goebel's study, any subset of 27 variables from the initial 266 would serve as well as any other subset. Further, since the expected decrease in relationship to the experience variable (the "shrinkage" of predictability in another sample) is based on the initial set size, any subset so derived would also have an expected cross-validity of essentially zero. It is thus possible, if sample size and measure set size do not "match," to become enmeshed in analyses which logically cannot provide any statistically dependable information at all about the measure set. Measure sets so developed are at risk of being, as Cureton (1950) expressed it, "baloney."

b) From the above discussions, it can be seen that almost any reasonably derived set of candidate measures for a complex skill is likely to be "too large" for the sample sizes typically available for performance measurement studies. For an initial candidate set size of 50 measures (fairly small by modern standards), a proper study would require at least 500 subjects

to reduce statistically by subset selection based on data from within the sample. Thus one of the ways of guarding against chance, the maintaining of sufficient sample size relative to the number of variables, may not always be possible.

As noted previously, another way to avoid chance effects is to somehow reduce the size of the initial set by using criteria for eliminating variables developed external to the present data. This procedure does not "use up" degrees of freedom from estimation of parameters. Alternatively, using only data that has to do with relationships among variables in the measure set, without considering their association with any criterion or outcome measures, will accomplish some savings in degrees of freedom, although there are still serious inferential risks. External bases for elimination of variables can vary widely, including expert judgment, perceived redundancy of content, or suspected unreliability, so long as these "rules" for elimination are not in any way based on values that must be computed from the sample.

A Validity Summary

Whatever the basis for its determination, a performance measure set should yield quantified information which is appropriate to the intended purpose of the measure set and has been demonstrated to be "valid" for that purpose. The reasoning presented above argues that the proper basis for demonstrating "validity" is through the operations associated with the accrual of measure credibility in the process of "construct validation." Measurement systems for which such demonstration is lacking or can not be accomplished due to time or resource constraints may still be of value, but should not be represented as providing measures of "performance" or "proficiency" for an individual operator or trainee.

SENSITIVITY OF MEASURES

The sensitivity of a measure reflects the extent to which the measure behaves "appropriately" in response to changes in the conditions under which the task is performed or to differences in individual capability to do the task. An "insensitive" measure tends to be of limited variability, and that variability is due primarily to factors other than those of interest, mostly measurement error. Changes in such scores do not relate in a lawful way to shifts in the nature or intensity of task variables; the strength of the stimulus required to bring about a threshold change in the measure is not proportional to the response.

Sensititivity is in one sense not a separate criterion for measures, but rather a characteristic which determines whether or not a measure can be shown to be reliable and valid. A measure which is weakly sensitive to influences that could reasonably be expected to affect performance is likely to be neither reliable (variability is limited and mostly error) nor valid (because of unreliability and because real differences are not reflected in the measure).

The importance of sensitivity in demonstrating other properties of measures has been previously discussed in several different contexts. Measures on a task which is not difficult enough for the group performing will be "insensitive," as in the example of highly experienced pilots flying instruments on a light plane simulator. Whenever members of a group are all capable of performing within the "level of resolution" of the measurement system, the measures will be composed principally of random variation, hence unreliable, and cannot be valid indicators of the skill. In Kelley and Prosin's (1969) example of evaluating tracking displays with experienced operators, there were no reliable between-subjects differences. Similarly,

for a task which is too difficult, the most and least skilled individuals are equally unable to control the process, variances are mostly random, and reliability and validity outcomes are the same as if it were too easy.

Sensitivity of measures should be considered in evaluating other properties of measure sets. An otherwise useful measure with highly relevant apparent content could be eliminated from a set because of a hidden sensitivity problem that might be readily corrected. Poulton (1965) suggests a variety of methods for increasing measure sensitivity, among them the adjustment of difficulty to improve threshold response, systematically selecting task conditions to induce greater variability, the use of that variability as a parameter in lieu of averaged scores, and the breaking down of a complex task into its separate components. Poulton notes that the greatest sensitivity is encountered when task difficulty is set so that the average performance falls in the midrange of possible scores. This has implications for the sensitivity (and reliability) of criterion-referenced measures based on mastery levels, for which terminal performance is by definition very near the upper limit attainable on a task.

COMPLETENESS AND COMPREHENSIVENESS OF MEASURES

Measure Dimensionality

The successful performance of any non-trivial task involves the coordination of many different task-related skills. The factor analyses previously cited (Fleishman & Ornstein, 1960; Wherry, Jr. & Waters, 1960; Zavala et al., 1965) showed that there were many "independent" dimensions of flying skill, even over a limited domain of training performance. Lane (1975), for example, showed that throughout basic, advanced and operational flying, evaluators made consistent and reliable distinctions

between such aspects of proficiency as basic airwork, instrument flying and ability to use weapons.

The "multidimensional" nature of performance is reflected also in the "criterion" literature of World War II (Flanagan, 1948) and after (Nagle, 1953), and in concerns for single vs. composite criteria noted by Dunnette (1963) and Guion (1961) among others. The need to tap a spectrum of important factors is central to the synthetic task approaches noted previously (Alluisi, 1967).

The importance of capturing all the relevant aspects of performance is most clearly seen in the measurement efforts organized around "taxonomic" concerns. Wheaton and Farina (1971) and Finley et al. (1970), among others, addressed the need for greater systematization of the measurement process, for anchoring measurements in an understanding of the components of behavior underlying the measurement system. There have been numerous efforts to develop a "catalog" of human skills as they are manifested in the operation of systems. The comprehensive volume by Fleishman and Quaintance (1984) reviews virtually all the suggestions of previous authors for classification schemes and taxonomic categories for human behavior. Perhaps because of the discouraging complexity of such work, there has been only limited emphasis on taxonomy and classification in recent years compared to those of the 1960's and 1970's (Fleishman [1982] is an exception). Such concerns remain critical in systematizing complex subject matter areas. Lane and Waldrop (1985), for example, describe the role of skill classification in improving the utilization of data about where and how to use computer-based instruction; more germane to the present effort, Lane (1986) reinforces the importance of "taxonomic" categorization of skills and abilities in understanding the process of complex skill acquisition. It is clear that some formal or informal "theoretical" structuring of the expected

"measurement space" is essential for initial development of the measure set, for understanding its "meaning," and for refining comprehensiveness or coverage, particularly for measures based on "objective" or criterion-related approaches.

The Content of Measures

The "completeness of dimensionality" of performance measures is to an extent equivalent to the problem of content/domain validity, or, as Messick (1980) prefers, "content relevance" and "content coverage." How well does a measure or measure set represent the complete domain of behaviors or skills important to task performance? This problem is somewhat differently defined for "objective" and "subjective" measures and for measures derived from different aspects of task behaviors.

Outcome measures -- For objective measurement based on task outcomes or "products" (kills, percent detection, impact deviation, etc.), the "relevance" of the measure is self-evident. At least in theory, outcome measures are clearly the most comprehensive and complete kind of quantification; all the important processes of the task operate in their natural context to produce varying degrees of success on the ultimate task objective. As previously noted, however, such measures are of inherently limited reliability, are poor indicators of individual performance and are nearly useless for diagnosis of why outcomes may have been unsuccessful.

The problems with outcome measures were well defined in the World War II literature, and were the subject of a powerful argument by Wallace (1965) about the "tyranny of the relevant criterion." Wallace noted the perceived need to use measures with poor metric properties but readily apparent relevance because such measures were important to the "customer." Such tendencies, he maintained, were a major obstacle to achieving

validity and reliability and to understanding the process of criterion development well enough to improve it. The apparent comprehensiveness of the measures was misleading, since outcomes were influenced not only by effects of interest, but by a host of other factors not germane to the performance involved.

Criterion-based objective measures -- Some objective measures make use of information from within the task performance process rather than the outcome. One major class contains those measure sets which depend on comparison of operator activity to pre-established criteria, what are traditionally called criterion-referenced measures. Completeness of such measures is primarily a function of whether or not the "criteria" elected contain all the important performance components. Previous discussions have identified several ways in which the measure set of criterion-based variables can be chosen. A key concern of those using such measures to represent performance has been with content issues (e. g. Waag and Knoop, 1977), and a careful attention to inclusion of what are presumed to be important factors can likely guard against major incompleteness problems. There is, however, a considerable risk of "criterion deficiency" resulting from reliance on the "book" if the book is not correct. Even if the behaviors included are the "right" ones to achieve the desired comprehensiveness, the measures of those behaviors may not have the other required properties, and the resulting set may be complete but not satisfactorily reliable or valid.

Subjectively-derived measures -- Observers ratings or estimates of performance are typically global assessments, either of overall proficiency or of how well an individual performs a specific operation or maneuver. Previous discussions have noted that such measures have both major weaknesses and significant strengths with respect to other measure properties. Their potential for completeness, however, in comparison to

other classes of measures, is relatively high, because of the ability of informed "experts" to combine judgmentally a set of dimensions that are inherently different in meaning and on different scales. Hakel (1986) noted that research actively continues on better ways to develop subjective measures. He suggests that ratings should be treated seriously as a source of performance information, that it is important to understand "...the determinants of what others say about a person..." and his/her performance.

Raters who are suitably experienced in the tasks performed, while they will differ on the importance (the weights) which they attach to various aspects of performance, are probably keying on the "correct" aspects. The rating "templates" applied by different raters are thus "comprehensive" in the sense that they capture most of the underlying factors important for useful individual differences. The different weights in the templates are not a "completeness" problem; they affect reliability and validity and diagnosticity, and require pooling of measures over raters and over occasions to "average out" rater bias. It was noted above, however, that many of the problems of subjective measures are "fixed" by pooling, and the resultant averages are relatively complete measures.

This "protection" against lack of comprehensiveness does not apply, however, to measures for which the components of performance to be observed are given to the raters or are otherwise selected "subjectively." Grunzke (1978), in an evaluation of a "measurement system" for air-to-air intercept performance, had experienced operators identify those dependent variables available from a training simulator which were judged as important for assessing task performance. Evaluators then assigned weights to each of the 28 criterion variables so selected, and a total score for each individual was computed from these weights. Grunzke noted that the resulting measures

were adequate as an "informational feedback tool," but did not discriminate among skill levels, and were only "minimally effective" as a performance measurement tool. It is likely that constraining the initial measure set to those physical variables "available" on the simulator, and the requirement to specify the weights to be used, prevented the "global" judgment of goodness or badness from operating while retaining both the variability in weighting within the template and the initial incompleteness of the measure set. The averaging of incomplete measures still produces incomplete measures.

Combining Measures for Comprehensiveness

Given the essential multidimensionality of task performance, there are likely to be, in any situation for which measurement is required, a variety of different kinds of measures which might reflect job proficiency. One way to obtain a single "number" as an indicator of performance would thus be to combine in some way all the various indicators that may be available or derivable. As noted previously, Guion (1961) and Dunnette (1963) took sharp exception to the combining of separate "criteria." They held that the multidimensional nature of performance was a realistic and unavoidable condition which must be dealt with through other methods than combination into a measure of unknown properties.

They noted further that job success involves what Ghiselli (1956) called the "individual dimensionality" of criteria, that equally "effective" individuals would differ in the components of the task that were emphasized to achieve "good" performance. A salesman who enjoys personal contact but not paperwork might increase his business base by investment of time in development of new clients, while one with opposite interests might achieve a similar outcome by increased attention to improving responsiveness and delivery in satisfying present clients.

Combining "marketing" and "follow-up" measures for the two individuals would create an overall measure which misrepresented performance for both and which further lost the utility of the individual performance components. Similarly, a pilot who recognizes himself as marginal at precision airwork might compensate in a weapons delivery mission by greater attention to planning the various stages of approach to release point so that he does not exceed his own personal envelope for making last-second adjustments to the flight path. In the paradigm advocated by Dunnette, separate measures of such component activity are more revealing and more meaningful than any combination of component scores.

Despite the theoretical correctness of the position of maintaining separate measures, there are still pressures for the production by a measurement system for a single "overall" representation of proficiency. As Thomas (1984) noted, "unitary" measures are important for a number of purposes, particularly in a training setting. They serve as general indicators for a) decisions about individuals (note earlier cautions about the strong requirements for validation in such uses), b) scaling the difficulty of training or practice to be given, c) evaluating the effectiveness of alternative training procedures, and d) general feedback to trainees.

Given that one or more of these purposes are "valid" requirements for performance measures, how should separate components be combined? The "global" assessment of performance described above, obtained from raters' observations of task behaviors without constraints on the structure of how they might assess and weight performance components, is one way to approach the combining of diverse criteria for goodness and badness. Allowing raters to decide how to combine components, however, bypasses the process of obtaining separate component measures, and considerably reduces the diagnosticity of the assessment.

Given a need to combine separate measures on other bases, there are a host of available methods of weighting them to produce a total "score."

The "bid system" -- Nagle (1953) presents several different operations for combining numbers, including weighting by reliability, by relationships to external variables, by factorial content, etc. He recommends a procedure familiarly known as the Toops (1944) "bid system," in which raters are given a fixed sum of "points" (usually 100) to distribute across the set of separate measures. Weights are derived from the pooled numbers of as many informed raters as can be obtained.

The "bid system" is also the choice of the present author, having been applied "successfully" in several different problem settings. One (unpublished) application involved the development of a pilot selection system for an Arabic-speaking population with no opportunity to acquire validity or criterion information prior to system use. A set of candidate tests was compiled. Using a "synthetic validity" approach, knowledgeable workers in the prediction field were given three hypothetical ability factors known to be important in training success for U.S. pilots and asked to estimate the relative importance of each factor in success of the new population by distributing 100 points of weights over the three factors. They were then given, in a separate evaluation, the candidate tests and asked to estimate the factor structure of each test by distributing 100 points of "factor loadings" over the three factors. The resultant estimates were then combined into composites for screening the new population. Because of the relatively high selection ratio that was necessary to fill available quotas, it later become possible to estimate, from subsequent training data, the effectiveness of the composites derived using the "bid system." It was found that composites related as well to training success as those that could be obtained by regression weights after the criterion data were available.

Policy-capturing techniques -- Another class of technique for combining separate criteria involves the analysis of global performance assessments to infer the weighting strategies used by raters or judges of performance. These "policy capturing" methods work backward from a collection of single judgments and from the separate criteria that might have been employed by raters to isolate the "rating policies" involved in assessment decisions. The effect of such approaches is to capitalize on the "completeness" characteristic of global judgments but to improve on the template weights by determining both the factors deemed important by judges and "how important" each was considered in the actual production of a rating. Estimates so obtained are likely more representative of what judges actually do than those obtained by directly asking them what they consider important, since the latter outcomes require a "self-conscious" attention that may not correspond to their actual practice. In addition, by virtue of the statistical operations involved, such techniques give higher weight to those criteria on which the degree of "consensus" among raters is greater, based on their rating behavior.

One of the earliest of the quantitative policy-capturing methods was the Judgment Analysis (JAN) technique described by Bottenberg and Christal (1961), which involved in essence the clustering of individual raters' prediction equations. This was expanded and further developed theoretically by Naylor and Wherry, Sr. (1965). Interest in such approaches, using other logics for decomposing raters' decisions, has continued. Hobson and Gibson (1983) review the recent literature on policy capturing as a means of improving performance appraisals in a business context. Much of this work has employed techniques from current decision theory. Zedeck and Cascio (1984) review some of these efforts in the context of performance theory as do Pitz and Sachs (1984), who expand on the underlying decision-theoretic bases involved.

Metric Properties of Combined Measures

Combining different (theoretically independent) measures of job performance has distinct effects on the properties of the resultant overall measure. In most circumstances, combined measures are likely to have higher potential for "validity" than any of their components, since they should represent more of the effects which influence the variation of the theoretical "true" performance, and have more variability in common with the "ultimate criterion."

There are some byproducts of the combination operation that are sometimes overlooked in considerations of theoretical validity. One, previously noted, is that the "meaning" of the combined measure is only determinable at the most general level, i. e., as a measure of overall performance. A second effect is that reliability as determined by some of the classical methods is likely to decrease. There are a number of assumptions associated with internal-consistency reliabilities, all of which are related to the presumption that the individual "items" which make up the measure are all estimating the same ability (see Wherry, Sr., 1984). This is clearly not the case with composite measures, and the use of such techniques as split-half and the various Kuder-Richardson formulas are not appropriate with composites. If they are used, they are likely to suggest somewhat lower reliabilities for the measures than are warranted, because the various "halves" being related to obtain the coefficients are not in fact equivalent. Such a reduction is not likely to be encountered in "test-retest" coefficients, which should be in theory increased for a composite over its individual components. Whether the reduction in reliability obtained from internal-consistency coefficients is a real loss of measure reliability will depend on the use to be made of the measure. "Intra-mission" reliability is likely to show a "real"

loss from composite measure use, while the generalizability of measures across missions will typically be enhanced.

SEPARABILITY OF OPERATOR CONTRIBUTIONS FROM THE MEASUREMENT CONTEXT

If "comprehensiveness" is the inclusion of all the relevant components of performance, then the concern for "separability" is for the omission or exclusion of irrelevant components. There are three sources of such irrelevance, those associated with the operator, those associated with the system on which the task is being performed, and those associated with the environment in which measures are being obtained. Operator-associated irrelevance contains such factors as instability of individual performance, momentary shifts in strategy, and so forth (unless such transients are the subject of interest). System-associated irrelevance includes variations among specific sets of equipment or platforms that change performance systematically but in ways that are not relevant to operator performance. Environment-associated irrelevance involves such factors as weather, target variables and other uncontrolled aspects of the task situation.

Throughout most of the above descriptions of desirable measure characteristics, measures have for simplification been treated as if they were, in varying degrees of completeness and fidelity, faithful representatives of the operator's influence on some ultimate performance variable. This is obviously not always the case. Earlier discussions about the nature of measurement and the extraction of meaning from measures dealt at length with the extensive literature on the confounding of sources of variability and its influence on the composition of obtained measures. Technique- and condition-specific variance will be encountered in virtually all measurement systems, even in those carefully developed and controlled, e. g., as in

high-fidelity simulators. An important part of measure "validation" is to look for such effects as part of the evaluation of the measurement system. Some creativity may be required in this process. Carter and Dudek (1947), for example, correlated scores obtained in the same seat for different navigators to estimate system contribution to the measures. Similarly, one could correlate a trainee's "performance" in one block of trials on a simulator with those of other trainees on subsequent trials to estimate the system-specific component.

It should be recalled that irrelevant contributions to measures can make reliability both higher and lower than the "true" value, depending on the time frame in which measures are taken and the degree to which different equipment is involved. Estimates of reliability from within a given performance sequence will be spuriously high to the extent that system and task conditions hold for that sequence but not for others. Estimates across sequences will be lower than the true stability of performance warrants unless the confounding variation is experimentally or statistically removed.

DIAGNOSTICITY AND SPECIFICITY

Every measurement system is (or at least should be) developed for some purpose. Throughout previous discussions of the characteristics of measurement systems runs a continuing emphasis that the properties most important in evaluating a system will vary as a function of that purpose. By far the bulk of efforts reviewed in prior discussions have had as a primary objective the measurement of an <u>individual's</u> capability to perform a task, and most of these, with only a few exceptions, were intended for use in evaluation in a training-related context.

While global measures (either ratings or constructed composites) have some practical uses in training (see Thomas, 1984), they are not helpful in determining the reasons for a particular performance being deficient or proficient. If measures are to be used for guidance of a novice or for detection and remediation of a specific difficulty, the variables contained in a measure set must be diagnostic. Virtually any measure of individual performance can be viewed as composed of two major components, a) how well an individual understands (has an appropriate model of) what he needs to do, and b) his skill in execution of that understanding. Training is concerned with both components, but the bringing about of improvements in each will normally involve very different training regimens (communication of knowledge vs. practicing the skill). An important element in measures for training is thus a sufficient refinement to "shred out" the part of performance attributable to each for diagnostic purposes.

Requirements for Diagnosticity

To be effective in such diagnostic use, variables must satisfy three general requirements: a) They must provide a level of detail which allows differentiation among skill and knowledge components, b) they must be sufficiently distinct in the content they measure, and c) the measures must be capable of being mapped with a reasonable degree of correspondence into those specific components. Restated, the constructs estimated by the measures should represent conceptually different aspects of skill, the obtained measures of those concepts should not correlate too highly, and each skill construct should be directly linked to some distinct score or score combination.

Diagnosticity and the Validitation Context

Diagnosticity/specificity is primarily a validity (rather than reliability) issue, in the sense that a measure set must be

evaluated in the context of its intended purpose. A set that is poorly diagnostic may, as noted above, be acceptably valid for some purposes (e. g., minimum standards), and not acceptably valid for other purposes (e. g., isolating the reasons for poor performance). There is an additional caveat involved in the determining of validity for diagnostic purposes. Proper use of a measure set requires that it be employed in the context in which it was validated. Diagnosticity is particularly affected by this distinction. It is in essence a property of individual rather than group data, and should be anchored in individual differences. Measures validated on the basis of group or unit differences should be used on groups or units, but not to assess individual deficiencies in performance.

UTILITY AND VALUE OF MEASUREMENT

A measurement system may be reliable and valid and possess all the other properties required of performance measures and still be of limited utility. The utility of a measurement system for a stated purpose must be judged against a set of criteria for value that are on a set of judgment axes independent from those described in preceding sections. To be "useful," a method must produce results that represent "true" performances more closely than any other available and affordable way of achieving that objective. If better information about performance is obtained, is it enough better to justify additional cost and time in obtaining it? It is obvious that, other things equal, operator performance measured repeatedly on a high-fidelity simulator is a better estimate of operational flying skill than is a single flight check to evaluate the attainment of minimum proficiency standards. One may, however, be feasible within given constraints of time and resources, the other may not. Information on validity, reliability and other properties does not resolve such questions; there remains the question of "utility for what purpose and against what alternative."

The principal determinants in evaluating utility are: a) Effectiveness of the measurement system against other available alternatives, b) the practicality and feasibility of implementing the system, and c) embedded in both the above criteria, the cost-benefit obtained from using the system.

Effectiveness against Alternatives

Quality of decisions -- How much better are the decisions reached using the measurement system than those made without it? Depending on the purpose of measurement, existing methods may be adequate for the bulk of measure uses. There is a definite tradeoff (not always examined) associated with implementation of a measurement system. In addition to the direct consideration of affordability and practicality of implementation, there are implicit concerns for whether the measurement system is capable of improving the functioning of the training or operational unit, and to what extent. There may be alternative ways of obtaining estimates of performance, either existing or potential. Subjective appraisal or minimum standards systems, for example, may already be in place. It is important to evaluate a new measurement approach in terms of such questions as to whether its use results in better-trained students in a given training period or reduces the time to reach a stated proficiency, compared to those existing alternatives and separate from cost concerns. Two measurement systems which lead to the same or nearly the same decisions are equivalent, regardless of any differences in sophistication or elegance.

Cost/benefit -- Given a differentiation between alternative systems with respect to improved outcomes, some judgment is required as to whether the improvement is "worthwhile" relative to costs. Depending on the nature of the application, the investment required to develop and validate a "proper"

measurement structure may or may not be recoverable over its projected life-span. Such cost recovery is only to an extent dependent on the quality of the measurement system; it is for the most part a function of the relative cost of correct and incorrect decisions and of the total frequency of system use over which cost must be amortized.

Practicality of Implementation

Regardless of how well the measurement system functions as an assessment tool, the degree to which it can change decisions and produce cost-effective results is heavily influenced by factors that affect its likelihood of being used in the context in which it must operate. The two key factors in that context are the feasibility of obtaining needed data (primarily engineering constraints) and the acceptability of and support by the user for the measurement system.

Feasibility -- Complex measurement systems tend to require extensive and relatively well-controlled data collection. Such capability almost certainly must be based on data recording and reduction equipment distinct from that already in place in the application setting. Additional instrumentation packages may be necessary. In addition, in order to realize the benefit from measures, particularly from the Automated Performance Measurement (APM) packages intended for training use, they must be available in real time or very nearly real time; this requires sophisticated computational capability.

The more equipment required and the greater its complexity, the more likely it is to malfunction and the heavier is the burden imposed on personnel to calibrate and maintain it. Systems at operational sites are likely to spend large amounts of time in a down status. Semple et al. (1981) describe some of the difficulties encountered with implemented APM systems, as do

a number of writers in an instructional features workshop proceedings edited by Ricard, Crosby and Lambert (1982). Several authors noted that the sheer weight of data is a threat to successful functioning of such systems. Semple and Cross (1982) refer to the "voluminous performance data" associated with APM's, and Charles (1982) suggests that "...The quantity of data far exceeds the capacity of any IP [instructor pilot] to utilize effectively during training and still be able to monitor and evaluate...pilot performance." (p. 17, brackets added).

A measurement system for which instrumentation and computation requirements exceed the capability of a user to support will not be feasible for application in that environment, regardless of its power to assess. An important consideration in evaluating measure systems, over and above their desired metric properties, is the extent to which the intended using organization is both able and willing to invest the additional effort and resources required for implementation and support of the physical components of the system.

Acceptability to user -- In addition to the feasibility issues associated with a measurement system's imposition of increased requirements on the user, the additional load can serve as a point of irritation and reduce benefit from the system by decreasing user acceptance. The author's observations of APM's (particularly simulator-based ones) suggest that they are used only when the benefits from improved or simplified training procedures materially exceed the investment of time required to make use of the measures, i. e., the measures either reduce workload or provide more effective use of an instructor's time. If such a balance of return for effort is not present for a measurement system, it will be studiously ignored or (as happened in one case) used as a source of spare parts.

The user should be able to understand the output of the system and should be able to integrate measure use into an

ongoing training or operational flow without major revisions to procedures. These needs suggest that some form of summary or "top level" description is far more likely to be used (despite its possible metric limitations) than the large quantities of "undigested" parameter data often provided. As Semple and Cross (1982) note, such "data recording" systems are of low utility -- "...such capabilities have found little acceptance for performance evaluation and learning problem diagnosis in day-to-day training. In other words, such systems are not used by instructors. The volume of data produced...often is overwhelming and is difficult to integrate and interpret" (p. 30, emphasis added).

Cost-benefit and utility tradeoffs -- There is obviously little return on investment available if a system is not or cannot be used due to impractical engineering or to a measure set design which fails to achieve user interest and support. It is also possible for the measurement requirements in a situation to be so complex that a system which has high "scores" on all the required metric properties and is acceptable to the user is simply unaffordable given present technology. The opportunity for system use may be so limited, or data collection so expensive, that no return of investment can approach justification of development and operating costs. In such a case, some tradeoffs of otherwise desirable requirements against affordability and complexity will be necessary.

Each of the "criteria" for measurement described above (with the exception of some reasonable demonstration of validity) can be "relaxed" to some extent. Diagnosticity, for example, can within limits be traded for reliability by pooling individual component measures, any one of which is insufficiently reliable to be useful. The reduced cost and complexity of obtaining these less reliable component measures may make feasible or affordable a system which otherwise is unacceptable on either

basis. Such systematic trading does not reduce in any sense the importance of examining measure systems for each of the characteristics they should possess. It is still critical to know, for example, the diagnosticities of measures, even if it is necessary to accept a reduced diagnostic capability. Similar trades can be made among other desirable characteristics if the interplays among measure properties are understood and considered in planning the measurement system design.

THE COMPOSITION OF PERFORMANCE MEASURES

PROCESS VS. PRODUCT MEASURES

To a considerable extent, the confusion and inconsistency seen in the use of the terms "performance" and "measurement" is attributable to a lack of proper distinctions between the product of a task (an outcome) and the process which acts to produce that outcome. A task or behavior is usually perceived as a continuous event, with a discrete outcome. Singer and Gerson (1978) illustrate this distinction with the analogy of hitting a baseball. One hits the ball (the product), but the swing before and after (the process) is the determinant of where the ball goes, and thus the swing is the most direct object of study, not the impact on the ball. This analogy can be extended to clarify another difficulty with product variables; the path taken by the ball and the distance it travels are determined by the wind and other ballpark conditions, and have very little to do with the phenomenon of interest, how well the bat is swung. The impact point on the ball may be worth studying, but only because it can be used to infer characteristics of the swing if the swing is not otherwise observable.

THE NATURE OF PROCESS VARIABLES

Of more relevance to the present study is the example of a weapons delivery task. One can observe three related aspects of the delivery process: a) The impact point, b) the release point, and c) the path by which the release point is reached. The latter two are what are considered in these discussions as process variables. Release point could be treated as an outcome variable in the sense that it is a discrete "outcome" of the previous flight path; it is considered here as a process

variable since it is only the last in a series of "waypoints" or key events in the flight path that influences impact point.

The three aspects of delivery "performance" are more or less in increasing order with respect to the degree of direct linkage between the variable and the activities of the individual pilot. Flight path is most directly responsive to what the pilot does at a moment in time; release point is a function both of the immediate inputs from the pilot and of the effectiveness of planning for the future flight path, since some errors accumulate over the course of performance and restrict the pilot's options. Impact point is only indirectly linked to flight path events since any deviation from actual impact and that predicted from release point is error for purposes of individual skill measurement.

Release point and path are also the sole determinants of the reliable variance in deviation of impact point from target. It was noted repeatedly above, and by many of the authors previously cited (e. g., Dunnette & Borman, 1979), that such terminal outcome measures have several attributes that make them "bad" measures in terms of the criteria defined in the previous section. Their reliability is low because they are influenced by numerous factors not under control of the individual; they are not diagnostic of where training is needed; their validity is low because (among other reasons) they confound operator and system performance. For most of the purposes for which measures are intended, it is necessary to look to some aspect of the process by which an outcome is achieved, rather than to the outcome itself.

Process variables have some practical disadvantages as measures. Systems that rely heavily on operator judgments that are not well understood are not always well suited to process measures (with the exception of a special kind of process

variable, the "proxy" measure to be discussed later). It may take considerable prior "homework" to derive candidate measures for tasks with extensive decision-making components. Even when such data are available, it is often not immediately obvious how the process of performance can be decomposed into appropriate intermediate components that are both reliably quantifiable and obtainable cost effectively.

In earlier discussions, the work of Brictson and his colleagues was pointed out both as a model of careful measure development and an illustration of the time and effort required to produce measures that have all the desired properties. They decomposed the carrier landing process into a series of "waypoints" in the approach, collected radar data for many hundreds of approaches, and analyzed the "patterns" at each waypoint for a number of conditions of the task (weather, day or night, carrier class) and for characteristics of individual pilots (overall experience, type of experience, recency of experience) that had potential for affecting the observed patterns (see Brictson, et al., 1971).

By the conclusion of the series of studies, the approach determinants and task conditions that "made a difference" in the outcome of a carrier landing were sufficiently well defined to begin the process of systematically eliminating process measures that were less important to achieve a more readily usable summary measure. As Brictson et al. (1973) noted, their full measuring system required considerable manpower and cost resources to use, and, while that was acceptable for a single study or evaluation, it virtually precluded routine acquisition of data for continuing feedback on training (the original measurement objective). After "working backward" from the end of the approach to understand the process, they then "worked forward" within the process to find a simpler, more readily obtainable measure that was sufficiently sensitive and reliable

and retained acceptable "validity." The Landing Performance Score which emerged was simply an empirical weight assigned to the arresting gear wire "caught" on a particular landing. In terms of the criteria defined above, the "utility" of the measurement system was enhanced. It is compelling in retrospect (but incorrect) to suggest that wire number as a measure could have been selected a priori without the time-consuming examination of hundreds of landings. To have done so without the understanding of the process would provide no basis for representing the wire number as a measure of anything; evidence for validity of the simpler measure was derived only through the operation of process decomposition.

The procedures employed in the Brictson et al. studies also illustrates another complication in the development of process measures -- no two people use exactly the same processes to perform a task. It has been discussed in several previous sections that different people arrive at equally effective performances in different ways, and may not even be consistent about how they do the task from one performance to another. Connelly (1982), in discussing weapon delivery measurement, describes a "confluence of paths" that can be flown to an acceptable "hypersurface" of release points. Each point on the hypersurface produces a satisfactorily accurate delivery. Connelly notes that there is no single path to be derived, and that the conceptual objective of a measure system is to define the envelope of paths and the hypersurface of release points that lead to successful performance. His concept is functionally equivalent to that employed by Brictson et al. in plotting and examining the coordinates at each waypoint of successful and unsuccessful approaches.

Some parts of any process have more influence (greater "causal" effect) than others. There is, in any decomposition of task sequences, a common thread of identifying the

activities or points in a task which actually determine ultimate outcomes. For carrier landing and weapons delivery, process importance involves an element of proximity to the end point of the task. Lyon et al. (1980) showed that activity during the "last few seconds" of a delivery maneuver was the most important predictor of delivery accuracy. In the Brictson et al. studies, as the end point of an approach got closer, there were progressively smaller windows within which the pilot was free to vary if the landing was to be successful. The predictability of outcome from position at a waypoint systematically increased as distance from touchdown decreased; early positions related less well to outcome because, within the population of trained career aviators, it matters very little where the pilot is early in the approach.

Such proximity relationships would not be expected for other performances. In some tasks, ACM for example, the outcome is sometimes determined from actions very early in the sequence. If a pilot fails to get first visual acquisition of the adversary aircraft, he is unlikely to win, even though the process of the task may continue for some time past that point (Ciavarelli, 1982). It is clear from the above that there are few generalizations that can be made about important segments in tasks. Each task must be analyzed and decomposed based on data about the impact of each segment on the eventual outcome (presumably a reliable one).

PROXY VARIABLES

There are, in many situations for which measurement would be useful, a whole host of factors, controllable and uncontrollable, which have an influence on the ultimate outcome. Some of these can be isolated and identified, others are likely to remain undetected despite the most thorough analysis. For many tasks, there are certain specific variables

available at points within the process which are sensitive to all or most of these outcome-shaping factors. When specific data on underlying factors is impractical to obtain, or the effect of factors can not be conveniently isolated, such internal variables can sometimes serve as what economists call "proxy" measures.

A proxy measure is a single quantity that reflects the combined effects of all the important determinants of performance on a complex task and is influenced in consistent ways by all such relevant sources of variation. It can be viewed as a composite variable which aggregates with unknown weights some unknown components of performance, both those which might be separately measured and those that are "invisible" in the process. Proxies are thus composites, but are "naturally existing" ones, differing from those constructed by deliberate weighting of separately determined component measures.

Proxy variables are relatively common. Any outcome variable is a proxy measure. The Gross National Product (GNP) is a proxy for numerous underlying and unknown processes which affect the value of goods and services. Many of the measures discussed in previous sections function as proxies. Impact point is a proxy for variables such as aircraft status (airspeed, bank, skid, etc.) at release, the release point itself, and the environmental conditions at the time of release, as well as the inherent accuracy of the weapon system. Release point is in turn a proxy for all the variables, including individual proficiency, that control the pilot's arriving at the right point at the right time. The Landing Performance Score of Brictson et al. (1971) and Brictson et al. (1973), the weighted wire number, is a proxy for the numerous process variables examined by the authors. In one sense, any subjective rating which is based on the "right"

observations, and which maps those observations in a consistent way into the assessment, is also a proxy, but it is not a process variable.

Variables which act as proxies for unknown processes are frequently seen in the experimental literature and in measurement for system evaluation, both with and without an awareness of the measures as a reflection of combined processes. Bahrick, Fitts and Briggs (1957) addressed one of the key aspects of what is termed here as proxy measures in their explicit attention to finding the best single measure of performance on multiparameter tracking tasks. They concluded that RMS error best reflected the underlying processes in a consistent way. Similar awareness of the need to find a single score that combined other less accessible measures is seen in the treatment of RMS by Mengelkoch, Muckler and Monroe (1959). Billings and Eggspuehler (1970) focused on variability in helicopter rotor RPM as an appropriate single measure for studying the effects of experimental conditions. They noted that RPM variability reflected pilot skill automatically adjusted for maneuver difficulty (removing the need to use difficulty as a factor in analysis), and was also sensitive to differences in stress effects.

Another form of proxy variable is seen in the logic of adaptive training and adaptive measurement. Adaptive logic by its nature depends on a single measure as an indicant of how well the task is being performed (Kelley & Kelley, 1970; Kelley & Prosin, 1968). The structure of the adaptive approach is to hold that performance constant, adjusting the system or task conditions as required to make the task easier or more difficult. The adaptive variable must be one which is sensitive in an orderly way to all the factors which can be "adapted" to hold performance at a constant level, and as such is clearly an aggregate measure functioning in the role of proxy as defined above.

There are distinct advantages in using the concept of proxy variables as a paradigm for thinking about the composition of a measure, even one in common use. It clarifies to an extent the need to determine the components of factors which cause the proxy to vary. Some forms of proxies are well suited to methods of analysis which decompose their variance into components by systematically relating other variables from within the process to the selected proxy variable (correlational or discriminant analyses, path analysis, etc.). Such decomposition can indicate which events from within the performance sequence matter most in determining outcomes. Similar procedures were used in the Brictson studies, in those reported by Ciavarelli (1982) and in the approach suggested by Connelly (1982).

SURROGATE MEASURES

It sometimes happens, despite the most careful efforts, that satisfactory measures of performance simply cannot be obtained because of resource constraints or because of the nature of the setting in which the phenomemon must be measured. This is particularly likely in operational or field environments. Prior discussions have defined mechanisms that can lead to the very low reliabilities typically encountered in operational measurement. The phenomenon of interest may be too unstable to assess without many repeated measures, but the nature of the task (activities not readily observable) or the cost of each data point (targets, ordnance, ranges, etc.) may preclude multiple performances. It may also not be possible in field settings to employ instrumentation and data recording equipment that would improve accuracy.

It was noted above that a reliability of .30 is high for field measures, and .00 to .10 is typical for a single

performance, leaving at least 90 percent of the field measure (more likely 98 or 99 percent) that cannot be related to any other measure since it does not relate to itself. An increasingly common use of field measures is for determining the effect on performance of unusual task conditions or stressor agents (drugs, altitude, chemical warfare agents, etc.). It is obviously important to know, for example, if a drug intended to neutralize the effects of chemical agents or equipment to protect against those agents causes greater disruption in performance than the agents themselves. Given the almost certain very low reliabilities of field measures, the power of those measures to detect real differences among conditions is for all practical purposes nonexistent. Further, even if the performance measure is theoretically of high validity, its low reliability places a mathematical limit on its validity in use that is far below its theoretical value. Consider a measure with an empirical reliability of .10 (not unlikely for outcome variables). Suppose further that its "real" validity, defined by its theoretical relation to the hypothetical performance construct which it estimates, is relatively high, say .70. Since the correlation of the construct with itself is perfect (one of the nice things about constructs), its reliability is 1.00. Using the well-known equations for correction of correlations for attenuation (unreliability), it can be shown that the maximum value for an obtained validity is .22. Because of the unreliability of the measure, no more than 4.8 percent of its variance can be "valid," i. e., held in common with the "true" performance on the task. It would require an extremely powerful effect for conditions or stressors to be detectable with so little overlap with true performance.

Kennedy, Jones and Lane (1986) further develop the logic of the above arguments and expand on the psychometric basis underlying the limited sensitivity of unreliable measures to

detecting shifts in true performance. They introduce the concept of "surrogate measure" systems for use in situations when operational measures cannot be measured with acceptable reliability. Surrogate measures are those which are related to or predictive of a construct of interest (such as "true" field performance), but are not direct measures of that construct. They involve the use of tests or batteries that are specially developed to have four major characteristics: They must a) "correlate" reasonably well to the performance construct, b) be sensitive to the same factors that affect the unobtainable performance, i. e., they change in the same way in response to varying conditions as the performance variable would if it were accessible, c) be much more reliable than the field measures, and d) involve minimal time to learn, so that they can be used without extensive practice. Kennedy et al. show that for any estimated values of a field measure's reliability and true validity, the reliability and overlap with the true performance required of surrogate measures can be determined.

The development of surrogates involves the demonstration of measure properties through a series of operations similar in nature to those required for establishing construct validity. Surrogates differ from proxies in that, while they are (hopefully) sensitive to the same factors, they are entirely separate from the process of task performance itself. They differ from conventional performance measures in that the tests need not involve operations in common with the performance measures, only components or factors in common. They also differ on similar grounds from another class of measure substitutes, those involving "synthetic task" or controlled job sample approaches, in that a key aspect of surrogate development is that the tasks or tests on which measures are taken require little time to learn, so that practice effects are not introduced into repeated administrations. Synthetic or job-sample tasks often require considerable practice for

operators to master, and are in addition likely to be as difficult to "score" as the field performances which they emulate.

PERFORMANCE MEASURES AND "MOE'S" AS PROCESS VARIABLES

There are recurring tendencies in the literature to treat as if they were equivalent two kinds of "measures" with somewhat different meanings. Previous discussions have repeatedly noted that the term "performance" measure should be reserved for application to assessments of individual proficiency. Performance measures for such use involve different evaluation and validation operations than those for what are typically called "measures of effectiveness" (MOE's). Both, however, can be viewed as "process" variables as the context of observation is expanded.

MOE's can be considered within the validation paradigms previously described as system-level "performance" measures (how well does the system taken as a whole do its job), and are appropriately used in that context. The "system" involved is ordinarily a single operator using one aircraft or weapons system to perform one task. The distinction between conventional performance measures (i. e., for individual use) and MOE's illustrates the importance of context of validation. Many traditional MOE's are of the class of outcome variables (impact error, detection probability) that have been shown to be of near-zero reliability and limited diagnosticity as individual measures. As defined in earlier discussions of measurement properties, such variables contain large components of error and irrelevance (for individual performances) which tend to average out when measures are aggregated to the next-highest level. The elements of system contribution to variance which harms the validity of performance measures are appropriately included in an MOE for that system taken as a

single human/machine unit, and such MOE's can appropriately be pooled over units and groups, gaining reliability from the accumulation of scores. Cavaluzzo (1984) reported that impact error scores for weapon delivery were satisfactorily consistent and sensitive measures of squadron effectiveness when aggregated across multiple deliveries for a number of pilots over a period of time. Such measures are not, however, generalizable to the sorts of individual and group measures ordinarily considered as performance measures, unless they are separately validated for that purpose.

While MOE's as traditionally employed may not be appropriate as measures of individual proficiency, that does not, however, invalidate their use as potentially important measures in the context of the system in which they are defined. Any observation or data that can be gathered about a system during its operation can, as we have noted above, be viewed as a process variable within some context. The choice of the context to be used for organizing observations about processes determines the appropriate employment of an MOE.

Systems, particularly as they are defined in the world of military operations, exist in a hierarchy of degrees of abstraction ranging from the single human performing a single task to the complete battle plan for a fleet operation. Each level in the hierarchy has distinct measures which reflect how well the system is functioning, and each of these measures is a process variable within a still broader context. The mission as defined for a given pilot on a single weapons delivery run is embedded in a larger context for that pilot performing a single ground attack mission of multiple attacks, which is in turn embedded in a squadron-level ground attack mission of multiple units making multiple runs, still further aggregated into multiple squadrons for sustained periods of performance, perhaps combined with a coordinated ground operation against

the same targets. Measures which are not satisfactory for estimating effectiveness at one level of that hierarchy may be the correct ones for the next level of aggregation. Thus the "outcome" measures derived from pilot/system combinations, which were shown to be poor measures for the individuals, are important numbers for assessing the combined effectiveness of pilot and aircraft.

MEASUREMENT OF AIR COMBAT MANEUVERING PERFORMANCE

Most of the information and discussion in the preceding sections of this report has dealt with general measurement issues. There has been a modest tailoring of comments and illustrations to focus more heavily on measuring performance for complex tasks in aviation settings, particularly those in operational or near-operational contexts. In this section, the focus is further narrowed to one of the most challenging and technically complex areas of aviation, that of ACM. Discussions will not describe the process of ACM and its associated geometries (Breidenbach et al [1985] and Wooldridge et al. [1982] provide thorough descriptions), nor will they review in depth the existing ACM measurement systems (Breidenbach et al.[1985] and Vreuls et al. [1985] taken together summarize essentially the complete literature on that topic). The intent is rather to relate the broader measurement considerations brought out in previous sections to the special measurement problems encountered in ACM.

Although interest in the assessment of air combat proficiency goes back to at least World War II (Henneman, 1946; Jenkins et al., 1950), the lack of technology for data recording limited measurement efforts to partial measures such as gunnery performance, to global judgments and ratings, and to clinical evaluations of "personality factors" important in combat effectiveness. Youngling et al. (1977) review the extensive literature from World War II and Korea on these various initiatives. It was only in the late 1960's that the beginning developments in air-combat simulators and the capability to record and process large amounts of positional data from airborne instrumentation and ground radars made possible algorithmic attempts to capture and analyze the maneuvering activity of fighter pilots.

THE ACM MEASUREMENT PROBLEM

Determining how proficient a pilot is at air combat is an extraordinarily complicated task. In ground attack missions (for example) the "target" is relatively fixed; in air combat, the adversary moves with approximately the same agility and speed as the attacking aircraft, and the task conditions, by the inherent nature of the task, are changing every fraction of a second. Engagements involving multiple attackers and multiple adversaries (so called "many vs. many") are common, and the activities of a pilot which are appropriate in a single engagement (a "one vs. one") may reflect a serious lack of proficiency in a many vs. many hassle. To further complicate the problem, the weapons available to a pilot (rear-hemisphere, all-aspect, multiple-targeting capability) and to the adversary will also determine the tactics that constitute "good" performance in a specific engagement. The "rules of engagement" can also change as a function of the degree to which the pilot perceives or is instructed in the importance of platform survivability vs. pressing home the attack at risk of platform and pilot loss, and the definitions of "proficient performance" will be modified accordingly.

Measures taken in the ACM environment also suffer from the problems of being very nearly the "ultimate criterion" for fighter pilot proficiency. For many other measurement situations, there may be measures from a later stage or from controlled simulator studies against which measures may be compared. This is not the case for the bulk of ACM measurement requirements. There are thus serious limitations on the ways in which measure sets intended for use in ACM can be validated. It is a costly environment in which to study a phenomenon (repeated measurements may be too expensive) and a constantly changing one (conditions are often uncontrollable and rarely repeatable). Under such conditions, examining the "performance construct" in conventional ways may be difficult if not impractical.

ACM COMPONENTS

The geometries and mathematics of aircraft maneuvers in air combat are immensely complex, and the various algorithms developed to convert maneuver data to measures of ACM performance are correspondingly complicated. Breidenbach et al. (1985) review and illustrate in depth these varying geometries, and describe the quantities of interest that form the potential base for measurement. They, along with most workers in the field, identify two critical components of success in ACM. The Angular Geometry Component (AGC) reflects the relative spatial orientations of fighter and adversary with respect to a "favorable" firing position, based on the presumption that it is better to have the adversary within the angular envelope of a weapon than to be within his angular envelope. AGC is derived in turn from two quantities, Angle Off the Tail (AOT) and Antenna Trial Angle, which indicate respectively the degree to which a fighter is "behind" the adversary and the relative bearing of the flight paths of the two aircraft. AGC indicates whether or not one aircraft or the other is in a position to successfully fire a weapon with respect to angular orientation separate from distance. Weapon Range Component (WRC) indicates the distance between aircraft relative to the capabilities of their weapons systems, i. e., are they close enough for successful firing.

Both AGC and WRC are affected in part by a third component in a somewhat different "dimension," the extent to which one pilot manages his available "energy" resources better than the other. An aircraft at altitude has potential energy which can be converted to airspeed by "diving." Likewise, an aircraft with greater airspeed than required can trade that airspeed for altitude by "climbing." There are complex "energy envelopes" within which a given aircraft can operate at a moment in time

that are a function of its design characteristics and its state at that moment. There are specific profiles for turning, increasing or decreasing altitude, and so forth, based on potential energy possessed, power setting, wing loading and other variables. Determining how a pilot's actions relate to those profiles reveals the extent to which he is making most effective use of the energy available to him at any point within the flight regime.

Displaying the changes in an aircraft's location within its energy envelope with the Energy Maneuverability Display (Moroney, Pruitt & Lau, 1979; Pruitt & Moroney, 1980) has been used for some time as a method of feedback to pilots to improve ACM technique. Because energy management reflects (is a proxy for) the effects of so many aspects of ACM capability (knowledge of aircraft limits, preciseness of airwork, platform-specific tactics, etc.), it has also been suggested as a potential basis for measuring ACM proficiency. Breidenbach et al. (1985) expand on this concept of quantifying the efficiency of energy utilization as a supplement to the AGC and WRC components, with primary emphasis on energy-use measures as aids to diagnosis and training.

ACM MEASURES AND MEASUREMENT SYSTEMS

Breidenbach et al. review and compare in a common framework four measurement models and variations on each: a) The Maneuver Conversion Model of Oberle (1974) and some historical descendants (Oberle & Naron, 1978; Oberle, 1983); b) the Performance Index (Simpson, 1976; Simpson & Oberle, 1977; Oberle, 1983); c) the All-Aspect Maneuvering Index (AAMI) (McGuinness, Bouwman & Puig, 1982; McGuinness, Forbes & Rhoads, 1983); and d) the TACSPACE model of Wooldridge et al. (1982). They describe the approaches used in each model and contrast the strengths and weaknesses of the models and the measurement scales they produce; those discussions will not be repeated here.

Essentially all of the measurement systems reviewed by Breidenbach et al. are "positional advantage" models, that is, they equate good performance to being within a firing envelope or in an offensive posture as much of the time as possible. As such, they use considerably more of the data from an engagement than measurements that utilize kills and losses or the kill/loss ratio, or indices such as time to kill/loss. Such outcome and time measures, in addition to the known problems with outcome data, are almost completely insensitive to the dynamic and evolving nature of an engagement, and provide little diagnostic assistance. (Note that measures such as "time to first kill" are what has been called earlier a proxy variable for the dynamics of the engagement process). The measures developed from the ongoing engagement are likely also to have better metric properties. The broader base of data used for measurement affects validity through the mechanism of measure comprehensiveness, and the compilation or averaging of measures over a greater number of "estimates" (more time slices) can have a major impact on reliability.

An additional positional-advantage model not reviewed by Breidenbach, but which was seminal to many later approaches, is that of Burgin and Fogel (1972). It, like several later efforts (i. e., TACSPACE), is a state-transition model. Situation matrices are constructed defined by a series of aircraft status variables important to successful outcomes (ahead/behind, visible/not visible, in/out of firing position). A "payoff" is assigned to each cell, reflecting the value of being in that state for achieving a successful outcome. A series of aircraft state variables (attitudes, energy states) are evaluated to determine into which cell each of the possible actions that are available would place the aircraft if those actions were taken, and the action or maneuver with the highest expected value is implemented. The model was originally developed as a computer adversary to fly against a pilot on a simulator. It apparently

performed its task so well that it was necessary to "detune" the model in order for the human pilot to be competitive in the engagement. Although not initially intended as a measurement model, the potential for extracting values which reflect correspondence of human decisions to "ideal" ones is obvious.

Not all the indices of ACM performance use positional advantage information exclusively. The Good Stick Index (GSI) (Moore et al., 1979) is a composite score of ACM performance on gun-equipped aircraft (vice missiles). It is a linear combination of simulator measures on a) time in the pointing angle envelope, b) average error inside kill range with trigger depressed, c) time in offensive vs. defensive posture, and d) time to first kill. Note that components a) and c) involve positional advantage information.

Jenkins (1982) compared the GSI and other simulator-based measures from training to inflight outcomes in a Weapons Instructor Course to examine the influence of simulator training on performance. Scores in the course included several different ratings by instructors, gunfilm records of valid gun/missile firings, and "exchange" (kill/loss) ratios over a series of engagements. Jenkins also examined state transition probabilities for measurement potential using a structure similar to that of Oberle (1974). Of the conventional measures, there were no differences between simulator and non-simulator groups in GSI, instructor ratings or exchange ratios (although the latter were indicative and might have reached significance in a larger sample). Only the gunfilm records showed consistent differences. Several of the transition probabilities obtained from the simulator were judged to be "promising" for use as measures of inflight performance, but the insensitivity of inflight measures precluded further exploration of the probability-as-measure approach. Jenkins also concluded that "...Attempts to quantify success in AA [air-to-air] combat can

lead to an oversimplification of the problem as a result of measurement of only a few specific tasks or parameters" (brackets added).

An important analysis of factors leading to successful ACM outcomes was reported by Ciavarelli (1982) based in part on earlier analyses (Brictson, Ciavarelli & Jones, 1977; Ciavarelli, Pettigrew and Brictson, 1981). In this effort, the sequence of events or critical tasks in ACM engagements was systematically decomposed in the same way as for the carrier landing breakdown of Brictson et al. (e. g., Brictson, Ciavarelli & Wulfeck, 1969). Such intermediate events as initial radar contact, visual identification, first-shot opportunity, etc. (the "process" variables) were examined for their impact on later events and on ultimate engagement outcomes. Success probabilities resulting from possible event sequences or "paths" were determined.

Such systematic relationship of processes to later processes and to outcomes serves a twofold purpose: It allows a direction of training emphasis toward the "main drivers" of success. It also focuses efforts in measure development on those events or tasks that are critical to record or observe, and thus delimits what is otherwise a massive data collection and reduction problem. Many of the critical events determined in the Ciavarelli analysis (first contact, visual detection, etc.) have the advantage of being objectively verifiable without complex instrumentation, and thus serve as potential proxy measures when more precise data are unavailable. Because of the distinct differences among event paths in the probability of success, the event paths themselves could serve as "proxy" measures of the eventual engagement outcome, even when the outcome itself might be indeterminate.

A MEASUREMENT APPLICATION FRAMEWORK

The analyses of ACM performance measurement requirements by Breidenbach et al. (1985) and by others (Stoffer, 1981) place particular emphasis on the diagnosticity of measures for indicating both relative "goodness" of performance and the causes of any performance deficiencies. Breidenbach et al. address in detail the importance of usability and ready accessibility of ACM measures by both novices and instructors. They advocate attention to choosing a sound framework for reporting and displaying measurement results, distinct from but no less important than the development of the measurement system itself. They recommend a structure similar to the Performance Assessment and Appraisal System (PAAS) (Ciavarelli, Pettigrew & Brictson, 1981; Ciavarelli, Williams & Pettigrew, 1981).

The PAAS, in its broadest conceptualization, is a method for storage and retrieval of ACM measures in formats that relate individual performances to a variety of different kinds of comparison information to allow those performances to be viewed in readily interpretable contexts. A performance on some intermediate task might be, for example, a) compared to a specific objective that represents the acceptable value for that task, or b) displayed in relation to the distribution of performances by other individuals of similar experience on that task, or c) compared to that particular individual's performances on previous engagements, or d) an individual's standing on that task relative to other individuals, compared to his standing on other important tasks, to indicate particular strengths or weaknesses.

MEASUREMENT ON SIMULATORS AND INSTRUMENTED RANGES

All of the measurement systems and supportive research described above are intended for use either on ACM simulators or on specially instrumented aircraft flying on instrumented ACM

ranges (in some cases on both). The advantages of simulator measurement have been well understood for many years (Grodsky, 1967; Smode, Gruber & Ely, 1962). Environmental and task conditions can be controlled, target behavior can be standardized and scenarios can, if desired, be repeated. Properly developed simulator measurement systems can be highly effective in evaluating trainee progress; simulator measurement capability offers as well the opportunity for controlled evaluation and "shakedown" of a measurement system intended for range application. It is clear from the management of ACM training within the military services, however, that simulators, despite their value for learning ACM skills, are intended primarily as an augmentation to conventional ACM practice time rather than as a substitute. The commitment of the operational forces to evaluating capability but to attaining that capability under near-operational conditions is reflected in the investment in the Navy Tactical Air Combat Training System (TACTS) and the Air Force Air Combat Maneuvering Instrumentation system (ACMI).

The TACTS/ACMI systems acquire data through instrumentation pods attached to each aircraft, which extract and transmit information about flight dynamics, weapons firing and other aircraft-specific parameters to a series of ground tracking stations. The ground stations receive and forward aircraft data to a master tracking station. Aircraft operate in a controlled airspace. Data from the pods and the tracking stations allow for the precise positioning of each aircraft within the airspace and for the recording of a variety of flight parameters that describe in detail what the aircraft is doing at each moment in time within its maneuvering regime. Other subsystems of TACTS/ACMI reduce data and display maneuver histories and outcomes for debriefing purposes. Detailed descriptions of the TACTS/ACMI systems are given by Breidenbach et al. (1985) and by Hooks and Kress (1984), who also discuss some specific applications of range data for measurement purposes.

COMMENTS ON EVALUATING ACM MEASUREMENT SYSTEMS

The extraordinary capability of the TACTS/ACMI to produce accurate and precise data about fighter and adversary position and location offers powerful opportunities for measurement of individual ACM skills. Although a number of the efforts identified above have generated candidate approaches to quantifying ACM performance, the potential offered by the ranges has thus far not been fully realized. This has resulted in part from the difficulty of developing and validating comprehensive measure sets for a skill as multidimensional as ACM in a multiple adversary environment, and in part from the willingness of system users to rely on subjective interpretations of mission replay displays to provide debriefing guidance. Several of the measurement approaches noted above appear to be tapping important aspects of how well the pilot uses his machine for ACM, and have the potential for yielding acceptable metrics, but none to date has been examined for measurement properties in systematic paradigms like those suggested in earlier sections.

THE VALIDATION PROCESS

The process of validation in ACM is perhaps the most difficult measurement challenge within aviation. Discussions above have noted the complexity and multidimensionality of the task and the degree to which "good" (successful) performance is contingent on events that evolve dynamically within a given engagement. Outcome of a single ACM flight, for example, is in part determined by the skill of the attacker and in part by the skill of the adversary, in addition to such factors as visibility, sun position, and the usual day-to-day variations characteristic of performance measures in general. The "ultimate outcome," i. e., who wins the engagement, is thus not a particularly reliable measure taken by itself, and it does not

serve well as an "ultimate criterion" against which to validate an ACM measure system. It is of course possible, and desirable, to pool outcomes over a series of engagements under varying conditions, increasing the stability and reliability in accordance with the greater number of "samples" of performance. This allows a greater confidence in relating variables in a measure set to outcomes as part of the validation process, but does not improve the utility of single engagement measures; they still retain the instability disadvantages of single measures.

THE PURPOSE OF THE MEASUREMENT SYSTEM

One way to view the ACM measurement problem is from the standpoint of primary purpose of a measurement system. Measures on the ranges are most appropriately used for diagnostic purposes, that is, to identify any weak areas in a pilot's tactics or responses. From this viewpoint, a measure system should be strongly process oriented; the most valuable information for diagnosis is that which indicates whether or not a "correct" or "optimal" response was made to the specific set of conditions in force at some specific point within the evolving engagement. The best (most "valid") measurement system will be one which takes into account, with the correct emphasis, all the important variables involved in a pilot's momentary decision on what to do next, and compares his decision to some template of what his "best" action would have been.

Such comparisons are inherently probabilistic; for most of the hundreds of individual "decisions" made by a pilot during an engagement, the optimal choice is not clearcut, and different actions may vary in their influence on ultimate outcome only in subtle ways. Although there are certain key events that can determine the outcome regardless of later activities, on the average the pilot who can consistently identify and execute the higher probability action in response to conditions at each

decision point is more likely to win the engagement. Measure systems which involve "payoff matrices" (Burgin & Fogel, 1972), or state-transition probabilities to more favorable or less favorable states (Oberle, 1974), tend to incorporate such "correctness of decision" information, and are likely to be more sensitive to the sustained quality of performance throughout the engagement. Outcome measures may represent either a high sustained performance or the result of a single key event that determines an outcome even when performance during the balance of the engagement was not particularly proficient.

A SUMMARY OF ISSUES

Evaluation of an ACM measure system must be approached with particular care:

Reliability

Because of the multitude of factors that influence ACM performance, reliability of measures is especially critical, and no proposed system for ACM measurement has to date determined or reported reliability values.

Focus on Intended Purpose

As previous sections have repeatedly emphasized, the operations appropriate to validation are to a great extent determined by and derived from the intended purpose or proposed usage of the measures.

Diagnosis and remediation -- If measures are to be used for detecting and remedying particular skill deficiencies in individual pilots, they must detect aspects of performance that are the important ones on two bases. The diagnostic measures should matter in determining ultimate outcome (but recall the

metric problems with outcomes), preferably over a sequence of several engagements, and they should "flag" inappropriate tactics or poorly understood procedures that may not be crucial in the outcome of some specific engagement but constitute "bad habits" that will over time reduce chances of success or survival. Further, each diagnostic measure should desirably be associated with a logical means of correcting the deficiency, i. e., the system should diagnosis "problems" about which something can be done.

Overall proficiency -- If measures are intended for assessment of overall individual skill, outcomes pooled over engagements may be sufficient, even though considerable information is lost about the processes involved in the win or loss, i. e., all "winners" will appear equally proficient even when there are clear differences in how the outcome was achieved.

Minimum standards -- If measures are intended for establishing some minimum level of capability to produce desired outcomes, further relaxation of validation requirements is allowable. As noted earlier , minimum standards use of "measures" requires only that the measure used be a satisfactory representative of the skill being "checked," and that the minimum performance established be appropriate to levels of proficiency expected from the population at that point in time.

Context of Validation

Whatever the purpose of the measure set, it should be emphasized that the set is only properly applied within the context of that purpose for which it is validated. Diagnostic measures are not necessarily good or appropriate measures of overall proficiency, and conversely, proficiency measures based on aggregation of outcomes do not necessarily indicate the presence or absence of potentially serious skill deficiencies for an individual.

Usability of Measures

A further consideration in validation of measures based on purpose is the nature of the scores yielded by the measure system. Intended use of measures implies "use by whom." The scale and level of aggregation of measures should match as closely as possible the needs of the potential user of the information contained in the measures. Separately from the appropriateness of use, instructors are not materially aided in identifying and correcting individual deficiencies by outcome scores or other measures which are combined or aggregated at a level which reduces understanding of the score in terms of the direct activities of ACM. As Breidenbach et al. (1985) note, some measure sets, particularly those involving reduction in the number of measures by the creation of statistical composites (factor or discriminant analysis), may produce scores which are satisfactory measures of performance but are essentially uninterpretable by instructors and thus are not helpful in improving the training process.

Verification Outside the Development Sample

Finally, any measure set, whatever the apparent relevance of its content or definitional operations, or the validity values demonstrated within a sample, must be "cross-validated," i. e., the operations used to determine validity must be repeated in another sample. This is particularly essential for measure sets derived from methods of statistical reduction or combination. The credibility of measures is not defined by evidence from the sample on which the measures were developed.

THE "CONSTRUCT(S)" OF ACM PERFORMANCE

Earlier sections on the important properties of performance measures dealt at length with the need to understand any task

"performance" as a construct and to "validate" proposed measures. Validation was achieved through the accumulation of evidence which would predispose a rational observer to both a) believe in the existence of the performance construct as defined by the measurement operations, and b) believe that the measurement operations do in fact provide an estimate of that performance. It was noted that validation of the construct is a relative weighting of evidence; there are no coefficients which indicate the believability of a measurement system.

There are several different aspects of ACM performance to which a measurement system might be addressed. Each of these purposes of measuring (diagnostic, overall proficiency, etc.) requires in a sense a separate "performance construct," and thus involves a somewhat different sequence of operations to establish the credibility of measures. The problem of measurement in ACM differs from that in other settings primarily in the complexity of the phenomenon. The diversity and quantitative sophistication of approaches presented to date suggests that the technology of ACM measurement is still on the margin of conceptual, economic, and implementation feasibility. It is important to recognize that there are as yet no "magic bullets" for untangling the ACM measurement problem. The different meanings of "performance" implied by the various "constructs" identified above suggests that no single measurement system is likely to satisfy all the ACM measurement needs. It is clear from the literature that different proposed systems have been focused on somewhat different constructs, and cannot be compared on the same criteria of effectiveness. The use of the term "ACM performance" as if it were a unitary concept may be a critical obstacle both in defining what should be in a "measurement system" and in communicating the worth and applicability of that system to the user.

REFERENCES

Allen, M. J. & Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks-Cole.

Alluisi, E. A. (1967). Methodology in the use of synthetic tasks to assess complex performance. Human Factors, 9, 375-384.

Alvares, K. M. & Hulin, C. L. (1973). An experimental evaluation of temporal decay in the prediction of performance. Organizational Behavior and Human Performance, 9, 169-185.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. Washington, DC: Author.

American Psychological Association. (1985). Standards for educational and psychological tests. Washington, DC: Author.

Anastasi, A. (1986). Evolving concepts of test validation. Annual Review of Psychology, 37, 1-15.

Anderson, J. R. (1982). Acquisition of cognitive skill. Psychological Review, 89, 369-406.

Ash, P. & Kroeker, L. P. (1975). Personnel selection, classification and placement. Annual Review of Psychology, 26, 481-507.

Bahrick, H. P., Fitts, P. M. & Briggs, G. E. (1957). Learning curves - facts or artifacts? Psychological Bulletin, 54, 256-268.

Baum, D. R., Smith, J. F. & Goebel, R. A. (1973). Selection and analysis of UPT maneuvers for automated proficiency measurement development. AFHRL-TR-72-62. Williams AFB, AZ: Air Force Human Resources Laboratory.

Ben-Avi, A. (1947). Studies of subjective measures of flying proficiency. In N. E. Miller (Ed.), Psychological Research on Pilot Training. Army Air Force Aviation Psychology Program Report No. 8. Washington, DC: Government Printing Office.

Billings, C. E. & Eggspuehler, J. J. (1970). Studies of pilot performance in helicopters. OSURF-7857-6. Columbus, OH: The Ohio State University.

Bilodeau, E. A. & Bilodeau, I. McD. (1961). Motor skills learning. Annual Review of Psychology, 12, 243-280.

Bingham, W. V. (1949). Today and yesterday. Personnel Psychology, 2, 267-275.

Bittner, A. C., Carter, R. C., Kennedy, R. S., Harbeson, M. M. & Krause, M. (1984). Performance Evaluation Tests for Environmental Research (PETER): Evaluation of 112 measures. Report No. NBDL-84R0006. New Orleans, LA: Naval Biodynamics Laboratory.

Bottenberg, R. A. & Christal, R. E. (1961). An iterative technique for clustering criteria which retains optimum predictive efficiency. WADD-TN-61-30. Wright-Patterson AFB, OH: Wright Air Development Division.

Breidenbach, S. T., Ciavarelli, A. P., Sievers, R. & Lilienthal, M. G. (1985). Methods and metrics for aircrew assessment during close-in air-to-air combat. R-85006. San Diego, CA: Cubic Corporation.

Brictson, C. A., Burger, W. J. & Kennedy, R. S. (1971). Predicting the quality of pilot performance during night carrier recovery. Aerospace Medicine, 42, 16-19.

Brictson, C. A., Burger, W. J. & Wulfeck, J. W. (1973). Validation and application of a carrier landing performance score: the LPS. Santa Monica, CA: Dunlap and Associates.

Brictson, C. A., Ciavarelli, A. P. & Wulfeck, J. W. (1969). Operational measures of aircraft carrier landing system performance. Human Factors, 11, 281-290

Brictson, C. A., Ciavarelli, A. P. & Jones, T. N. (1977). Development of aircrew performance measures for the air combat maneuvering range (U). NAMRL L53001. Pensacola, FL: Naval Aerospace Medical Research Laboratory (Confidential).

Brictson, C. A., Hagen, P. F. & Wulfeck, J. W. (1967). Measures of carrier landing performance under combat conditions. Santa Monica, CA: Dunlap and Associates.

Buckhout, R. & Cotterman, T. E. (1963). Considerations in the design of automatic proficiency measurement in simulators. AMRL Memo P-40. Wright-Patterson AFB, OH: Aerospace Medical Research Laboratories.

Burgin, G. H. & Fogel, L. J. (1972). Air-to-air combat tactics synthesis and analysis program based on an adaptive maneuvering logic. Journal of Cybernetics, 4, 60-68.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Carter, L. F. (Ed.) (1947). Psychological research on navigator training. Army Air Force Aviation Psychology Program Report No. 10. Washington, DC: Government Printing Office.

Carter, L. J. & Dudek, F. J. (1947). The use of psychological techniques in measuring and critically analyzing navigators' flight performance. Psychometrika, 12, 31-42.

Cavalluzzo, L. (1984). Optempo and training effectiveness. Professional Paper 427. Alexandria, VA: Center for Naval Analyses.

Charles, J. P. (1982). Operational problems in instructor operator station design. In G. L. Ricard, T. N. Crosby & E. Y. Lambert (Eds.), Workshop on instructional features and instructor operator station design for training systems. NAVTRAEQUIPCEN IH-341. Orlando, FL: Naval Training Equipment Center.

Charles, J. P. & Johnson, R. M. (1971). Automated training evaluation (ATE). NAVTRADEVCEN 70-C-0132-1. Orlando, FL: Naval Training Device Center.

Chiles, W. D. (1967). Conference proceedings: Assessment of complex operator performance. Discussions and conclusions. Human Factors, 9, 385-392.

Ciavarelli, A. P. (1982). *Methodology* to assess in-flight performance for air-to-air combat training. In Proceedings of the 4th Interservice/Industry Training Equipment Conference: Volume I. Washington, DC: National Security Industrial Association.

Ciavarelli, A. P., Pettigrew, K. W. & Brictson, C. A. (1981). Development of a computer-based air combat maneuvering range debrief system. NAVTRAEQUIPCEN 78-C-0136-1. Orlando, FL: Naval Training Equipment Center.

Ciavarelli, A. P., Williams, A. M. & Pettigrew, K. W. (1981). Performance assessment and appraisal system (PAAS). NAVTRAEQUIPCEN 78-C-0136-1. Orlando,FL: Naval Training Equipment Center.

Connelly, E. M. (1982). Performance measures for aircraft carrier landings as a function of aircraft dynamics. PMA5-1-81. Vienna, VA: Performance Measurement Associates.

Connelly, E. M., Bourne, F. J., Leontal, D. G., Migliacco, 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.

Connelly, E. M., Schuler, A. R., Bourne, T. J. & Knoop, P. A. (1971). Application of adaptive mathematical models to a T-37 pilot performance measurement problem. AFHRL-TR-70-45. Wright Patterson AFB, OH: Air Force Human Resource Laboratory.

Conway, E. J. & Norman, D. A. (1974). Adaptive training: New directions. In 7th NTEC/Industry Conference Proceedings. NAVTRAEQUIPCEN IH-240. Orlando, FL: Naval Training Equipment Center.

Cook, S. W. (Ed.) (1947). Psychological research on radar observer training. Army Air Force Aviation Psychology Program Report No. 12. Washington, DC: Government Printing Office.

Crawford, M. P., Sollenberger, R. T., Ward, L. B., Brown, C. W. & Ghiselli, E. E. (Eds.) (1947). Psychological research on operational training. Army Air Force Aviation Psychology Program Report No. 16. Washington, DC: Government Printing Office.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 92, 281-302.

Cureton, E. E. (1950). Validity, reliability, and baloney. Educational and Psychological Measurement, 10, 94-96.

Danneskiold, R. D. & Johnson, W. (1954). An evaluation of an experimental flight grading method for use in the naval air basic training command. Joint Project Report, The Psychological Corporation and U.S. Naval School of Aviation Medicine. New York: The Psychological Corporation.

Demaree, R. G., Marks, M. R., Smith, W. L. & Snyder, M. T. (1962). Development of qualitative and quantitative personnel requirements information. AMRL-TDR-62-4. Wright-Patterson AFB, OH: Aerospace Medical Research Laboratory.

DeMaio, J., Bell, H. H. & Brunderman, J. (1983). Pilot-oriented performance measurement. In Proceedings of the 5th Interservice/Industry Training Equipment Conference. Washington, D. C.: National Security Industrial Association.

Dickman, J. L. (1982). Automated performance measurement: An overview and assessment. In Proceedings of the 4th Interservice/Industry Training Equipment Conference. Washington, DC: National Security Industrial Association.

Dunnette, M. D. (1963). A note on the criterion. Journal of Applied Psychology, 47, 225-251.

Dunnette, M. D. & Borman, W. C. (1979). Personnel selection and classification systems. Annual Review of Psychology, 30, 477-525.

Ebel, R. (1979). Essentials of educational measurement (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Edwards, J. M., Bloom, R. F., Oates, J. F., Jr., Sipitowski, S., Brainin, P. A., Eckenrode, R. J. & Zeidler, P. C. (1985). An annotated bibliography of the manned systems measurement literature. ARI Research Note 85-18. Fort Benning, GA: Army Research Institute.

Ericksen, S. C. (1947). Objective measures of multi-engine instrument flying skill. In N. E. Miller (Ed.), Psychological Research on Pilot Training. Chapter 8. Army Air Force Aviation Psychology Program Report No. 8. Washington, DC: Government Printing Office.

Ericksen, S. C. (1952). A review of the literature on methods of measuring pilot proficiency. Research Bulletin 52-25. Lackland AFB, TX: Human Resources Research Center.

Farina, A. J. & Wheaton, G. R. (1971). Development of a taxonomy of human performance: The task characteristics approach to performance prediction. Technical Report 74. Washington, DC: American Institutes for Research.

Finley, D. L., Obermayer, R. W., Bertone, C. M., Meister, D. & Muckler, F. A. (1970). Human performance prediction in man-machine systems. Vol. I -- A technical review. NASA CR-1614. Canoga Park, CA: Bunker-Ramo.

Fiske, D. W. (1947). Naval aviation psychology. IV. The central research groups. American Psychologist, 2, 62-72.

Flanagan, J. C. (Ed.) (1948). The aviation psychology program in the Army Air Forces. Army Air Force Aviation Psychology Program Report No. 1. Washington, DC: Government Printing Office.

Fleishman, E. A. (1967). Performance assessment based on an empirically derived task taxonomy. Human Factors, 9, 349-366.

Fleishman, E. A. (1982). Systems for describing human tasks. American Psychologist, 37, 821-834.

Fleishman, E. A. & Ornstein, G. N. (1960). An analysis of pilot flying performance in terms of component abilities. Journal of Applied Psychology, 44, 146-155.

Fleishman, E. A. & Quaintance, M. K. (1984). Taxonomies of human performance: The description of human tasks. Orlando, FL: Academic Press.

Freda, J. S., Hall, E. R. & Ford, L. H. (1982). Relationships among student ability, school performance and fleet supervisor ratings for Navy "A" school graduates. Technical Report 136. Orlando, FL: Training Analysis and Evaluation Group.

Fuller, J. H., Waag, W. L. & Martin, E. L. (1980). Advanced simulator for pilot training: Design of automated performance measurement system. AFHRL-TR-79-57. Brooks AFB, TX: Air Force Human Resources Laboratory.

Gardlin, G. R. & Sitterley, T. E. (1972). Degradation of learned skills - A review and annotated bibliography. Report D180-15080-1. Seattle, WA: Boeing Aerospace.

Ghiselli, E. E. (1956). Dimensional problems of criteria. Journal of Applied Psychology, 40, 1-4.

Glaser, R. & Klaus, D. J. (1962). Proficiency measurement: Assessing human performance. (Chapter 12). In R. M. Gagne (Ed.), Psychological principles in system development. New York: Holt, Rinehart & Winston.

Glaser, R. & Klaus, D. J. (1971). Criterion referenced measurement. In M. D. Merrill (Ed.), Instructional design readings. Englewood Cliffs, NJ: Prentice-Hall.

Gleason, J. G. (1947). Fixed gunnery as an objective measure of flying skill. Chapter 11. In N. E. Miller (Ed.), Psychological Research on Pilot Training. Army Air Force Aviation Psychology Program Report No. 8. Washington, DC: Government Printing Office.

Goldstein, I. L. (1978). The pursuit of validity in the evaluation of training programs. Human Factors, 20, 131-144.

Grodsky, M. A. (1967). The use of full-scale mission simulation for the assessment of complex operator performance. Human Factors, 9, 341-348.

Grunzke, P. M. (1978). Evaluation of the automated adaptive flight training system's air-to-air intercept performance measurement. AFHRL-TR-78-23. Williams AFB, AZ: Air Force Human Resources Laboratory.

Guion, R. M. (1961). Criterion measurement and personnel judgments. Personnel Psychology, 14, 141-149.

Guion, R. M. (1965). Personnel testing. New York: McGraw-Hill

Guion, R. M. (1974). Open a new window: Validities and values in psychological measurement. American Psychologist, 29, 287-296.

Hakel, M. D. (1986). Personnel selection and placement. Annual Review of Psychology, 37, 351-380.

Hawley, J. K., Howard, C. W. & Martellaro, A. J. (1982). Optimizing operator performance on advanced training simulators: Preliminary development of a performance assessment and modeling capability. Technical Report 573. Alexandria, VA: Army Research Institute.

Hayes, K. J. & Pereboom, A. C. (1959). Artifacts in criterion-referenced learning curves. Psychological Review, 66, 23-26.

Hemphill, J. K. & Sechrest, L. B. (1952). A comparison of three criteria of aircrew effectiveness in combat over Korea. Journal of Applied Psychology, 36, 323-327.

Henneman, R. H. (1946). Proficiency measures for fighter pilots at the operational level of training in the Army Air Force. American Psychologist, 1, 293 (Abstract).

Hill, J. W. & Goebel, R. A. (1971). Development of automated GAT-1 performance measures. AFHRL-TR-71-18. Williams AFB, AZ: Air Force Human Resources Laboratory.

Hobson, C. J. & Gibson, E. W. (1983). Policy capturing as an approach to understanding and improving performance appraisal: A review of the literature. Academic Management Review, 8, 640-649.

Hooks, J. T. & Kress, G. (1984). Application of air combat maneuvering instrumentation (ACMI) data for performance measurement. AFHRL-TP-83-36. Brooks AFB, TX: Air Force Human Resources Laboratory.

Horner, W. R., Radinsky, T. L. & Fitzpatrick, R. (1970). The development, test and evaluation of three pilot performance reference scales. AFHRL-TR-70-22. Williams AFB, AZ: Air Force Human Resources Laboratory.

Jenkins, D. J., Jr. (1982). Simulator training effectiveness evaluation. Final Report, TAC Project 79Y-001F. Nellis AFB, NV: Tactical Fighter Weapons Center, Tactical Air Command.

Jenkins, J. G., Ewart, E. S. & Carroll, J. B. (1950). The combat criterion in naval aviation. Report No. 6. Washington, DC: National Research Council Committee on Aviation Psychology.

Jones, M. B. (1959). Simplex theory. Monograph No. 3. Pensacola, FL: Naval School of Aviation Medicine.

Keenan, J. K., Parker, T. C. & Lenzycki, H. P. (1965). Concepts and practices in the assessment of human performance in Air Force systems. AMRL-TR-65-168. Wright-Patterson AFB, OH: Aeromedical Research Laboratory.

Kelley, C. R. (1969). What is adaptive training? Human Factors, 11, 547-566.

Kelley, C. R. & Kelley, E. J. (1970). A manual for adaptive techniques. Santa Monica, CA: Dunlap and Associates.

Kelley, C. R. & Prosin, D. J. (1968). Adaptive performance measurement. Santa Monica, CA: Dunlap and Associates.

Kelley, C. R. & Prosin, D. J. (1969). Adaptive techniques in measuring complex performance. Santa Monica, CA: Dunlap and Associates.

Kelly, M. J., Wooldridge, A. L., Hennessy, R. T., Vreuls, D., Barneby, S. F., Cotton, J. C. & Reed, J. C. (1979). Air combat maneuvering performance measurement. AFHRL-TR-79-3. Williams AFB, AZ: Air Force Human Resources Laboratory.

Kemp, E. H. & Johnson, A. P. (1947). Psychological research on bombardier training. Army Air Force Aviation Psychology Program Report No. 9. Washington, DC: Government Printing Office.

Kennedy, R. S., Jones, M. B. & Lane, N. E. (1986). Overcoming unreliability in operational measures: The use of surrogate measure systems. Paper submitted for presentation at the 30th Annual Meeting of the Human Factors Society.

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

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

Krendel, E. S. & Bloom, J. W. (1963). The natural pilot model for flight proficiency evaluation. NAVTRADEVCEN 323-1. Port Washington, NY: U.S. Naval Training Device Center.

Lane, N. E. (1971). The influence of selected factors on shrinkage and overfit in multiple correlation. NAMI Monograph No. 17. Pensacola, FL: Naval Aerospace Medical Institute.

Lane, N. E. (1975). Operational criteria of pilot performance: Development and evaluation of a postgraduate rating form. Paper presented at Annual Meeting of Aerospace Medical Association, San Francisco, CA.

Lane, N. E. (1986). Skill acquisition curves and military training. Final Report, Institute for Defense Analyses Order No. 5540. Orlando, FL: Essex Corporation.

Lane, N. E. & Waldrop, G. W. (1985). Computer-based instruction (CBI): Considerations for a user-oriented technology data base. EOTR 85-5. Orlando, FL: Essex Corporation.

Lepley, W. M. (Ed.) (1947). Psychological research in the theaters of war. Army Air Force Aviation Psychology Program Report No. 17. Washington, DC: Government Printing Office.

Leuba, H. R. (1964). Quantification in man-machine systems. Human Factors, 6, 555-583.

Lintern, G., Nelson, B. E., Sheppard, D. J., Westra, D. P. & Kennedy, R. S. (1981). Visual Technology Research Simulator (VTRS) human performance research: Phase III. NAVTRAEQUIPCEN 78-C-0060-11. Orlando, FL: Canyon Research Group.

Lintern, G., Sheppard, D. J., McKenna, D., Thomley, K. T., Nolan, M., Wightman, D. C. & Chambers, W. (1985). Simulator design and instructional features for air-to-ground attack: Transfer study. NAVTRASYSCEN 85-C-0044-1. Orlando, FL: Naval Training Systems Center.

Livingston, S. A. (1972). Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 9, 13-26.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Lyon, D. R., Eubanks, J. L., Killion, T. H., Nullmeyer, R. T. & Eddowes, E. E. (1980). Pop-up weapon delivery maneuver: Use of pilot self-assessment data in analysis of critical components. AFHRL-TR-80-33. Brooks AFB, TX: Air Force Human Resources Laboratory.

Martin, E. L. (1984). Practice makes perfect. AFHRL-TR-84-32. Williams AFB, AZ: Air Force Human Resources Laboratory.

Marks, M. R. (1961). Development of human proficiency and performance measures for weapons systems testing. ASD-TR-61-733. Wright-Patterson AFB, OH: Aeronautical System Division.

Matheny, W. G. (1969). The effective time constant: A new technique for adaptive training. Human Factors, 11, 557-560.

McGuinness, J., Bouwman, J. H. & Puig, J. A. (1982). Effectiveness evaluation for air combat training. In Proceedings of the 4th Interservice/Industry Training Equipment Conference: Volume I. Washington, DC: National Security Industrial Association.

McGuinness, J., Forbes, J. M. & Rhoads, J. E. (1984). Air combat maneuvering performance measurement design. AFHRL-TP-83-56. Brooks AFB, TX: Air Force Human Resources Laboratory.

Mengelkoch, R. F., Muckler, F. A. & Monroe, R. D. (1959). Pilot performance measurement equipment for an electronic flight simulator. Report 10977. Baltimore, MD: Martin Company.

Messick, S. (1975). The standard problem: Meaning and value in measurement and evaluation. American Psychologist, 30, 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35, 1012-1027.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. Psychological Bulletin, 89, 575-588.

Miller, N. E. (1947)(a). Psychological research on pilot training. Army Air Force Aviation Psychology Program Report No. 8. Washington, DC: Government Printing Office.

Miller, N. E. (1947)(b). The problem of measuring flying proficiency. In N. E. Miller (Ed.), Psychological research on pilot training (pp. 73-82). Army Air Force Aviation Psychology Program Report No. 8. Washington, DC: Government Printing Office.

Miller, N. E. (1947)(c). Summary of main results and recommendations for future work. Chapter 15. In N. E. Miller (Ed.), Psychological research on pilot training. Army Air Force Aviation Psychology Program Report No. 8. Washington, DC: Government Printing Office.

Mixon, T. R. (1982). An annotated bibliography of objective pilot performance measures, Part 2. NP555-81-010PR. Monterey, CA: Naval Postgraduate School.

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

Moore, S. B., Madison, W. G., Sepp, G. D., Stracener, J. T. & Coward, R. E. (1979). Air combat training: Good stick index validation. AFHRL-TR-79-15. Brooks AFB, TX: Air Force Human Resources Laboratory.

Morgan, B. B., Jr. & Alluisi, E. A. (1972). Synthetic work: A methodology for the assessment of human performance. Perceptual and Motor Skills, 35, 835-845.

Moroney, W. F., Pruitt, V. R. & Lau, C. (1979). Utilization of energy maneuvering data in improving in-flight performance in air combat maneuvering. In Proceedings of the Human Factors Society 23rd Annual Meeting, 503-507.

Muckler, F. A. (1977). Selecting performance measures: "Objective" versus "subjective" measurement. In L. T. Pope and D. Meister (Eds.), Productivity enhancement: Personnel performance assessment in Navy systems. San Diego, CA: Navy Personnel Research and Development Center.

Nagle, B. F. (1953). Criterion development. Personnel Psychology, 6, 271-288.

Naylor, J. C. & Wherry, R. J., Sr. (1965). The use of simulated stimuli and the JAN technique to capture and cluster the policies of raters. Educational and Psychological Measurement, 25, 969-986.

Newell, A. & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), Cognitive skills and their acquisition. Hillsdale, NJ: Erlbaum.

Nunnally, J. C. (1967). Psychometric theory. New York: McGraw-Hill.

Oberle, R. A. (1974). An air combat maneuver conversion model. Report No. CRC 274. Washington, DC: Office of Naval Research.

Oberle, R. A. (1983). Air combat evaluation: The reduced dimension measures. RES 83-6-2. Orlando, FL: Naval Training Equipment Center.

Oberle, R. A. & Naron, S. E. (1978). The air combat maneuvering range readiness estimation system (ACMR RES). Volume I. Project overview. CRC-355-Vol.-1. Arlington, VA: Center for Naval Analyses.

Obermayer, R. W., Vreuls, D., Muckler, F. A., Conway, E. J. & Fitzgerald, J. A. (1974). Combat-ready crew performance measurement system. AFHRL-TR-74-108(I). Brooks AFB, TX: Air Force Human Resources Laboratory.

Parker, J. F., Jr. (1967). The identification of performance dimensions through factor analysis. Human Factors, 9, 367-373.

Pfeiffer, M. G. & Browning, R. F. (1984). Field evaluations of aviation trainers. Technical Report 157. Orlando, FL: Training Analysis and Evaluation Group, Naval Training Equipment Center.

Pitz, G. F. & Sachs, N. J. (1984). Judgment and decision: Theory and application. Annual Review of Psychology, 35, 139-163.

Popham, W. J. & Husek, T. R. (1969). Implication of criterion-referenced measures. Journal of Educational Measurement, 6, 1-9.

Poulton, E. C. (1965). On increasing the sensitivity of measures of performance. Ergonomics, 8, 69-76.

Prophet, W. W. (1972). Performance measurement in helicopter training and operations. Professional Paper 10-72. Alexandria, VA: Human Resources Research Organization.

Pruitt, V. R. & Moroney, W. F. (1980). Energy maneuverability displays for air combat training. SAE Technical Paper Series 801182. Paper presented at the 1980 Aerospace Meeting of the Society of Automotive Engineers.

Rehmann, J. T. (1982). Pilot performance measurement: An annotated bibliography. DOT/FAA/CT-82/24. Atlantic City, NJ: Federal Aviation Administration.

Ricard, G. L., Crosby, T. N. & Lambert, E. Y. (Eds.) (1982). Workshop on instructional features and instructor operator station design for training systems. NAVTRAEQUIPCEN IH-341. Orlando, FL: Naval Training Equipment Center.

Ronan, W. W. & Prien, E. P. (1966). Toward a criterion theory: A review and analysis of research and opinion. Greensboro, NC: The Richardson Foundation.

Ronan, W. W. & Prien, E. P. (Eds.) (1971). Perspectives on the measurement of human performance. New York: Appleton-Century-Crofts.

Rusis, G., Spring, W. G. & Atkinson, J. M. (1971). Future undergraduate pilot training (UPT) system study. Final report, Appendix XIV, Performance measures. Northrop Operational Report 70-149. Wright-Patterson AFB, OH: Aeronautical Systems Division.

Ryack, B. L. & Krendel, E. S. (1963). Experimental study of the natural pilot flight proficiency evaluation model. NAVTRADEVCEN 323-2. Port Washington, NY: Naval Training Device Center.

Schneider, W. A. (1985). Training high performance skills: Fallacies and guidelines. Human Factors, 27, 285-300.

Semple, C. A., Cotton, J. C. & Sullivan, D. J. (1981). Aircrew training devices: Instructional support features. AFHRL-TR-80-58. Wright-Patterson AFB, OH: Air Force Human Resources Laboratory.

Semple, C. A. & Cross, B. K. III. (1982). The real-world and instructional support features in flying training simulators. In G. L. Ricard, T. N. Crosby & E. Y. Lambert (Eds.), Workshop on instructional features and instructor operator station design for training systems. NAVTRAEQUIPCEN IH-341. Orlando, FL: Naval Training Equipment Center.

Simpson, W. R. (1976). Development of a time-variant figure of merit for use in analysis of air combat maneuvering engagements. TM 76-1-SA. Patuxent River, MD: Naval Air Test Center.

Simpson, W. R. & Oberle, R. A. (1977). The numerical analysis of air combat engagements dominated by maneuvering performance. TM 77-2-SA. Patuxent River, MD: Naval Air Test Center.

Singer, N. & Gerson, R. F. (1978). Cognitive processes and learner strategies in the acquisition of motor skills. Report No. TR-78-TH-10. Alexandria, VA: Army Research Institute.

Smith, J. F., Flexman, R. E. & Houston, R. C. (1952). Development of an objective method of recording flight performance. Technical Report 52-15. Lackland AFB, TX: Human Resources Research Center.

Smode, A. F., Gruber, A. & Ely, J. H. (1962). The measurement of advanced flight vehicle crew proficiency in synthetic ground environments. AMRL-TDR-62-2. Wright-Patterson AFB, OH: Aerospace Medical Research Laboratory.

Smode, A. F., Hall, E. R. & Meyer, D. E. (1966). An assessment of research relevant to pilot training. AMRL- TR-66-196. Wright-Patterson AFB, OH: Aerospace Medical Research Laboratory.

Spears, W. D. (1983). Processes of skill performance: A foundation for the design and use of training equipment. NAVTRAEQUIPCEN 78-C-0113-4. Orlando, FL: Naval Training Equipment Center.

Stoffer, G. R. (1981). Performance measurement and the Navy's tactical aircrew training system (TACTS). NAVTRAEQUIPCEN IH-333. Orlando, FL: Naval Training Equipment Center.

Swezey, R. W. (1978). Aspects of criterion-referenced measurement in performance evaluation. Human Factors, 20, 169-178.

Tenopyr, M. L. & Oeltjen, P. D. (1982). Personnel selection and classification. Annual Review of Psychology, 33, 581-618.

Thomas, G. S. (1984). Close air support mission: Development of a unitary measure of pilot performance. AFHRL-TR-84-39. Williams AFB, AZ: Air Force Human Resources Laboratory.

Thorndike, R. L. (1947). Research problems and techniques. Army Air Force Aviation Psychology Program Report No. 3. Washington, D. C.: Government Printing Office.

Thorndike, R. L. (1949). Personnel selection: Test and measurement technique. New York: Wiley.

Thorndike, R. L. (Ed.) (1971). Educational measurement (2nd ed.). Washington, DC: American Council of Education.

Toops, H. A. (1944). The criterion. Educational and Psychological Measurement, 4, 271-297.

Van Cott, H. P. & Altman, J. W. (1956). Procedures for including human engineering factors in the development of weapon systems. WADC-TR-56-488. Wright-Patterson AFB, OH: Wright Air Development Center.

Vreuls, D. & Goldstein, I. (1976). In pursuit of the fateful few: A method for developing human performance measures for training control. In 9th NTEC/Industry Conference Proceedings. NAVTRAEQUIPCEN-IH-246. Orlando, FL: Naval Training Equipment Center.

Vreuls, D. & Obermayer, R. W. (1971)(a). Emerging developments in flight training performance measurement. In Naval Training Device Center 25th Anniversary Commemorative Technical Journal. Orlando, FL: Naval Training Device Center.

Vreuls, D. & Obermayer, R. W. (1971)(b). Study of crew performance measurement for high performance aircraft weapon system training: Air-to-air intercept. NAVTRADEVCEN 70-C-0059-1. Orlando, FL: Naval Training Device Center.

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

Vreuls, D., Obermayer, R. W., Goldstein, I. & Lauber, J. K. (1973). Measurement of trainee performance in a captive rotary-wing device. NAVTRAEQUIPCEN 71-C-0194-1. Orlando, FL: Naval Training Equipment Center.

Vreuls, D., Obermayer, R. W., Wooldridge, A. L. & Kelly, M. J. (1985). Performance measurement guidelines for research. AFOSR-TR-85-0642. Thousand Oaks, CA: Vreuls Research Corporation.

Waag, W. L., Eddowes, E. E., Fuller, J. H. & Fuller, R. R. (1975). ASUPT Automated Objective Performance Measurement System. AFHRL-TR-75-3. Williams AFB, AZ: Air Force Human Resources Laboratory.

Waag, W. L. & Knoop, P. A. (1977). Planning for aircrew performance productivity enhancement R&D: U. S. Air Force. In L. T. Pope and D. Meister (Eds.), Productivity enhancement: Personnel performance assessment in Navy systems. San Diego, CA: Navy Personnel Research and Development Center.

Wallace, S. R. (1965). Criteria for what? American Psychologist, 20, 411-417.

Weiss, D. J. & Davison, M. L. (1981). Test theory and methods. Annual Review of Psychology, 32, 629-658.

Westra, D. P. (1981). Investigation of simulator design features for the carrier landing task: An in-simulator transfer-of-training experiment. NAVTRAEQUIPCEN 81-C00105-1. Orlando, FL: Naval Training Equipment Center.

Westra, D. P., Lintern, G., Sheppard, D. J., Thomley, K. T., Mauk, R., Wightman, D. C. & Chambers, W. (1986). Simulator design and instructional features for carrier landing: Transfer study. NAVTRASYSCEN 85-C-0044-2. Orlando, FL: Naval Training Systems Center.

Wherry, R. J., Jr. & Waters, L. K. (1960). Factor analysis of primary and basic stages of flight training: Advanced jet pipeline students. NSAM-254. Pensacola, FL: Naval School of Aviation Medicine.

Wherry, R. J., Sr. (1957). Past and future of criterion evaluation. Personnel Psychology, 10, 1-5.

Wherry, R. J., Sr. (1984). Contributions to correlational analysis. Orlando, FL: Academic Press.

Wilcoxon, H. C., Johnson, W. & Golan, D. L. (1952). The development and tryout of objective check flights in pre-solo and basic instrument stages of naval air training. Joint Project Report, The Psychological Corporation and U.S. Naval School of Aviation Medicine. Pensacola, FL: Naval School of Aviation Medicine.

Woodrow, H. (1938). The relation between abilities and improvement with practice. Journal of Educational Psychology, 29, 215-230.

Wooldridge, A. L., Breaux, R. & Weinman, D. G. (1976). Statistical issues in the use of multivariate methods for the selection of flight simulator performance measures. NAVTRAEQUIPCEN 75-C-0091-1. Orlando, FL: Naval Training Equipment Center.

Wooldridge, A. L., Kelly, M. J., Obermayer, R. W., Vreuls, D., Nelson, W. H. & Norman, D. A. (1982). Air combat maneuvering performance measurement state space analysis. NAVTRAEQUIPCEN IH-342/AFHRL-TR-82-15. Brooks AFB, TX: Air Force Human Resources Laboratory.

Youngling, E. W., Levine, S. H., Mocharnuk, J. B. & Weston, L. M. (1977). Feasibility study to predict combat effectiveness for selected military roles: Fighter pilot effectiveness. MDC E1634. St. Louis, MO: McDonnell-Douglas Astronautics.

Youtz, R. P. (1947). Objective measures of flying skill for the primary level of training. Chapter 6. In N. E. Miller (Ed.), Psychological research on pilot training. Army Air Force Aviation Psychology Program Report No. 8. Washington, DC: Government Printing Office.

Zavala, A., Locke, E. A., Van Cott, H. P. & Fleishman, E. A. (1965). The analysis of helicopter pilot performance. AIR-E-29-6/65-TR. Washington, DC: American Institutes for Research.

Zedeck, S. & Cascio, W. F. (1984). Psychological issues in personnel decisions. Annual Review of Psychology, 35, 461-518.

DISTRIBUTION LIST

Commanding Officer
Naval Training Systems Center
Code 711
Orlando, FL 32813-7100

Defense Technical Information
Center
Cameron Station
Alexandria, VA 22310

Edward A. Martin
Technical Advisor
ASD/ENETS
Wright-Patterson Air Force Base
Dayton, OH 45433

B. G. Williams
Naval Training Systems Center
Code L02
Pensacola, FL 32508

Jeff Robson
Naval Air Systems Command
Code 5313Y
Washington, DC 20361

Naval Personnel Research and
Development Center
ATTN: Russel M. Vorce, Code 31
San Diego, CA 92152

Air Force Human Resources Laboratory
ATTN: Thomas H. Killion, Ph.D
Williams Air Force Base
Chandler, AZ 85224-5000

AFHRL/FTR
ATTN: Robert S. Kellogg, Ph.D.
Williams Air Force Base
Chandler, AZ 85224-5000

Air Force Human Resources Laboratory
OT Division
ATTN: CDR M. R. Wellick
Williams Air Force Base
Chandler, AZ 85224-5000

Naval Aerospace Medical Institute
Code 00L
ATTN: COL F. S. Pettyjohn
Naval Air Station
Pensacola, FL 32505

Dr. Stan Collyer
Office of Naval Technology
MAT-0722
800 N. Quincy Street
Arlington, VA 22217

Mr. Chuck Gainer
Chief, ARI, Field Unit
ATTN: PERI-SR
Fort Rucker, AL 36362

ARMRL/HEF
ATTN: Dr. Grant McMillan
Wright Patterson Air Force Base
Dayton, OH 45433

Naval Training Systems Center
Code 002
(ATTN: MAJ L. Rohloff)
Orlando, FL 32813-7100

Dr. John Chippendale
PERI-SR
Bldg 501
Fort Rucker, AL 36362

LT David Gleisner, MSC, USNR
Naval Air Systems Command
Code 5313X
Washington, DC 20361

Dr. William E. Dawson
Psychology Department
Haggar Hall
University of Notre Dame
Notre Dame, IN 46556

Mr. Franklin G. Hempel
Office of Naval Research
Code 1141
800 N. Quincy Street
Arlington, VA 22217

Dr. John Casali
Dept of Industrial Engineering and
Operations Research
Virginia Polytechnical Institute
& State University
Blacksburg, VA 24061

LCDR Thomas Crosby, MSC, USN
Naval Air Systems Command
ATTN: Code 933G
Washington DC 20361-3300

Dr. Michael Lentz
Naval Aerospace Medical Research
Laboratory - Bldg 1811
Naval Air Station
Pensacola, FL 32508

CAPT Thomas Gallagher, MSC, USN
Naval Air Development Center
Code 60A
Warminster, PA 18974

LT. C. Barrett
Naval Air Development
Code 6021
Warminster, PA 18974-5000

Dr. George Anderson
Dept. of Psychology
University of Illinois
603 E. Daniel Street
Champaigne IL 61820

CDR Chuck Hutchins, MSC, USN
Naval Post Graduate School
Code 55MP
Monterey, CA 93940

CAPT James Goodson, MSC, USN
Operational Psychology Dept
Naval Aerospace Medical Institute
Code 11
Naval Air Station
Pensacola, FL 32508-5606

CAPT William Moroney, MSC, USN
Naval Air Development Center
Code 602
Warminster, PA 18974-5000

CAPT Michael Curran, MSC, USN
Office of Chief of Naval Operations
Director, Naval Medical (OP-939)
Pentagon - Room 4D461
Washington, DC 20350-2000

LCDR Larry Frank, MSC, USN
Pacific Missile Test Center
Code 4025
Point Mugu, CA 93042-5000

CDR Wade Helm, MSC, USN
Naval Aerospace Medical Research
Laboratory (Code 05)
Naval Air Station
Pensacola FL 32508-5700

CAPT Joseph Funaro, MSC, USN
Naval Training Systems Command
Code 71
Orlando, FL 32813-7100

LT James Hooper, MSC, USNR
Naval Air Systems Command
ATTN: APC205-ON
Washington, DC 20361-1205

LT Lee Goodman, MSC, USN
Naval Air Development Center
Code 6022
Warminster, PA 18974-5000

LCDR Dennis McBride, MSC, USN
Pacific Missile Test Center
Code 4025
Point Mugu, CA 93042-5000

CDR Thomas Jones, MSC, USN
Office of Naval Research
Code 125
800 N. Quincy Street
Arlington, VA 22217-5000

LCDR Tom Singer, MSC, USN
Naval Air Development Center
Code 60B5
Warminster, PA 18974-5000

LCDR Dave Norman, MSC, USN
DTDAC
3280 Progress Drive
Orlando, FL 32826-3229

CAPT Paul Chatelier, MSC, USN
OUSR7E (R&AT)
The Pentagon (Room 3D129)
Washington DC 20301

Commander
Naval Air Force
U.S. Pacific Fleet (J. Bolwerk)
Naval Air Station North Island
San Diego, CA 92135

Commanding Officer
Air Force Office of Scientific
Research
Technical Library
Washington, DC 20301

National Defense University
Research Directorate
Fort McNair, DC 20319

Dr. Jesse Orlanski
Institute for Defense Analyses
Scienceand Technology Division
400 Army-Navy Drive
Arlington, VA 22202

Commanding Officer
405TIW/SEF
Luke Air Force Base AZ 85309

CDR Jerry Owens, MSC, USN
Naval Air Systems Command
ATTN: Code APC205-OM
Washington, DC 20361-1205

American Psychology Association
Psyc. Info, Document Control Unit
1200 Seventeenth Street
Washington DC 20036

LCDR Ed Trautman, MSC, USNR
Psychology Department
Human Factors Laboratory
University of South Dakota
Vermillion, S.D. 57069

Cubic Corporation
9233 Balboa Avenue
Technical Library
San Diego, CA 92123

Technical Library
Naval Training Systems Center
Orlando, FL 32813-7100

Naval Research Laboratory
ATTN: Library
Washington, DC 20375