AD-A099 947   COURSEWARE INC SAN DIEGO CALIF   F/G 5/9

RECOMMENDATIONS FOR THE F-16 PERFORMANCE MEASUREMENT SYSTEM.(U)

MAR 81 R F SCHMIDT, A S GIBBONS, R JACOBS   F02604-79-C-8875

UNCLASSIFIED   NL

1 OF 1
AD
A099947

END
DATE
FILMED
6 81
DTIC

F-16 AIRCREW TRAINING DEVELOPMENT PROJECT

Contract No. F02604-79-C8875

RECOMMENDATIONS FOR THE F-16 PERFORMANCE MEASUREMENT SYSTEM

DEVELOPMENT REPORT No. 14
MARCH 1981

Prepared in fulfillment of CDRL no. B020
and partial fulfillment of CDRL nos. B031 and B050

by

R.F. Schmidt (Courseware, Inc.)
A.S. Gibbons (Courseware, Inc.)
R. Jacobs (Hughes Aircraft Company)
G.W. Faust (Courseware, Inc.)

COURSEWARE, INC.
10075 Carroll Canyon Rd.
San Diego, CA 92131
(714) 578-1700

PREFACE

This report was created for the F-16 Aircrew Training Development Project contract no. F02604-79-C8875 for the Tactical Air Command to comply with the requirements of CDRL nos. B020, B031, and B050. The project entailed the design and development of an instructional system for the F-16 RTU and instructor pilots. During the course of the project, a series of development reports was issued describing processes and products. A list of those reports follows this page. The user is referred to Report No. 34, A Users Guide to the F-16 Training Development Reports, for an overview and explanation of the series, and Report No. 35, F-16 Final Report, for an overview of the Instructional System Development Project.

F-16 AIRCREW TRAINING
DEVELOPMENT PROJECT REPORTS

Copies of these reports may be obtained by writing the Defense Technical Information Center, Cameron Station, Alexandria, Virginia 22314. All reports were reviewed and updated in March 81.

Gibbons, A.S., Rolnick, S.J., Mudrick, D. & Farrow, D.R. Program work plan (F-16 Development Report No. 1). San Diego, Calif.: Courseware, Inc., September 1977, March 1981.

Thompson, A., Bath, W., & Gibbons, A.S., Previous ISD program review (F-16 Development Report No. 2). San Diego, Calif.: Courseware, Inc., September 1977, March 1981.

Wild, M., & Farrow, D.R. Data collection and management forms report (F-16 Development Report No. 3). San Diego, Calif.: Courseware, Inc., September 1977, March 1981.

Gibbons, A.S. Review of existing F-16 task analysis (F-16 Development Report No. 4). San Diego, Calif.: Courseware, Inc., June 1977, March 1981.

Gibbons, A.S., & Rolnick, S.J. Derivation, formatting, and use of criterion-referenced objectives (CROs) and criterion-referenced tests (CRTs) (F-16 Development Report No. 5). San Diego, Calif.: Courseware, Inc., September 1977, March 1981.

Rolnick, S.J., Mudrick, D., Gibbons, A.S. & Clark, J. F-16 task analysis, criterion-referenced objective, and objectives hierarchy report (F-16 Development Report No. 6). San Diego, Calif.: Courseware, Inc., October 1978, March 1981.

Gibbons, A.S. Task analysis methodology report (F-16 Development Report No. 7). San Diego, Calif.: Courseware, Inc., October 1978, March 1981.

Gibbons, A.S. Objectives hierarchy analysis methodology report (F-16 Development Report No. 8). San Diego, Calif.: Courseware, Inc., October 1978, March 1981.

Mudrick, D., Gibbons, A.S., & Schmidt, R.F. Goal analysis report (F-16 Development Report No. 9). San Diego, Calif.: Courseware, Inc., February 1978, March 1981.

Rolnick, S.J., Mudrick, D., & Thompson, E.A. Data base update procedures report (F-16 Development Report No. 10). San Diego, Calif.: Courseware, Inc., October 1978, March 1981.

Mudrick, D., & Pyrz, K.E. Data automation of task and goal analysis: Existing system review and recommendation (F-16 Development Report No. 11). San Diego, Calif.: Courseware, Inc., September 1977, March 1981.

O'Neal, A.F., & Smith, L.H. Management System needs and design concept analysis (F-16 Development Report No. 12). San Diego, Calif.: Courseware, Inc., December 1977, March 1981.

Gibbons, A.S., Thompson, E.A., Schmidt, R.F., & Rolnick, S.J. F-16 pilot and instructor pilot target population study (F-16 Development Report No. 13). San Diego, Calif.: Courseware, Inc., September 1977, March 1981.

Schmidt, R.F., Gibbons, A.S., Jacobs, R. & Faust, G.W. Recommendations for the F-16 performance measurement system (F-16 Development Report No. 14). San Diego, Calif.: Courseware, Inc., October 1978, March 1981.

Thompson, E.A., & Gibbons, A.S. Program/system constraints analysis report (F-16 Development Report No. 15). San Diego, Calif.: Courseware, Inc., October 1978, March 1981.

Gibbons, A.S., & Rolnick, S.J. A study of media production and reproduction options for the F-16 project (F-16 Development Report No. 16). San Diego, Calif.: Courseware, Inc., February 1978, March 1981.

O'Neal, A.F., & Kearsley, G.P. Computer managed instruction for the F-16 training program (F-16 Development Report No. 17). San Diego, Calif.: Courseware, Inc., July 1978, March 1981.

Wilcox, W.C., McNabb, W.J., & Farrow, D.R. F-16 implementation and management plan report (F-16 Development Report No. 18). San Diego, Calif.: Courseware, Inc., October 1978, March 1981.

Sudweeks, R.R., Rolnick, S.J., & Gibbons, A.S. Quality control plans, procedures, and rationale for the F-16 pilot training system (F-16 Development Report No. 19). San Diego, Calif.: Courseware, Inc., October 1978, March 1981.

Gibbons, A.S., Axtell, R.H., & Hughes, J.A. F-16 media selection and utilization plan report (F-16 Development Report No. 20). San Diego, Calif.: Courseware, Inc., October 1978, March 1981.

Thompson, E.A., Kearsley, G.P., Gibbons, A.S., & King, K. F-16 instructional system cost study report (F-16 Development Report No. 21). San Diego, Calif.: Courseware, Inc., October 1978, March 1981.

Jacobs, R.S., & Gibbons, A.S. Recommendations for F-16 operational flight trainer (OFT) design improvements (F-16 Development Report No. 22). San Diego, Calif.: Courseware, Inc., October 1978, March 1981.

Gibbons, A.S. F-16 instructional sequencing plan report (F-16 Development Report No. 23). San Diego, Calif.: Courseware, Inc., October 1978, March 1981.

Farrow, D.R., & King, K. *F-16 coursewares and syllabi delivery schedule* (F-16 Development Report No. 24). San Diego, Calif.: Courseware, Inc., September 1979, March 1981.

Rothstein, L.J., Hibian, J.E., & Mudrick, D. *F-16 instructor/course manager training requirements report* (F-16 Development Report No. 25). San Diego, Calif.: Courseware, Inc., October 1978, March 1981.

O'Neal, A.F., & O'Neal, H.L. *F-16 pilot media selection* (F-16 Development Report No. 26). San Diego, Calif.: Courseware, Inc., March 1979, March 1981.

Gibbons, A.S. *F-16 instructional system design alternatives* (F-16 Development Report No. 27). San Diego, Calif.: Courseware, Inc., September 1979, March 1981.

Gibbons, A.S. *F-16 instructional system basing concept* (F-16 Development Report No. 28). San Diego, Calif.: Courseware, Inc., September 1979, March 1981.

O'Neal, H.L., & Rothstein, L.J. *Task listings and criterion-referenced objectives for the instructor pilot F-16 training program* (F-16 Development Report No. 29). San Diego, Calif.: Courseware, Inc., September 1979, March 1981.

Bergman, D.W., & Farrow, D.R. *F-16 training system media report* (F-16 Development Report No. 30). San Diego, Calif.: Courseware, Inc., September 1979, March 1981.

Gibbons, A.S., O'Neal, A.F., Farrow, D.R., Axtell, R.H., & Hughes, J.A. *F-16 training media mix* (F-16 Development Report No. 31). San Diego, Calif.: Courseware, Inc. October, 1979, March 1981.

Farrow, D.R. *F-16 training media support requirements* (F-16 Development Report No. 32). San Diego, Calif.: Courseware, Inc., September 1979, March 1981.

Gibbons, A.S. *F-16 training media constraints and limitations* (F-16 Development Report No. 33). San Diego, Calif.: Courseware, Inc., September 1979, March 1981.

Farrow, D.R., & Kearsley, G.P. *A user's guide to the F-16 training development reports* (F-16 Development Report No. 34). San Diego, Calif.: Courseware, Inc., January 1981, March 1981.

Farrow, D.R., & Clark, J. *F-16 Final Report* (F-16 Development Report No. 35). San Diego, Calif.: Courseware, Inc., January 1981, March 1981.

EXECUTIVE SUMMARY

The major purpose of this report is to present "state of the art" recommendations for developing the F-16 PMS. Before the recommendations are made, theoretical and practical concerns that a PMS must address are presented. Theoretical issues such as reliability, validity, and the rule linking measurement to grades are seen as fundamental to the measurement process. But these concepts must be implemented within real world constraints.

Since regulations, for example TACR 50-31, determine the structure and content of performance measurement practice, these documents were reviewed and their guidelines evaluated in terms of existing systems within the Air Force. The A-10, F-15, and F-4 systems were reviewed. The direction provided by TACR 50-31 ranges from precise prescriptions to very broad guidelines. This has both good and bad points. On the one hand, the broadness of the guidelines allows flexibility for individual training systems to adapt to local needs. However, the lack of specifics on critical matters like grade interpretation, the remediation process, and the function of gradeslips are seen as ambiguous areas that might lead to confusion.

The final section of the report presents a proposal for the F-16 PMS. Although specific proposals are made on the tools to be used, personnel involved, and record keeping incorporated in the PMS, this summary will present only the highlights of the proposed system. The major innovations are as follows:

1. Use of automated academic tests and quizzes.
2. Concern with higher level evaluation rather than rote memorization (where appropriate).
3. Use of a comprehensive student progress report.
4. Improved gradeslips that will identify student strengths and weaknesses.
5. Procedures for proficiency advancement.
6. Integrated affective indicator behaviors.
7. Measurement of both part and whole tasks.
8. IP instruction in performance measurement.

9. Automated record keeping.

At this time, the final decisions on what the PMS will look like have not been made. The PMS that will be used for the F-16 training program will be described in report number 18, F-16 Implementation and Management Plan Report.

CONTENTS

CONTENTS (cont.)

RECOMMENDATIONS FOR THE F-16 PERFORMANCE MEASUREMENT SYSTEM

1.0 INTRODUCTION

The purpose of this paper is to outline the principles and design for a performance measurement system to be developed for F-16 pilot and instructor pilot training. A complete performance measurement system must embody all of the measurements which (a) identify when a student is ready to advance to a new learning activity or graduate from a curriculum, (b) produce administrative instruments, reports, and records vital to management of the instructional system, and (c) assist in maintaining system quality control. A performance measurement system is directly involved in prescribing measurement instruments and procedures. However, the quality of an instructional system's performance measurement indirectly affects every aspect of system function, including the ability of the system to allocate its resources, perform self-maintenance, and respond to changing output or input requirements.

The process of designing a performance measurement system requires an interaction between theoretical and real-world concerns. Measurement theory should serve as a model or form which the real-world limitations must be mixed with in order to produce a workable measurement system. Within the instructional development approach, criterion-referenced objectives (CROs), along with their standards and conditions, serve as the foundation for instructional measurement and thus constitute the challenge of real world. Measurement theory is employed to produce a system which measures the attainment with the highest possible consistent degree of accuracy and reliability.

A performance measurement system is closely tied to the overall instructional system design. It is dangerous to conceive of the performance measurement system as an entity distinct from the total instructional system. The instructional development process produces a series of closely interdependent products which support each other both in form and function. Both the performance measurement system and the instructional system are founded on the same logically derived documentation base of task listings and objectives. Both derive their structure and content from these documents which are created for that specific purpose and which are carefully maintained and updated to provide that information on a current basis. Changes in that base must affect performance measurement as well as instruction. Operationally the instructional and performance measurement systems complement each other in a cycle which produces feedback for instructional system direction and maintenance. Because performance measurement is an integrated component within an instructional system, there are many factors within the system which are affected and which must be taken into consideration in making the system practical and usable.

In designing the performance measurement system for the F-16, a broad range of questions and considerations were addressed and are reported in this document. Section 2.0 reviews the basic principles of measurement theory upon which measurement systems should be based. In Section 3.0 existing representative TAC performance measurement systems are examined. Each system is analyzed for both strong and weak points, considering both theory and practical limitations. Section 4.0 describes key measurement problems in performance measurement for fighter pilot training. Section 5.0 enumerates recent advances in automated performance measurement in anticipation of ideas proposed for F-16. Section 6.0 combines the information from the preceding sections into an F-16 performance measurement plan. Included in this section is a detailed description of system characteristics and rationale along with ideas regarding system implementation.

2.0 PERFORMANCE MEASUREMENT SYSTEM PRINCIPLES

The general principles prerequisite to all performance measurement systems are enumerated in this section. First, considerations common to all performance measurement systems are named. Second, principles important to measurement within pilot training are discussed, and after that the requirements for performance measurement arising from the application of instructional technology.

2.1 General Performance Measurement Principles

A performance measurement system must distinguish between grades and observation/measurement. Performance measurement within training occurs in two basic steps: (a) a prespecified behavior is observed and recorded and (b) a value is placed on that measurement in the form of a grade. The measurement step is defined as the process of acquiring and recording data concerning the dimensions, capacity, or amount of something regulated by a standard. A grade is an index attached to a measurement to indicate the extent to which a measured value satisfies a criterion. A grade gives meaning or interpretability to a number obtained from measurement.

Past performance measurement seems to have overlooked this separate-but-symbiotic relation between measurement and grades. As a case in point, simulators and complex electronic measuring devices have been made to generate literally thousands of data points for individual skills, but the collective meaning of these measures is often uninterpretable by the user, since no rule is devised for converting measures to grades. The F-16 performance measurement system must employ a combination of measurement and grading rules which have demonstrated training value.

A performance measurement system must have potentially meaningful grades and scores. Depending upon the use of scores, the interpretation which can be placed upon them may vary. Scores which are to be used for the purpose of ranking students and spreading them out along a continuum of performance for comparison purposes need not be descriptive of a student's actual level of performance. They function most effectively when they can be cast along a single dimension of a given scale. On the other hand, scores which are meant to indicate the adequacy of a student's performance with respect to a set criterion must indicate whether or not the criterion was reached. Additionally, they should inform the instructional system when the criterion was not reached and describe the specific deficiencies in the performance. Such scores serve an important diagnostic function. For F-16, an attempt will be made to choose methods of reporting scores and grades that have interpretability and are not opaque to the reader.

A performance measurement system must measure valid behaviors. Validity is a prerequisite of good performance measurement. It is defined as the quality of measuring that which you intend to measure. If a performance measurement system states that it is measuring a student's ability to perform full-range mission-related behaviors and then measures specific behaviors which are not representative, then that system does not make valid measurements.

One of the greatest strengths of the instructional development approach is that its generative processes are designed to provide a closer link between what is intended, what is taught, and what is measured. The mechanism for this lies in correctly performing five steps. First, careful consideration is made of the desired terminal performance expected in training graduates. This includes careful specification of end-of-course expectations for all courses. Second, the process of task analysis compiles an inventory of behaviors involved in the terminal performance. These behaviors, directly derived through a logical process, are themselves potential measurement points for the performance measurement system. Third, in the face of limited resources, performance measurement decision rules are implemented to determine in a systematic way those behaviors which are most job-valid and thus have the highest priority for being measured. This insures that behaviors not measured are carefully selected out and not omitted through chance or through selecting only the easiest behaviors to measure. Fourth, scenarios and problem situations are constructed for use in performance measurement at higher levels. These problem settings help insure that a full range of behaviors is being tested, including specifically the more job-relevant behaviors. Fifth, specific validity checks before and after implementation of the measurement system form a quality control backup to help insure that a full range of valid behaviors is being measured.

A performance measurement system must employ valid measurement points and parameters. In order to evaluate a valid behavior one must choose measurement points and parameters which are indicative of "correct" performance of the behavior. In the choice of measurement points and parameters, there are several approaches which can guide the user toward collecting valid data. One approach may specify steps of tasks and measure performance along prespecified parameters at each step. Another approach may record a single index or a set of indices when the task has been completed. This type of measurement is used most often when a product is being measured or when a process for task execution cannot be unequivocally stated. Milestone measurement schemes are still a third approach. They take measurements of specified variables at set points during task execution. Not all stages and not all values involved in performing the task are measured, but rather selected values at selected points. Finally, a tolerance band approach may be used which requires continuous rather than isolated measurement of the student's ability to keep within certain prespecified limits of performance--a flight path defined as a tunnel in the sky, for instance.

When measurement schemes are constructed, it must be guaranteed that methods of measurement and measurement points and values are chosen which can tell if the task is being performed correctly.

The F-16 instructional system will attempt to insure the validity of its measured behavior by choosing the most appropriate parameters to measure and by attempting to devise sufficient recordkeeping systems to handle the data generated by these procedures. The difficulties which are anticipated from taking this approach are twofold: (a) the criterion problem which is discussed at length later in this report and (b) the critical problems which come from imposing a tremendous information processing burden upon the performance measurement system. In many performance measurement settings the generation of vast amounts of data prevents timely individual task performance analysis for most tasks in a form that is usable by the student and instructor. Solutions to these problems are addressed in Section 6.0.

A performance measurement system must measure reliability between judges. The problem of inter-rater reliability in evaluation of most types of performance is well documented. Well defined CROs can provide the basis for measuring inter-rater reliability because of clearly stated standards of acceptable behavior. The role of the performance measurement system is to measure as accurately and reliably as CROs allow. Subjective judgments of student performance by instructors or raters generate highly unreliable data.

The task of the F-16 raters will be to eliminate such sources of rating error without harming the usefulness of the measures. Sources of rating error include subjective feelings of the judge toward the student being rated, variability of judge temperment from day to day, lack of understanding on the part of the judges as to specific criterion on which to grade students, the requirement to perform rating in an extremely fast-paced environment, difficulty in constructing standard measurement scenarios and problem situations, and others.

One of the major goals of the F-16 performance measurement system will be to aid instructor pilots (IPs) to overcome these problems, not to replace IP judgment, but to augment and support it so that it may become more reliable between raters. These problems can be attacked through several means. Automated measurement coupled with IP interpretation appears to be capable of measuring a limited number of performances reliably. The training of instructors will have some effect on the standardization of the judgments and knowledgeablility concerning measurements to be made. Increased emphasis on the grade as an instructional tool and diagnostic aid also appears to improve judgment reliability and usefulness.

A problem bearing on inter-rater reliabilities to be discussed later in this paper is the threat created by specific grades given by the instructor and the possible subsequent consequences (as in accident investigations). Such issues must be confronted by the performance measurement system in order to create an atmosphere hospitable to reliable measurement.

A performance measurement system must measure reliability within judges. Within-judge reliability is a difficult problem related to inter-rater reliability. Little research has been conducted on the problem, despite the fact that many of the problem factors described in the preceding paragraph also affect within-rater reliability.

It is suspected, but not supported, that less total error occurs in this area. In light of the fact that student pilots are "rotated through" many different IPs, within-judge unreliability is probably less likely to disrupt performance measurement. A partial solution to between- and within-judge reliability problems is to instill in the rater a high degree of interest in making correct calls and a willingness to comply with published standards. It is unthinkable that a judge at the Olympics would undertake to revise the scoring practice by making his own point-giving system. The same mentality must exist among IPs. The standards for such judging must be published and public. To increase within-judge reliability, care must be given by individuals to making correct judgments and recognize established standards.

A performance measurement system must be able to measure student behavior reliably. The design of a performance measurement system must include provisions for measuring student performance frequently enough to insure the behavior is in fact a stable level of achievement rather than a momentary fluctuation. When tests are given only once the danger exists of determining proficiency inaccurately. A student may have a good day, or if the grade is low, a particularly bad day, causing measurements to predict a level of behavior higher or lower than than actually exists.

Safeguards against this problem include scheduling of spaced-interval repeated performance testing. This technique, already implemented at the knowledge level with certain emergency procedures, must be implemented at the performance level within the F-16 measurement system.

2.2 Performance Measurement Systems in Pilot Training

The characteristics that must be present in a performance measurement system related to pilot training are discussed individually below. These characteristics arise mainly out of practical considerations which are related specifically to the environment in which the pilot is trained and in particular, the constraints and pressures which are placed on pilots and instructors in the high performance flying environment.

The performance measurement system must have simplified preflight prepartion. Prior to each flight exercise, IPs must refresh themselves on the activities to be engaged in during the flight. Reviewed items include types of performance measures to be made, the critical values and steps to be observed during measurement, and special contingencies that should be expected during the flight. It is a practical consideration in the design of a performance measurement system that preparation for flight be kept as simple as possible. There are several reasons for this necessity. First, if the preparation task becomes too time-consuming or difficult, it may not be done well. Second, IPs usually have auxiliary duties and are working under heavy work loads. They have a minimum amount of time to devote to flight preparations.

The following measures can be adopted to simplify the preparation for flights: (a) standardization of flight format when practical and the adoption of standard terminologies to represent complex sets of activities, (b) special emphasis on the pre-training of instructors in an intensive instructor training course which stresses the criteria and ability to identify criterion performance, (c) IP familiarization of sequencing based on a hierarchical view of skill learning, and (d) review or refresher programs to assist IPs in upgrading and maintaining their knowledge of judging standards.

A performance measurement system must have a simplified inflight data recording procedure. Very often the speed with which flight related activities are carried out and the distractions which occur in the flight situation make it difficult or impossible for an IP to record all of the data necessary for evaluation or remediation. The section reviewing present performance measurement systems will cover this topic and its potential in detail. (See Section 3.0.)

A pilot training performance measurement system must have simple and useful debrief guides to insure complete coverage. The principle of feedback following an exercise in an aircraft is very important not only from the instructional technology standpoint but also for insuring safety in future flights. Debrief guides will be created for the F-16 performance measurement system which will assist the IP in making a thorough debrief to help insure complete coverage of all relevant topics and issues. This guide will provide subsequent IPs and students with a complete index of progress and needed help reduce time spent in retraining of already-mastered skills.

A pilot training performance measurement system must have a simplified flight recordkeeping procedure. The primary purpose of data produced by inflight and simulator recording would be for use during the debrief. The pressure of time, which can tend to short circuit debriefs and make them less effective, will also have an adverse affect on postflight recordkeeping procedures. The recordkeeping system must therefore be simplified so that the data are both accurate and useful for the IPs and the student. The same data would also provide adequate documentation for any future evaluation/examination of flight performance. The task of designing recordkeeping policy and procedures will require somewhat difficult tradeoffs in view of the fact that somewhat detailed records are recommended by this report in sections dealing with diagnostic feedback to the student.

A pilot training performance measurement system must have the capability of measuring the full range of job behaviors. A primary measurement challenge for pilot training includes the requirement to design a system which can manage behaviors ranging from the very basic and observable to high level, complex and lengthy behavioral sequences. The grading of a preflight aircraft checklist procedure would be considered very simple, whereas grading air combat manuevering against unpredictable targets is much more complex, though possible. Few problems are anticipated regarding methods for making performance measurements in the areas which are more basic and easier to measure. Historically these behaviors have been well-measured. Perhaps the ease of making them has caused them to be over-measured as well. There is, however, a tremendous lack of guidance to the instructional developer in measuring or grading more complicated levels of behavior such as air combat maneuvering. The F-16 performance measurement system will treat each main level of performance measurement complex separately. Management techniques and grading schemes will be selected which are best suited to each level of behavior being measured. Organizational guidelines will also be provided for the IPs and the recordkeeping system to retain conceptual simplicity, an essential for usefulness of grades.

A pilot training performance measurement system must remove antimotivating or threat producing practices from grading and data recording. It is possible that an IP can be held responsible for accidents and damage that occur far beyond his immediate control. Very often the practices of grading and commenting on student performance provide the occasions on which unrealistic retribution for instructor contribution hangs. The IP who makes himself visible by giving comments or grades which are different from the expectations produced by previous grades may be exposing himself to unwarranted consequences. If an unexpected grade is awarded, IPs are usually required to make detailed rationales and justifications. IPs are also often held responsible for assessment comments which are only advisory in nature.

The performance measurement system of F-16 must endeavor to relieve IPs from the threat of retributions by providing the opportunity for comment and by requiring feedback on student performance in a structured fashion which provides sufficient gradeslips will attend to exceptional performance and address specific criterion achievement. Comments will enumerate deviations from performance standards and evaluate in an objective fashion, perhaps through directed questioning, overall levels of performance.

2.3 Performance Measurement System Characteristics Derived From Instructional Development or Technology Principles

Principles of instructional development or instructional technology define what characteristics, good training, and measurement related to that training must possess. Since all measures are integrated in some way into the operation of a total instructional system, the following principles for performance measurement are stated resulting from the tenets of instructional technology.

A performance measurement system should have diagnostic properties. The role of feedback from the performance measurement system to the instructional system is tremendously important. The performance measurement system is an extension of the instructional system and is created to serve that system. Tests are not administered strictly for the purpose of grading and giving a summary score, but for maintaining progress and quality control checks within the instructional system as well. All practice of behaviors during instruction can be viewed as an informal type of performance test. For a test to act both in guiding and evaluating instruction, feedback from tests must (a) point out specific areas of student behavior which are weak and in need of remediation and (b) tie to specific instruction or practice exercises which will achieve that remediation.

A performance measurement system should be criterion-referenced and avoid the tendency toward norm-referencing. The purpose of training is to produce individuals capable of performing a specified set of job behaviors. Without a criterion-referenced testing scheme, the scores and grades produced by the training system cannot be interpreted to indicate whether students have reached job-level skills. The performance measurement system for the F-16 training program must be criterion-referenced, and students' scores and grades must rise out of a comparison between student performance and a standard criterion.

Though normative scores will function at some places within the F-16 training system for increasing motivation, achieving goal analytic aims, and in performing placement services, the portions of the performance measurement system used to assign grades and mark progress will be based on criterion rather than normative comparisons.

A performance measurement system must be able to discriminate between super-criterion performances in a criterion-referenced testing program. It is often desirable to discriminate between testing program. It is often desirable to discriminate between students who have already performed at criterion level. This function is critical for improving performance above the minimum standards and for inducing some motivational force to insure that students will continue to seek higher and higher levels of performance once they have passed minimum criteria. Normative ranking of scores provides incentives for competition within one's own performance, among classmates and with previous classes. Provision for such comparisons within the F-16 measurement system will be made.

A performance measurement system must be able to measure incoming capabilities of students. One of the first tasks of an operational instructional system is to assess the entry behavior of the students. Frequent changes which occur in the instructional system and in systems which contribute students to it are factors making this asessment difficult. When a new weapon system is introduced, it is usual practice to convert highly experienced crew members from other weapon systems to be crew members for the new weapon system. The training of these personnel is not as extensive as that which is necessary for the personnel trained in later phases of that weapon system's existence. This later group usually has only basic training and little experience, and requires more basic presentations and a greater range of training experiences.

The performance measurement system for the F-16 must be able to determine the entry skill level of the student in such a way that his assignment to curricula meets his specific needs to the extent allowed by the system. It can be expected that such a program will save student time and system resources while maintaining consistent levels of performance.

A performance measurement system should provide students and the system with an index of progress toward instructional objectives. The criterion behaviors measured in a performance measurement system are most often not independent, unsequenced behaviors, but consist of carefully constructed sequences building toward increasingly more complex and difficult behaviors. This structured sequence of behaviors is the result of analysis and careful design employed to maximize the impact of the instructional program and minimize the time and resources consumed. The performance measurement system must produce an index of progress which can be made available to the student, the IP, and the system. These indices must be based upon the structured sequence of measured behaviors and act as a motivating and self-management device for the student and a recordkeeping tool for the system.

In addition to monitoring progress and determining the adequacy of student behavior, this feature can be used to project the likely future of the student during training and detect patterns in learning and the use of strategies. Also, progress indices inform the instructional system of its training effectiveness at each checkpoint. When several students fail to complete a phase within an expected period, an analysis of the instruction would be warranted. The next section addresses this quality control procedure in more detail.

A performance measurement system must provide data for use by the instructional system for self-evaluation and the quality control process. Because the performance measurement system exists as a main data generating process in an instructional system, the data generated on student performance may be used to evaluate the effectiveness and efficiency of the system's instruction and its testing. Item analysis, records of progress adequacy, instructor and student evaluations, and level of terminal competency are some of the self-evaluating measures to be utilized. These data can be used for deciding where and how revision must be made to maintain the performance standards.

The procedures conducted by the quality control system must effectively interface among instructional, evaluation and management needs in a coordinated manner. A system data management processor utilizing centralized control is best performed with computer assistance.

A performance measurement system must select performance measures with instructional use in mind. The choice of measurement points and values must be made from the standpoint of the entire instructional system function rather than from consideration of the performance measurement system as a separate entity. The validity of measurement must first be determined by identifying the objective to be accomplished and then determining what information resulting from performance evaluation may be used for instructional purposes. In some cases, such as extremely complex maneuvers, measurement may only be possible on a holistic evaluation basis: a single score of satisfactory or unsatisfactory. However, for other tasks a measurement approach may be selected which allows sufficient diagnostic or part-task information to be gathered. In both these cases the approach was selected which allowed the maximum of instructionally-usable measurement information.

A performance measurement system must measure whole behaviors as well as behavior fragments. When analytic instructional development techniques are used to derive lists of behaviors to be taught and measured the comprehensive behaviors are systematically broken into individual elements of a much smaller scope. Care must be taken that these fragmented behaviors are not the only behaviors which are measured. When fragmented behaviors are measured, certain higher-order integrated skills are left unevaluated. These higher-order skills normally tie together or give coherence and sequence to the fragmented behaviors and are usually among the paramount objectives of the instruction. Although higher-order behaviors are more difficult to measure reliably, it should be noticed that measuring at this level increases validity, as the integrated skill is more relevant to real-world job performance.

A performance measurement system must take goal analysis factors into account. In order that the instructional system design for F-16 consider the full range of behaviors critical to task performance, affective behaviors as well as cognitive and psychomotor skills must be accounted for. The F-16 performance measurement system must therefore incorporate observation and evaluation of the indicator behaviors yielded by the goal analysis for the assessment of such critical but difficult to measure fighter pilot attributes as confidence, judgment, situational awareness, and aggressiveness. The F-16 goal analysis is one of the first attempts to objectively define observable behaviors reflective of affective states, and should provide both predictive and diagnostic information which may add significantly to training effectiveness. The F-16 performance measurement system will accommodate the measurement of indicator behaviors for these goals which are deemed of greatest worth to the student in a systematic way.

3.0 REVIEW OF EXISTING PERFORMANCE MEASUREMENT SYSTEMS

This section summarizes onging and developing performance measurement systems within Tactical Air Command (TAC). To explain the content and structure of existing performance measurement practice it has been necessary to work from the regulation documents which underlie, and in large part, determine them. For this purpose a thorough study of existing performance measurement documentation has been made. Additional data on existing systems have been gathered through interviews conducted with individuals at various levels of TAC in performance measurement responsibilities. An exhaustive review of performance measurement within TAC is not intended here. More time and contact would have been necessary for such a study than was provided in the present project's scope. An attempt has been made to provide a representative rather than exhaustive analysis of performance measurement by sampling specific weapons systems. Moreover, the concern of this report is primarily those practices and organizations which are related to F-16 performance measurement. Each document or organization will be reviewed in three major areas: (a) tools for performance measurement, (b) personnel required, and (c) the recordkeeping system for measurement and management.

This section will present first a discussion of performance measurement as it exists in the current TAC regulations and manuals, then a review of some current TAC measurement systems in operation (F-4, F-15, A-10), and finally the results of interviews on performance measurement from various perspectives.

3.1 Review of Air Force Regulations and Manuals

This section presents a review of current TAC regulations and manuals pertaining to performance measurement practices.

3.1.1 TAC Regulation on Performance Measurement--TACR 50-31

TACR 50-31 provides all training and operational TAC wings with a general outline by which performance measurement is to be conducted. These regulations serve as the conceptual foundation upon which specific measurements must be made, with specificity of direction ranging from precise to very broad, depending upon the specific measurement topic. This is both a good and a bad feature. The greatest strength of the TACR 50-31 is the flexibility allowed individual training systems for innovation and adaptation to local needs. Its weakness is the nondirective format on critical matters. This lack of specifics can be abused when ambiguous regulation is interpreted in several ways, leading to confusion.

3.1.1.1 Tools. Tools for performance measurement and their manner of construction are specified in TACR 50-31 as described below.

3.1.1.1.1 Screening and Admissions. The TACR 50-31 provides significant detail regarding the tools to be used for student screening and admission. Documents and forms to be used are listed, however, no information is provided as to how the screening and admissions procedures are derived, or what principles they are designed to implement. The assumption is made that individual weapons system programs will create the tools for further screening at the wing level. Whether these additional criteria can actually influence the acceptance of an individual is still unspecified.

3.1.1.1.2 Gradeslips. Content of the Air Force form 1363 (gradeslip) is deferred to the wing level without substantive comment. Presently, gradeslips are reviewed by an IP prior to a flight and are insufficient as instructional tools for directing the IP to problem areas of student performance. For example, no diagnostic recommendations concerning performance and its improvement are required. Such information must be gleaned from voluntary IP comments.

3.1.1.1.3 Academic Testing. The TACR 50-31 calls for the use of criterion referenced testing and academic measurement. All intructional systems development (ISD) teams of the same weapons system are to cooperate in preparing the academic tests. However, no guidelines are provided for establishing "major phases" for which grades are to be assigned or how academic measurement is to impact.

3.1.1.1.4 Grades. For inflight and simulator performance measurement, IPs are required to assign grades as soon as possible following flights and are instructed to grade against an absolute scale specified by the CROs. The meaning attached to the seven possible grades remains ambiguous, however, and does not easily correlate with the mechanism of objective-based grading. Grades are awarded to students according to a "sliding scale" of expectations. Rather than measuring a student's performance against a set standard of acceptability, the performance is judged either adequate or inadequate for the student's level of experience. In this way the expectations of the IP are the real criterion, and the meaning of grades can vary widely.

The IP upgrade evaluation recommendations are even less specific and provide no information regarding the meaning of grades.

3.1.1.1.5 Remediation. When substandard performance occurs, the TACR 50-31 suggests several sources of input for identification and remediation: the IP, the student, the instruction itself, and the testing media. No direction is given, however, as to what data should be collected or how to use them in decision-making. Three additional sorties are allowed per syllabus phase for remedial purposes regardless of the length or complexity of the phase. A phase could therefore conceivably entail fewer missions than the remedial rides permitted. This deficiency could be remedied by determining the number of additional sorties by proficiency factors and the complexity of the phase rather than by an arbitrary number.

3.1.1.1.6 Course Critiques. Course critiques by students at mid-point and the end of the course are required. No mechanisms are established for acting upon the data collected, however.

3.1.2 Personnel

Most personnel assignments specific to performance measurement are made at the wing level, so the TACR 50-31 supplies minimal guidelines for making either selection decisions or assigning duties. The qualification required for individuals participating in the performance measurement system is the "best available". No specifics are supplied for determining who is the "best". All performance/achievement is monitored by the flight commander, which enables the IP to spend more time administering precise instruction on specific tasks. The TAD officer is designated for soliciting recommendations for phase improvements, but no guidelines or agencies for evaluation or corrective action are provided.

Student progression is determined by the IP and supervisor. Proficiency advancement has been approved for use when allowed by the syllabus. The concept of proficiency advancement will be discussed later in this report.

3.1.3 Recordkeeping

The TACR 50-31 provides specific information about the records to be maintained, but very little on how they are to be completed. The first three sections of the 50-31 discuss student admissions and recordkeeping. Sections 8 through 11 instruct the use of various required records and forms, their contents, and their function beyond graduation. The reasons for some policies are not clear, for instance, TAC form 180 "Flying Training Summary" must list the students' stengths and weaknesses for the graduate file, but no further suggestions are made regarding the use of these data.

Although the performance measurement system is technically responsible for measurement of deviation in the case of washout procedings, the TACR 50-31 specified that ultimately the IP and the squadron or wing commander perform the qualification checkrides, using procedures similar to normal grading. The strength of this position is that every effort is made to assist the student in meeting the criterion standards without harsh bias. This strength is detracted from by any idiosyncratic meaning of grades. The performance measurement system must document an absolute level of improvement if measurement is to take place at any more than a subjective level, and if subjective level evaluations are accepted, then it must be understood that the ability of the instructional system to track the progress and proficiency of students and graduates is seriously impaired and will involve much approximating and guesswork.

3.2 Aircrew Standard/Evaluation Program--TACR 60-2

The second Air Force Regulation document reviewed is the Standards Maintenance Program (TACR 60-2) utilized for internal quality control. The TACR 60-2 STAN/EVAL program is conducted by a separate noncommital organization of TAC, and consists of both wing and headquarters teams who conduct scheduled and unannounced evaluations. The TAC program, which supports AFR 60-1 policies, encompasses standardization of aircrew operations procedures, evaluation of the ability of aircrews to perform their assigned flying duties, and compliance with established directives related to flying operations. While the F-16 performance measurement system will have no direct impact on the content or format of STAN/EVAL, TACR 60-2 reflects the policies of TAC toward a performance measurement system.

3.2.1 Tools and Recordkeeping

The tools implemented in STAN/EVAL are used for evaluation purposes only, and are therefore not designed to carry any instructional or diagnostic value. The TACR 60-2 calls for an annual check on flight and instrument proficiency for all operational and training personnel. The written exam is derived from a sampling of all the CROs relevant to the specific weapons system created separately by STAN/EVAL. Special emphasis is placed on boldfaced procedures. Criterion performance on written exams is 85%, or 100% for boldfaced procedures.

The flight and instrument checks utilize their own checklist forms which are based directly on the criterion specified in TACR 60-2 CROs. The checklists are completed by the evaluator both during and following the checks. Grades, performance deviation, and comments are all summarized on TAC form 8, which serves as the documentation and record for the evaluation. A pilot is grounded when he receives an unsatisfactory grade for either the flight or instrument checks (using grades Q, Q-, and unsatisfactory/not combat ready). The pilot must successfully complete a reevaluation in order to qualify for flight duties.

3.2.2 Personnel

Personnel assignment in STAN/EVAL is conducted on a "best available" basis, with no further specification for flight or measurement experience. The procedures state only that selection be made on the experience with the specific weapons system to be evaluated and knowledge of STAN/EVAL procedures. The AFR 60-1 states who is responsible for conducting evaluations, documentation and reports associated with STAN/EVAL. The duties of the supervisors are thus well-defined, but the actual measurement capabilities of the evaluators either from prior experience or program training is left undiscussed.

3.3 Review of Existing Performance Measurement Systems

This section briefly reviews the existing performance measurement systems of the A-10, F-15, and F-4. The A-10 and F-15 programs represent recent performance measurement systems, while the F-4 demonstrates a combination of older measurement practices and innovation.

The differences among the three systems reviewed were surprisingly small, with measurement hardware being the greatest source of diversity. As new developments in measurement techniques appear and are approved by TAC, most weapons systems adopt the changes wherever possible. For example, many older aircraft as the F-4 have been retrofitted with gun cameras as instructional utility for such films has become apparent.

3.3.1 Academic Tests and Quizzes

Academic tests are generated primarily by the instructors following the training course CROs. The degree to which the questions are based solely on the objectives varies inversely with the age of the performance measurement system.

Quizzes are not presently an integral part of academic measurement, primarily due to the amount of effort required for administration and grading without data processing automation systems to help. F-4 tests occur only two or three times during the academic course, and thus partially ignore blocks or phases as logical testing intervals. The A-10 system does employ a number of tape/ slide series which provide diagnostic and remedial instruction.

3.3.2 Progress Reports

The progress of students is loosely monitored by their performance in mission advancement, and less so their preparedness in academic instruction.

3.3.3 Gradeslips

The gradeslip is presently the fulcrum of all performance measurement systems, both because it is the only record of flight proficiency, and because of its apparent ease of use. A review of more than 60 different gradeslips from most existing performance measurement systems indicated that they all share certain basic characteristics. Mission elements are listed in a column followed by seven grades: unknown, dangerous, and grades 0-4. Space is reserved for remarks on the front and back sides. A standard description of each grade's meaning is provided at the bottom of the back side. Each mission element is graded separately. In addition, a single cumulative grade is assigned for overall mission performance.

While the grades theoretically correspond to the criteria specified in the CROs, of which all IPs are knowledgeable, IPs all agree that the grade and its behavioral identity vary greatly among tasks and IPs. This review therefore contends that one of the critical problems with existing performance measurement is the gradeslip format. Studies have demonstrated that experienced IPs can very reliably judge and grade student performance over a variety of tasks, with greater reliability for standardized, predictable maneuvers (Reference 6). Nevertheless, experience cannot be bought, nor can IP training realistically provide the extent of experience needed. As has been the case with previous training programs, the F-16 will expect to receive less well-experienced IP candidates as the system develops, compounding the problem. A more effective, explicit gradeslip format would contribute materially to the solution of thrse problems.

Revision of gradeslips, like the revision of most instructional aspects of the systems reviewed, is prompted only by hardware or policy changes. In the case of sortie curtailment, entire missions are removed without adequate consideration for the instructional implications. To illustrate, if the development of cognitive and psychomotor skills is sequential and constructive, removing any link in the development chain should disrupt the learning process. Because the instructional development approach assumes such a hierarchy of learning, the present approach to instructional changes can prove ineffective and inefficient.

3.3.4 Proficiency Advancement

As an instructional tool, proficiency advancement is at present a program aimed at sortie elimination. All three performance measurement systems reviewed use proficiency advancement in this context rather than as a means of extending and expanding the students' experiences, as it is intended to be used. The misuse of proficiency advancement has two serious ramifications. First, with full knowledge that flying is a strong reinforcing agent, present programs follow above-standard performance by eliminating this rewarding event. Thus, students are rewarded with extra flights for not performing well. Second, the benefits derived from equipping the better students with optimal training reflecting their superior capability are removed. Rather, the good students actually fly less, forcing them into a mold of mediocrity created by the less capable pilots. This programmed regression toward the mean increases flight predictability for administrators but deters mission effectiveness and training efficiency. A combination of individualized advancement and lock-step instruction may prove more compatible with the opposing demands of both the system and the student.

3.3.5 Briefing/Debriefing Guides

The briefing and debriefing guides are tools created for the IP and generally reflect the orientation particular to the individual training program. For example, where terminal objectives have been well developed, as in the A-10 performance measurement system, more emphasis is placed in the guides on measurement by exception, or performance within or outside of the specified tolerance bands. However, many tasks are either too complex to reasonably describe, or no data gathering systems exist for applying a quantifiable score.

The weakness of the guides in all three programs is the lack of direction to the IP on instructional sequencing and attention to the hierarchy of skills being developed. This is to some extent an artifact of syllabus structure. So many behaviors are trained in RTU syllabi that little time may be spent in careful development of confidence and competence in one skill. To the extent that sequences of building skills are injected into syllabi, the instructor guides must reflect knowledge of that and give guidance to the IP.

3.3.6 Simulator Measurement

A part of the present question of simulator training effectiveness is taken up in questions regarding simulator performance measurement capabilities. This report recognizes that simulators are a permanent addition to pilot and aircrew training and that their use can be anticipated to increase in the future, very likely at a cost to airborne instruction. Presently, simulators serve very well in replacing aircraft for the practice and evaluation of some procedures, provide excellent preliminary practice and evaluation for other classes of behaviors, and for a third class offer no apparent instructional benefit. For the training and evaluation of complex air-to-air and air-to-surface combat, simulators may never entirely replace inflight exercises. It has not been determined whether that is even desirable from a training viewpoint.

The currency of simulator training within the F-4, F-15, and A-10 programs increases with the younger programs. The F-4 simulator has been found to have a number of training and measurement problems because it does not closely resemble actual flying. Often-times, the simulator behaves in ways very different from the aircraft and is in danger of producing counterproductive training and misleading measurements. The F-15 and A-10 simulators have incorporated more sophisticated technology and training design and are purported to be effective training and measurement tools for such maneuvers as low level flying with radar, emergency avionics, and interface instrumentation.

A distinction should be made between simulator characteristics desirable for training purposes and those desirable for performance measurement. These sets of characteristics are not mutually exclusive, but the need for simulators for training purposes often overshadows the closely related but somewhat different need for simulators for performance measurement purposes.

Design of a simulator performance measurement system implies the selection of a set of measures of interest and values from among massive amounts of data that the simulator is capable of producing. The selection should be guided by an appreciation of the properties of the data available and by the uses to which they are put.

Digital point simulators operate on a quantitative symbolic representation of the physical processes of flight. To completely close the loop of students, aircraft dynamics and environment, data representing aircraft and subsystem states are modified iteratively by modeling effects of the students' behaviors. These data are by the nature of the computation process quantitative rather than qualitative, and objective rather than subjective. A simulator performance measurement system can only monitor, select, process, and report on the subset of these data, thus can only be expected to be quantitative and objective in its operation. In this respect, it differs from the process of performance measurement by immediate evaluation of the IP. As distinguished earlier in this report, the simulator performance measurement system produces scores but cannot assign grades.

The range of variables available to the performance measurement system process in state-of-the-art simulators is quite broad. Everything from pilot behavior to mission performance level is represented quantitatively in some form of the device and may be assumed to be accessible. To capture histories on all these quantities for even a short duration mission would tax the largest of memories for even a mission of moderate length; it would be a questionable value for real-time interpretation. Thus, it seems clear that to be of use to the IP, a smaller but meaningful subset of this data base must be attended to and that a summary of process data rather than raw history, should be preferred. The identification of critical performance parameters for inclusion in this set is not easy. Studies have repeatedly failed to produce sets of quantitative parametric performance on aircraft state variables that can be demonstrated to correlate highly with IP subjective evaluation for any but the most simply described manuevers. For more complex tasks such as air-combat maneuvering, criteria for successful performance have yet to be non-controversially defined.

The problem is also complicated by the consideration of reference thresholds. Uniform standards of performance may not be appropriate as the simulated mission progresses from phase to phase. The use of alternative schedules and standards as a function of mission activities has been implemented with notable success in the Army's SFTS system at Fort Rucker, and in the Coast Guard Helicopter Training Simulator Complex at Mobile, Alabama. But there, missions are restricted to highly structured, rigidly standardized exercises. Such training curricula as may be built into the program fail to individualize remediation, thus we cannot move individual students along at the optimum pace. Instructor interaction to select alternative measure sets imposes an unworthy workload on the IP at a time when his attention should be focused on training mission management and curricular option selection. The IPs' proper function is executive rather then clerical: to interpret scores and perform diagnoses.

The identification of a set of summary measures implies sufficient a priori analysis to insure that scores produced will embody appropriate critical performance attributes upon which to base grades and to provide feedback for identifying weaknesses in the training system itself. Variablility in scores is attributable to several sources, including differences in student aptitudes, student's day to day variations, the quality of training, the degree of learning, and the stability of the performance measures themselves. The measures selected must enable the discrimination of the wheat from the chaff in this melange. Although all of these are of interest for various applications, the degree of learning is the only appropriate predictor of future performance by the students.

Performance measures, to be of use, must be presented properly. Not only should they be formatted to be easily interpretable by the IP, but if possible the computing power of the simulator should be exploited to apply meaningful data transformation, to supply interpretation of scores, and to point the training sequence in the direction of appropriate remediation. The hierarchical structural relationship between performance components and training skill objectives is presumed to be identified as part of the training system design. What is suggested here is that this structure be logically represented as an adjunct to the performance measurement system.

A final point concerning measures is that they should be compatible with growth functions of the simulation device. The properties required in such measures to drive automated training and/or adaptive training logic should be considered in selecting the measure set.

3.3.7 Inflight Performance Measurement Tools

The performance measurement systems being reviewed make use of several measurement tools during inflight exercises in addition to the gradeslips described earlier. Gun camera film is used to record a 20° field of vision of the pilot's cockpit view for measurement of air-to-surface delivery tasks and to provide some information for air-to-air task measurement. While these tapes provide useful measurement data which can be used in a thorough evaluation of performance, there is presently a 1-day delay in film availability. As a result, the film's instructional impact is reduced as both pilot and IP memory for the flight's contextual detail fades. Unfortunately, the IP is often kept from attending the delayed review because of other assignments. The F/15 is presently experimenting with video tape recorder (VTR) systems which provide instant playback capabilities so that videotapes can be used during the debrief session, supplying the same information as gun camera film, plus data from the heads-up display (HUD).

IPs sometimes use portable cassette recorders during complicated missions which enable them to generate auditory notes of measurement events during the mission. This record of comments, judgments, and descriptions is available for debriefing and provides excellent recall cues necessary for effective diagnostics. Memory for exact events and momentary values is believed to develop with good training and experience.

Mission data cards are also used by the IP for notes and comments, sketching engagements, and so on. These provide ready information especially valuable for during-flight reattempts or adjustments, as well as debrief information.

For air-to-surface missions, judgments made by the IP during flight are sometimes supplemented by range scores generated independently of human contamination. These scores are matched with, and added to, the IP evaluation and are included separately on the gradeslips. Range scores are sometimes available during flight via radio communication as well as following the mission in the form of written score reports.

IPs report frustration at the overwhelming amount of measurement data they wish to record for future examination but are unable to--even with recorders, gun cameras, handwritten notes, and range scores. The greatest weakness of measurement aids, however, (except videotape) is that as they become more informative, they become less available during critical analysis periods.

3.3.8 Personnel

Personnel placement and assigned duties in the systems studied conform to the TAC regulation documents described above. While individual responsibilities in each system are well delineated, a problem exists in that little or no training is given to instructors or the Director of Operations (DO) in regard to performance measurement. For example, academic courses generally employ multiple choice tests despite the fact that the instructor receives no training on how to write good multiple choice items. The training staff are more often asked to perform their jobs through ingenuity and past experience than by methods sponsored actively through policy disseminated through careful training. The F-16 performance measurement system will provide IPs with training regarding the effective use of measurement tools, especially in those areas where objective-based measurement techniques have been refined over past practice.

3.3.9 Recordkeeping

Records for the three ongoing systems reviewed vary little in that they mirror the requirements of TACR 50-31. Recordkeeping is conducted manually, which restricts the amount of data which can be handled. Records presently contain too little or improper information to detect with any precision the nature of instructional difficulties encountered by students. Records seldom function as the source of quality control analysis, partially due to ineffective organization and content of the records, the lack of time allotted to improving instruction, or the lack of a periodic scanning and corrections procedure. In essence, once an instructional (and performance measurement) system is up and going, it follows the path of least resistance as far as change and update go.

3.4 Performance Measurement Systems in Training--Interviews

The following review consists of comments and suggestions on general performance measurement made by individuals ranging from students to supervisors during structured interviews. Every effort has been made to solicit a representative cross section of opinions. However, individual interviews cannot be construed as reflective of all personnel of a given group.

3.4.1 Performance Measurement in STAN/EVAL

The emphasis of performance measurement in STAN/EVAL is to insure that aircrews are capable of performing within the guidelines of the published objectives of the TACR 60-2. These objectives are developed and revised by TAC Headquarters and its wing affiliates and serve only as evaluative indices. The objectives contain the conditions and standards associated with each task. A standard type of gradeslip is used to record performance. "Team" performance is also critically observed.

The opinion was expressed that more sophisticated inflight recording devices such as the air combat maneuvering range/instrumentation (ACMR/I) could provide excellent measurement information. The key factor in their use in training would be the efficiency of playback capabilities. The data collected by the system could be evaluated through a combination of automated scoring and human judgment. In this collaborative process, sections of the flight could be played back for review and the system could supply critical data unavailable to the IP during the mission or from his memory.

A problem expressed was that of the changing expectations of the examiner. Although in concept STAN/EVAL employs the same criteria for every evaluation, the view was expressed that in reality different levels of performance capability are expected from pilots and enter into grading decisions. Less experienced pilots tend to be examined "by the book" on safety and basic procedures. For more experienced pilots expectations go beyond this minimum level, and a better performance is required for a good grade.

In order to prevent added bias from entering an evaluation, whether for training or STAN/EVAL, it was suggested that IPs and evaluators be trained at a site different from the site of their assignment in order to remove personal relationships which might influence judgment. The need was also expressed for more frequent evaluation to prevent conclusions about pilot capabilities from being made on a very restricted sample of observed behaviors.

3.4.2. Performance Measurement at the Supervision Level

Measurement tools were mentioned in supervision interviews as being subject to changing expectancies in that grades do not reflect judgment by an absolute standard, but rather a proficiency displayed relative to that expected, given an experience level in the evaluated person. The opinion was expressed that IPs need to possess both a thorough knowledge of the criteria and have available as much independent data regarding student performance as can be readily interpreted. Systems such as ACMR/I were mentioned as an ideal tool for supplying supplementary information to the IP and SP because of its accuracy and objectivity. The use of permanent recording would also make judgments relative to specific criteria more precise in the case of complex tasks.

Because gradeslips are the focal point of simulator and inflight performance, special care should be taken to design them to facilitate ease, accuracy, and quickness of completion. It was also expressed that weaknesses in performance are better indicated by lack of improvement over time than by single observations of individual performances. It was suggested that the performance measurement system therefore track student progress and provide the IP with a history of student learning and not just information about one previous ride.

3.4.3. Performance Measurement at the Student Pilot Level

Perhaps the most useful interview feedback came from students who had nearly completed a training course. Students were generally concerned with fair and valid measurement of their capabilities and thought that grades should reflect actual performance, whether good or bad. Because academics was recognized as a familiarization of tasks to be performed in the cockpit, academic tests were found to be useful as motivators or progress milestones. Students further stated that important content is encountered in everyday routines. Academics was therefore seen as complementary to learning by doing, which is considered invaluable.

When questioned about the meaning of proficiency advancement, the response that it was "a means to eliminate scheduled missions from an individual's syllabus." Eliminating rides under the current proficiency advancement practice has served as a punisher and is avoided by the students.

Automated measurement systems as VTR and ACMR/I were considered by students to be extremely useful measurement devices. It was suggested that if such recording systems are not always available, SPs could either alternate among properly fitted aircraft or go TDY to ranges which have the capabilities to afford everyone the opportunity.

4.0 THE FUNDAMENTAL PERFORMANCE MEASUREMENT CRITERION PROBLEMS

Developments in instructional technology related to performance measurement suggest that performance measurement systems may be capable of addressing criterion problems which up until now have existed within measurement practices. The purpose of this section is to review the problem briefly. In some cases, F-16 attempts at solutions are advanced in general terms. Such solutions are explained in more detail in Chapter 6, which describes the proposed F-16 performance measurement system.

Ever since the inception of criterion-referencing in training, measurement systems have suffered from a lack of well-delineated criteria. The criterion problem which exists in performance measurement systems is that of defining the criterion against which performance will be measured. The problem involves the need to reach an agreement upon which specific variables will be measured (Section 2.1) at various stages of training. The problem is multidimensional and does not have a single simple answer.

For the F-16 training system, the criterion problem will exist at two levels. It will exist at the level of the individual task to be graded, and it will exist at the level of the terminal goals for the F-16 training courses.

4.1 The Individual Task Criterion Problem

The problem of setting criteria for individual tasks consists of finding an acceptable set of measures which will determine whether a performance has been accomplished satisfactorily. The criterion problem at this level must be recognized as an artifact of the evolution of the content itself and probably a permanent problem in performance measurement. Instructional content cannot be thought of as an absolute entity but rather an evolving and developing complex of information. The exact detail of the profile of a maneuver, for instance, is not born fully defined in the mind of pilots, complete with exacting standards of performance. The contrary is true. A given maneuver, which was often discovered by accident, becomes more and more widely used as time progresses. Then standards begin to form about it. Argument over these evolving standards precipitates the criterion problem.

The properties of the F-16 performance measurement system call for criterion-referencing and diagnostic feedback for the student, therefore, the criterion problem cannot be avoided. Selection of measurement points will be closely related to the topographical description of the procedure being performed. That is, measurement points will probably be related to steps when procedures are being measured, and the student's deviation from some standard at each of those points will be the main question. For nonprocedural measurements a similar approach will be used to the extent possible.

For the higher-level, more complex pilot and instructor pilot tasks, a similar problem occurs in generating mastery criteria. In the measurement of complex air-to-air combat behaviors, for instance, problem scenarios are constructed; after the opening set-up, no two follow the same script. Setting criteria for judging these encounters satisfactory or unsatisfactory may be accomplished in different ways. A "kill" or "advantage" score may be used analogous to the scoring of wrestling. Such scores are not diagnostic, however, of the causes of unsatisfactory performance. More elaborate scoring schemes may be evolved to gather such data, but the cost in manpower and instrumentation for such capabilities is very high, and agreement on standards of measurement will be difficult to come by.

In some such situations it may be necessary to resort to the time-honored custom of relying on subjective judgments to a greater extent. If this is done, to avoid the evils which accompany such practices, it must be acknowledged by all that subjectivity is being introduced and a broader range of judges must be called upon to evaluate performances for each person. The judges selected must be widely acknowledged as expert and impartial, and the negative results of adverse judgment must be applied more slowly and with greater deliberation.

In addition to the problems described above, both measurement and recordkeeping within the performance measurement system must reflect criterion attainment on an absolute, as well as a sliding scale. This will consist of multiple levels of achievement for one task, with attainment of each level measured against a well-defined criterion. Student progress must reflect fewer errors and progressively greater within-tolerance performance for the level of performance being mastered. This need is initiated by the requirement for graduating, in a minimum of time, students who have had basic experience in a variety of job tasks. The shortness of the time constrains the instructional system to graduate students (after measurement of criterion performance) who are not at fully-experienced pilot performance levels but who have met a preliminary, acceptable minimum standard of performance, with the expectation that future training (and measurement) will further the students' skills.

The F-16 solution for dealing with this problem will be to specify the minimum acceptable performance standards for each task to be mastered as well as progressively more stringent standards for more advanced measurement. Because diagnostic records will be kept on the growth of student skills for each task through each level of standard, descriptive data for each student's expected learning curve will be available. These data will be invaluable both for the management of instruction during the students' stay at the RTU and for those receiving the student after RTU training who will be responsible for his further training and skills maintenance.

Following this discussion of the criterion problem it may appear that the need for criterion measures imposed by the demands of instructional technology is rendered almost impossible by practical barriers (requirements for high-volume data processing, requirements for high-resolution observation tools, requirements for more standard IP measurement, bookkeepping loads related to the statement of multiple standards for each task, etc.) and the possibility of inundating the training system in a sea of data. It is true that there are many problems in the specification and use of criteria for measurement. The difficulty of the measure is almost always directly proportional to its importance and closeness to real-world behaviors. It nevertheless remains that the measurement capability is indispensible to an instructional system which intends to economize training resources and yet produce the best possible trained pilot. If students are to be advanced through training systematically and as they are ready, and if the instructional system is to be capable of monitoring its effect and changing to increase its own effectiveness, then the detailed measures are necessary and the criterion problem must be given a practical solution.

4.2 Mastery Model Criterion Problem

The mastery model criterion problem arises as measurement of individual task performance is left behind and larger, more lengthy complexes of simple tasks joined together must be measured. Mastery can be defined only to the extent that the specificity of criteria will allow, and for the more complex behaviors the question regarding when a jet pilot has "mastered" a particular skill or behavior can be answered with varying degrees of certainty. The problem of higher level criteria has two parts: statement of F-16 pilot career mastery models, and statement of F-16 RTU pilot mastery models. The F-16 RTU squadrons will train students to fly the F-16 to a given criterion or mastery (the RTU mastery model). At the conclusion of the course, students will be graduated into continuation training which will continue to improve pilot skills and combat-readiness until the designated terminal mastery model has been reached (the career mastery model).

The mastery model criterion problem centers in the fact that the level of attainment at which students will be graduated from the RTU squadron (or later said to have reached a career goal) cannot easily be thought of in terms of simple mastery or nonmastery, but is more appropriately a collection of mastery levels. Such a mastery model may require mastery of some skills, familiarity with other skills, and nonfamiliarity with still other skills. If mastery models are expressed in this way, a "profile" of desired student graduating behavior is possible which can itself be used as a criterion. This deals with the problem conceptually, but in the real world things are not so simply handled.

There is the problem, for instance, of continually-changing RTU resources and expectations. Among the major determinants of the amount a student can master while he is in a training squadron are (1) the resources applied to training (e.g., sorties, aircraft availability, simulator availability, instructor availability), (2) the amount of time a student spends in training, and (3) the incoming capabilities of students as they enter the training system. These variables will probably change during the lifetime of the F-16 instructional systems, as there are changes in budget, mission, or administration. As these variables change, the exact charactertistics of the pilot graduated from the training squadron must be expected to vary also unless there is a related change in the time and resources allocated to RTU training.

Using the "profile" notion introduced above as a means of describing student performance capability at a given point in time, it can be seen that as resources and time allocated to training change, the change in student output can be expressed with a change in the graduated student profile. This offers a solution to the RTU mastery model problem which differs from the present unrealistic "reduce-the-resource-consumption-without-changing-the-quality-of-the-graduate" philosophy. It also recognizes the fact that graduation from RTU represents only a conceptual, and not a real, plateau of performance capability being reached by students. Graduation from the RTU is really only a point on a continuum which represents no real or meaningful level of skills attainment but rather a switch from intensive to less intensive training on the same family of skills.

If the F-16 performance measurement system is to successfully interface measurement of mastery with the changeability of system environment, the user of the system must conceptualize a criterion referenced system as described above with increasingly demanding performance standards, any of which can be used to describe student performance levels at a give time. The F-16 performance measurement system will propose "profiles" where it is necessary to have break points in a student's training career, but these break points will be adjustable to meet the demands of the system's external environment and do not represent any absolute level of preparation.

5.0 TECHNICAL ADVANCES IN AUTOMATED PERFORMANCE MEASUREMENT

This section examines recent technical advances in performance measurement of interest to F-16 training. Performance measurement becomes increasingly difficult as the complexity of the task increases. In high complexity, high-speed air-to-air and air-to-surface performance measurement the amount of information to be registered and recorded by the measurement agent quickly expands beyond the level of human capabilities, and the number of individual task executions to be observed and graded also exceeds the limits of the human information processor. Even with the aid of memory extenders like note pads and audio recorders, the rapid, heavy flow of information to the measurer, most of which is germane to an objective performance measurement, is too much to be handled humanly without the loss of vast amounts of very important data.

Automation of performance measures is a way of extending the range of human memory and observation and judgment capabilities so that an unbiased, accurate measurement may be made of a given performance. Automated performance measurement devices aid their human users by doing one of the following:

1. Gathering and recording data that would normally escape human observers because of the rapid flow of incoming information.

2. Observing and recording values which are not observable to the human observer.

3. Recording trends and patterns in performances and/or noting patterns of deviations from set tolerance limits.

4. Summarizing and correlating vast amounts of data into more readily usable indices and comparison values.

Probably the greatest need for automated performance measures is in those areas where both automation and measurement are the hardest to accomplish: air-to-air and air-to-surface combat. Yet in these areas especially, and in most areas in general, automated measurement is a practice and a technology very much in its infancy. Preliminary studies suggest that automatically-collected data are effective measurement tools, and when properly implemented, assist students in mastering objectives (Reference 14, Reference 20), but the limited range of presently measurable behaviors and the difficulties many measurement systems encounter in producing readily and easily usable readouts has thwarted widespread use of them.

A publication of TAC Headquarters (Reference 18) recently stated:

> "Performance assessment techniques in tactical aircrew training programs have not incorporated recent technology

improvements in performance measurement. In addition, present TAC aircrew performance measurement systems do not provide adequate discrimination between skill levels (both cognitive and psychomotor) of tactical aircrew members."

Previous sections of this report have treated the manner in which the F-16 measurement system will attempt to discriminate between skill levels. The possibilities for F-16 use of automated, technology-based performance measurement aids are discussed below.

For the purpose of this discussion, automated performance measurement systems are divided into simulator and aircraft types. Aircraft-related systems are further broken down into within- and without-aircraft systems. Each of these areas is discussed separately.

5.1 Simulator Related Automated Performance Measurement Systems

The technology of performance measurement in simulators is an area in which there has been a great deal of research (Caro, 1973, Reference 6). The results of this research are highly favorable to further exploration of automated performance measurement systems as an alternative for some tasks, and as a supplement to other tasks for instructor- or judgment-driven measurement. The present state of the art is somewhat behind this optimistic projection. However, automated performance measurement systems for simulators allow for the measurement of a modest range of maneuvers with varying degrees of utility. Recent programs to study air combat automated measurement possibilities will extend the range greatly and no doubt challenge the problems of making automated measurement systems more usable by the worker in the field. The practice of performance measurement is still to a great extent the domain of scientists and researchers. There is yet to develop a general understanding of automation approaches, their use, and their integration with day-to-day training activities or a set of guidelines as to where and when automated measurement is most appropriate or effects the most efficiencies. Saying this is not meant to imply that there is nothing useful at present in the area of automated measures, but it does mean that each application at present is a major research undertaking and often entails exploration of unknown and untried methodologies.

Computer supported simulator systems are presently capable of recording and analyzing output created by any manipulated or indicator instrument in the cockpit and all pertinent environmental influences imposed by the system. General measurement models usually consider six determinants: (1) a maneuver segment, (2) a parameter, (3) a sampling rate, (4) a desired value if required, (5) a tolerance value if required, and (6) a tranformation. (See Reference 17, page 25.) Measures are defined as the end result of the measure production process, which starts with a raw data parameter and ends with a specific transformation of that parameter. Simulators are therefore able to measure observable behavior at almost any level of detail and it is left to the instructional system what is to be measured and what meaning the measure has (References 18, 19).

The issue which remains a controversy is that of transfer of training and fidelity. Micheli (Reference 8) pointed out that the greatest difficulty in simulator use has been the resistance of trainers to adopt simulators in instruction. The report further stated that "exact simulation" is not necessary for positive transfer. Reports cited earlier imply that negative transfer occurs when simulators behave in a manner other than the real aircraft. In those cases, simulation should either be deleted or improved. Fidelity, as stated earlier, is a training issue. It must be judged in that context and requires a great deal more study than it has received.

Prior to adopting any automated measurement system for F-16, several issues must be considered. First, the system must be capable of growth during the lifetime of the F-16 weapons system so that it will be possible to add new measurement devices as the technology of measurement itself advances. The F-16 will be in use for several years, and it is certain new devices and methodologies will emerge during that time. Second, the system must avoid interference with simulator operations stemming from operation of the performance measurement system. Performance measurement systems which are integrated with simulator systems in a unitary fashion often suffer when changes must be made in either system, or when operational difficulties are encountered. Thus, the interaction betwen simulator system and performance measurement system must be intimate but cancellable at any time. Third, all automated measurement systems must be justified by increased effectiveness and efficiency in performance measurement. A composite gain score generated by increased validity, reliability, cost- and time-effectiveness, and usefulness must justify the adoption of any measures adopted.

The adoption of automated performance measures for use in the F-16 instructional system beyond those of a research nature is recommended if the above criteria can be met. Prior to decisions to automate, it is recommended that a study be made of the anticipated costs and benefits resulting from adoption. It is recommended that increased command support be given to F-16 related research on automated measures and air-combat related measures as well.

5.2 Aircraft Automated Performance Measurement Systems

Aircraft related performance measurement systems are of critical importance to the training system because the most job-relevant behaviors and behavior settings occur in the aircraft. As cited above, performance measurement may be thought of as two activities, observing the performance, and measuring or scoring. An automated performance measurement system for use in aircraft measurement should ideally facilitate the observation portion of the performance measurement activity by producing a record of the details of the performance to be measured which could be viewed and reviewed by both student and instructor. It should also facilitate the grading and scoring process by allowing stop-action and slow analysis for greater deliberation and attention to detail. Presently both of these are impossible, and observations made are reconstructed as well as possible from the (sometimes imperfect, sometimes incomplete) memories of the participants. The resulting scoring must be assumed to contain a generous amount of subjectivity.

Automated performance measurement systems for aircraft may be subdivided into within-aircraft measurement systems and extra-aircraft measurement systems. Within-aircraft performance measurement systems include filming and VTR from the aircraft, usually through a cockpit perspective or through a gun camera. One extra-aircraft measurement system used for fighter pilot training is the ACMR/I for recording critical data on the relationship of the aircraft to each other. Each of these systems is discussed separately below.

5.2.1 Airborne Video Tape Recording

The second significant inflight recording device recently introduced is a VTR which provides gun camera-vantage information. The VTR can also record all data projected onto the HUD, a recently developed instrumentation summary device.

A summary of TAC project 76C-071F, Electronic Gunsight Sensor Evaluation (Reference 23) provides an excellent overview of the VTR's utility. Air-to-air and air-to-surface missions were used to determine the VTR system effectiveness for recording and documenting training and combat missions. The project report stated:

> "No malfunctions occurred during the flight test, and the aircrews stated the system was outstanding in all aspects. The ability to receive the electronic gunsight sensor picture on the rear cockpit digital scan coverter provided an excellent inflight instruction aid. The immediate postflight availability of the heads up display/gunsight and "real world" on video tape with full aircraft audio enhanced flight debriefing and aircrew training."
>
> "The sensor adequately satisfied the operational requirements to record and document training missions, and missile firings confirmed that it would adequately suffice as a combat

documentation system for recording air-to-air kills within the sensor field of view."

VTR overcomes the problem of delayed feedback and contributes precise review information which will directly contribute to increasing scoring objectivity. The evaluation and diagnostic information from this system could easily provide a somewhat complete reconstruction of any airborne exercise.

The VTR as a low-cost system has been judged very useful, but certain of its limitations should be noted. First, it is a bulky system and takes up cockpit room, which in the F-16 is limited. Second, the VTR is a sensitive device which if used in the rugged environment of the aircraft may be subject to reliability problems. Third, the VTR has a field of view which for air-to-surface recording may be considered sufficient, but which for air-to-air recording must be judged extremely limited. The area of observation is limited by the narrow lens angle of the VTR to a small corridor of air straight ahead of the aircraft. Consequently only those events which occur within that corridor can be recorded. In air-to-air engagements this means that outcome scoring is possible, but event recording and relative aircraft positioning information is not.

5.2.2 Air Combat Maneuvering Range/Instrumentation

An ACMR/I system is capable of recording a vast amount of data about air-to-air combat engagements. Specially designed instrumentation is mounted on the plane to simulate weapons and send measurement impulses to ground-based receiving stations. Two ground subsystems convert the transmitted data into a suitable form for display, and serve as a control center and data display station. Two important training economies are realized by the ACMR/I. It is a powerful tool for increasing combat effectiveness and readiness. In addition, it makes possible training expense savings. Sources of substantial savings are training missile expenditures, reduced target services, and reduced ACM-associated training accidents.

The ACMR/I system is capable of collecting a variety of performance indicators from each of the following classes: pilot performance data, aircraft performance data, aircraft teaching or tracking data, pilot training data (ground-transmitted instruction), aircraft position data, on-line display data, and off-line display data. The system can simultaneously generate the above data for eight aircraft in a high performance environment simultaneously, and specify such composite indices as time spent in offensive and defensive positions, kill attempts and success, and time and occurrence of dangerous and highly advantageous positions. Complete playback capabilities are also available for debrief.

The F-16 performance measurement system views the ACMR/I system as a technological breakthrough which can serve optimally in both instructional and evaluation functions. Instruction benefits from the completely objective and informative range scores for assisting and supplementing IP scoring and grading. The immediate feedback provided upon attaining or missing a kill gives the student an outstanding opportunity to acquire appropriate behaviors and correct errors. The multiple recording capabilities expand engagement exercises to realistic proportions without loss of data or confusion during reconstruction of the sortie events. From an evaluation standpoint, the objective scores generated by the system can assist and add to the process of scoring and grading student performance.

The ACMR/I system can be employed for recording of air-to-surface maneuvers with equal effectiveness, and may eventually serve to supplement IP data on even more basic maneuvers such as landing, pattern, and basic aerobatics. However, whether the use of such sophisticated measurement tools adds appreciably to the accuracy or effectiveness of instruction of more basic tasks is yet to be demonstrated. This point nevertheless reiterates the fundamental principle that the performance measurement system exists to serve the instructional system, not vice versa.

6.0 PROPOSAL FOR THE F-16 PERFORMANCE MEASUREMENT SYSTEM

This section contains a description of the measurement system proposed for use within the F-16 instructional system. The proposed F-16 performance measurement system will consist of a structure employing many of the concepts and principles discussed in previous sections of this paper. Recommendations are based on those principles and the realization that the performance measurement system must be practical and effective. Alternatives will be presented in those areas where the ideal approach is not likely to be feasible.

Three major headings divide the description: (1) tools used in performance measurement, (2) personnel involved in measurement tasks, and (3) recordkeeping. With each section, attention will be dedicated to considerations of academic, simulator, and inflight instruction.

6.1 Tools Used in Performance Measurement

The measurement tools to be proposed for the F-16 performance measurement system combine automation and human judgment in making precise, quick diagnostic evaluation decisions. Each tool is described separately, citing (1) how it is derived, (2) its measurement properties, and (3) how it is revised.

6.1.1 Automated Academic Tests and Quizzes

Academic tests will be created by the instructional systems development (ISD) team and will directly represent the objectives contained in the objectives hierarchies. The tests will consist of a mix of multiple choice, fill-in, listing, and identification formats. Careful review during test construction will insure that test item formats are appropriate to each objective type and are valid measures of objective behaviors. Each resulting test and alternate form will be used as a standard test at all F-16 training sites. Individual instructors will be encouraged to develop and use additional questions in the specific exams to a predetermined limit, perhaps 10-20% of test length. Instructor-generated items which prove over time to be functional will be considered by the operational training development (OTD) team for inclusion in the standard test to replace weak items and increase test face validity. The quality of standard test items will be evaluated by item analyses with input from all training sites into a single central facility at OTD team offices, hopefully, computer managed instruction (CMI) supported to accomplish the large amount of data processing which will be necessary because of this and several other demands. Items created and employed by the instructors will be examined separately until they are officially included as standard test items.

Academic tests will occur at the end of each instructional block and will contain questions specific to that block. In addition, either two or three comprehensive exams will be given to assure stability of retention and encourage recall of previously tested material. For the comprehensive tests, care will be taken to impose a higher-level evaluation environment to disallow pure memorization except in those areas where memorization is desired, such as for emergency procedures. Overevaluation will be avoided and will occur at levels tolerable to students. This is especially important when it is remembered that much feedback data on system operation will also be gathered from students. With automated academic evaluation such as a computer-based testing system, testing could simultaneously serve two functions. Given an instructional unit, the performance measurement system could introduce quizzes at the end of small segments of instruction and automatically provide corrective feedback, review weak points, and conduct a second qualifying test. The automatic presentation of diagnostics would make this system an extremely potent instructional tool. An automated system could also record the students' progress and performance on quizzes and generate evaluative data separate from block-end tests. A system of this type is recommended.

Such a system would be critical for alleviating the three problems from which academic instruction generally suffers: (1) unavailability of the instructor or trained personnel, (2) lack of time or support to improve instruction, and (3) lack of congruence between academics and flying. The task load of IPs and instructors is generally heavy, resulting in less frequent evaluation. It is proposed that a computer-linked opscan scoring device be used for scoring all multiple-choice type exams. Using an automated system, the instructor would not be required to spend time with such repetitive adminstrative tasks as test scoring. In addition, if testing could be conducted on-line, this scoring procedure could directly transfer the student's data to the test item bank for analysis, yielding reports on student performance in individual content areas, the effectiveness of specific test items, and comparative progress among class individuals, actual versus predicted progress projections, and class to class comparisons. Moreover, unlimited versions of each test could be created automatically by the computer, making it so that every test would be unique. This eliminates the expense and work of creating parallel test forms.

With data records centralized for use at all F-16 training sites, academic instruction could be fully coordinated, combining quality control and progress data for immediate update and change. Effective new items created by the instructors could be quickly added to the standard tests, and ineffective items could be quickly dropped. It is important to note that the proposed computer-based academic evaluation accounts for only a small part of instructional system functions to be served by automation.

6.1.2 Academic Tests and Quizzes--Nonautomated

A non-automated system of academic performance measurement would not differ from an automated system in the method of derivation or revision of tests and items but would be restricted in the production of diagnostic feedback to the student, the ease of analysis, the ease of test form generation and revision, and the production of reports. Immediate student questions would have to be deferred unless an instructor were present at testing. Frequent evaluations over small segments of instruction would be extremely difficult due to manpower requirements. The net effect would be fewer evaluations, less feedback to students and instructional system, and increased IP time spent administering tests rather than instructing. Given a non-automated system, some limited range of functions could be attained. Diagnostics following testing could be attempted using test feedback sheets or instructor directions at great cost of maintenance. Item analyses could be conducted by hand by learning center support personnel. Data generated from classes and across tests and items could be summarized, but not to as great an extent and only after some delay.

6.1.3 Progress Report

A management tool to be incorporated into the F-16 performance measurement system is a concise, comprehensive student progress document. The nature of this document will be unique because of the variety of functions it will serve. The progress document will act as an achievement file for the student, a reference for the IP as to the instruction needed by the student, a summary statement of student progress for the supervisor, and the foundation document for system assessment, future scheduling, and possibly subsequent student assignment. The document will create an achievement profile which will briefly describe the point the student has reached in proficiency in each task. The task-by-task description of student achievement will show in detail the effects of increase or decrease in commitment of resources to training. It would, for instance, enable system managers to see the precise ramifications of sortie cuts and syllabus changes. It will also enable continuation training to interface exactly with the "B" course, picking up the student's continuation training exactly where the RTU training left off. Proficiency in flying missions in the F-16 will be far too complex to be defined by a single descriptive index, and the profile of student capabilities in the progress report will deal directly with that problem by avoiding a single index. This does not mean that a band of competencies cannot be established across the profile to define that level of competence required of the student upon graduation from the RTU, but it does avoid the evils of a noninformation-bearing single index.

The progress report will also be a ready reference for the IP prior to flying a mission. In addition to seeing previous flight gradeslips and flight recommendations, the progress report will quickly tell the IP the strengths, weaknesses, and learning pace of the student. This information will make sorties more effective and allow the instructor to place special emphasis weaknesses. This achievement profile will also be necessary for adjusting the proficiency advancement mission, a concept explained later.

The progress document will much alleviate the problems encountered by the DO in tracking all the students under his direction. The task-specific format will provide detailed achievement information in a format which is easy to inspect and evaluate.

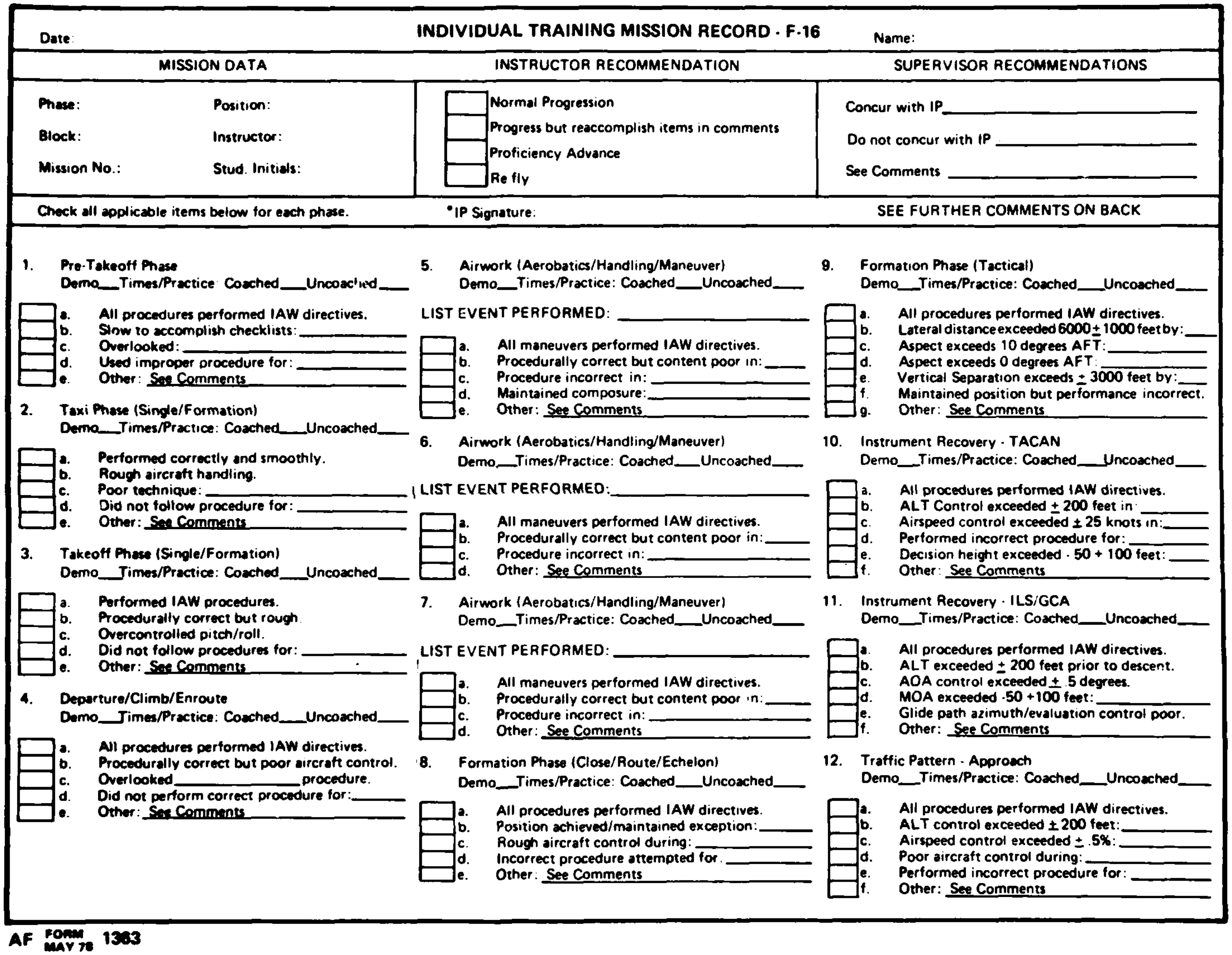
The quality control of the instructional system will depend in large part on the match-up between predicted and achieved progress. When an individual and/or group shows unusual progress, whether slow or fast, adjustments in the training will be made to rectify the deficiency. Group data will also be used to establish realistic expectations, especially during syllabus changes. Scheduling of all training support and maintenance can benefit from the progress document by previewing predicted demands based on past trends and present rate of progress. Finally, when a student demonstrates unusual proficiency at task acquisition in a criterion-referenced evaluation program, the progress document supplies indirect normative data which can be used to distinguish super-criterion performance and assist in making assignments following RTU training.

6.1.4 Gradeslips

The proposed gradeslip for the F-16 performance measurement system is intended to be more informative and easier to complete than those of previous systems.

6.1.4.1 Format. Sample gradeslips can be found on the following pages. (See Figures 1 and 2.) Each gradeslip will contain all the tasks to be completed on a given mission as specified by the syllabus (shown in Figure 1 in some cases as blanks). The tasks will be divided into two types: (1) common operations and (2) specific mission tasks. Common operations are those tasks which must be performed on a given mission but are not the focus of training or evaluation. For example, after the transition phase, tasks such as preparation, STTO, enroute, and pattern are standard tasks necessary for accomplishing the training objectives of any mission. Common operation tasks would be graded simply as qualified or unqualified (with dangerous and unknown under "other"). A grade of unqualified would indicate regressive performance and would require comment.

The specific mission tasks, on the other hand, are the focus of the mission, and need more meaningful and informative grading than is available using standard gradeslips. For every task, each critical segment would be graded on a descriptive scale provided directly on the gradeslip. For example, a landing could be evaluated by the "angle of approach", "touch down zone", and "vertical velocity". The gradeslip would contain the following descriptors: (1) "All procedures performed IAW directives," (2) "AOA control



Date:

INDIVIDUAL TRAINING MISSION RECORD - F-16

Name:

MISSION DATA		INSTRUCTOR RECOMMENDATION	SUPERVISOR RECOMMENDATIONS
Phase:	Position:	□ Normal Progression	Concur with IP ____________
Block:	Instructor:	□ Progress but reaccomplish items in comments	Do not concur with IP ____________
Mission No.:	Stud. Initials:	□ Proficiency Advance	See Comments ____________
		□ Re fly	

Check all applicable items below for each phase.

*IP Signature:

SEE FURTHER COMMENTS ON BACK

1. Pre-Takeoff Phase
Demo___Times/Practice Coached____Uncoached____

- □ a. All procedures performed IAW directives.
- □ b. Slow to accomplish checklists: __________
- □ c. Overlooked: __________
- □ d. Used improper procedure for: __________
- □ e. Other: See Comments

2. Taxi Phase (Single/Formation)
Demo___Times/Practice: Coached____Uncoached____

- □ a. Performed correctly and smoothly.
- □ b. Rough aircraft handling.
- □ c. Poor technique: __________
- □ d. Did not follow procedure for: __________
- □ e. Other: See Comments

3. Takeoff Phase (Single/Formation)
Demo___Times/Practice: Coached____Uncoached____

- □ a. Performed IAW procedures.
- □ b. Procedurally correct but rough.
- □ c. Overcontrolled pitch/roll.
- □ d. Did not follow procedures for: __________
- □ e. Other: See Comments

4. Departure/Climb/Enroute
Demo___Times/Practice: Coached____Uncoached____

- □ a. All procedures performed IAW directives.
- □ b. Procedurally correct but poor aircraft control.
- □ c. Overlooked__________procedure.
- □ d. Did not perform correct procedure for: __________
- □ e. Other: See Comments

5. Airwork (Aerobatics/Handling/Maneuver)
Demo___Times/Practice: Coached____Uncoached____

LIST EVENT PERFORMED: __________

- □ a. All maneuvers performed IAW directives.
- □ b. Procedurally correct but content poor in: __________
- □ c. Procedure incorrect in: __________
- □ d. Maintained composure: __________
- □ e. Other: See Comments

6. Airwork (Aerobatics/Handling/Maneuver)
Demo___Times/Practice: Coached____Uncoached____

LIST EVENT PERFORMED: __________

- □ a. All maneuvers performed IAW directives.
- □ b. Procedurally correct but content poor in: __________
- □ c. Procedure incorrect in: __________
- □ d. Other: See Comments

7. Airwork (Aerobatics/Handling/Maneuver)
Demo___Times/Practice: Coached____Uncoached____

LIST EVENT PERFORMED: __________

- □ a. All maneuvers performed IAW directives.
- □ b. Procedurally correct but content poor in: __________
- □ c. Procedure incorrect in: __________
- □ d. Other: See Comments

8. Formation Phase (Close/Route/Echelon)
Demo___Times/Practice: Coached____Uncoached____

- □ a. All procedures performed IAW directives.
- □ b. Position achieved/maintained exception: __________
- □ c. Rough aircraft control during: __________
- □ d. Incorrect procedure attempted for. __________
- □ e. Other: See Comments

9. Formation Phase (Tactical)
Demo___Times/Practice: Coached____Uncoached____

- □ a. All procedures performed IAW directives.
- □ b. Lateral distance exceeded 6000 ± 1000 feet by: __________
- □ c. Aspect exceeds 10 degrees AFT: __________
- □ d. Aspect exceeds 0 degrees AFT: __________
- □ e. Vertical Separation exceeds ± 3000 feet by: __________
- □ f. Maintained position but performance incorrect.
- □ g. Other: See Comments

10. Instrument Recovery - TACAN
Demo___Times/Practice: Coached____Uncoached____

- □ a. All procedures performed IAW directives.
- □ b. ALT Control exceeded ± 200 feet in: __________
- □ c. Airspeed control exceeded ± 25 knots in: __________
- □ d. Performed incorrect procedure for: __________
- □ e. Decision height exceeded - 50 + 100 feet: __________
- □ f. Other: See Comments

11. Instrument Recovery - ILS/GCA
Demo___Times/Practice: Coached____Uncoached____

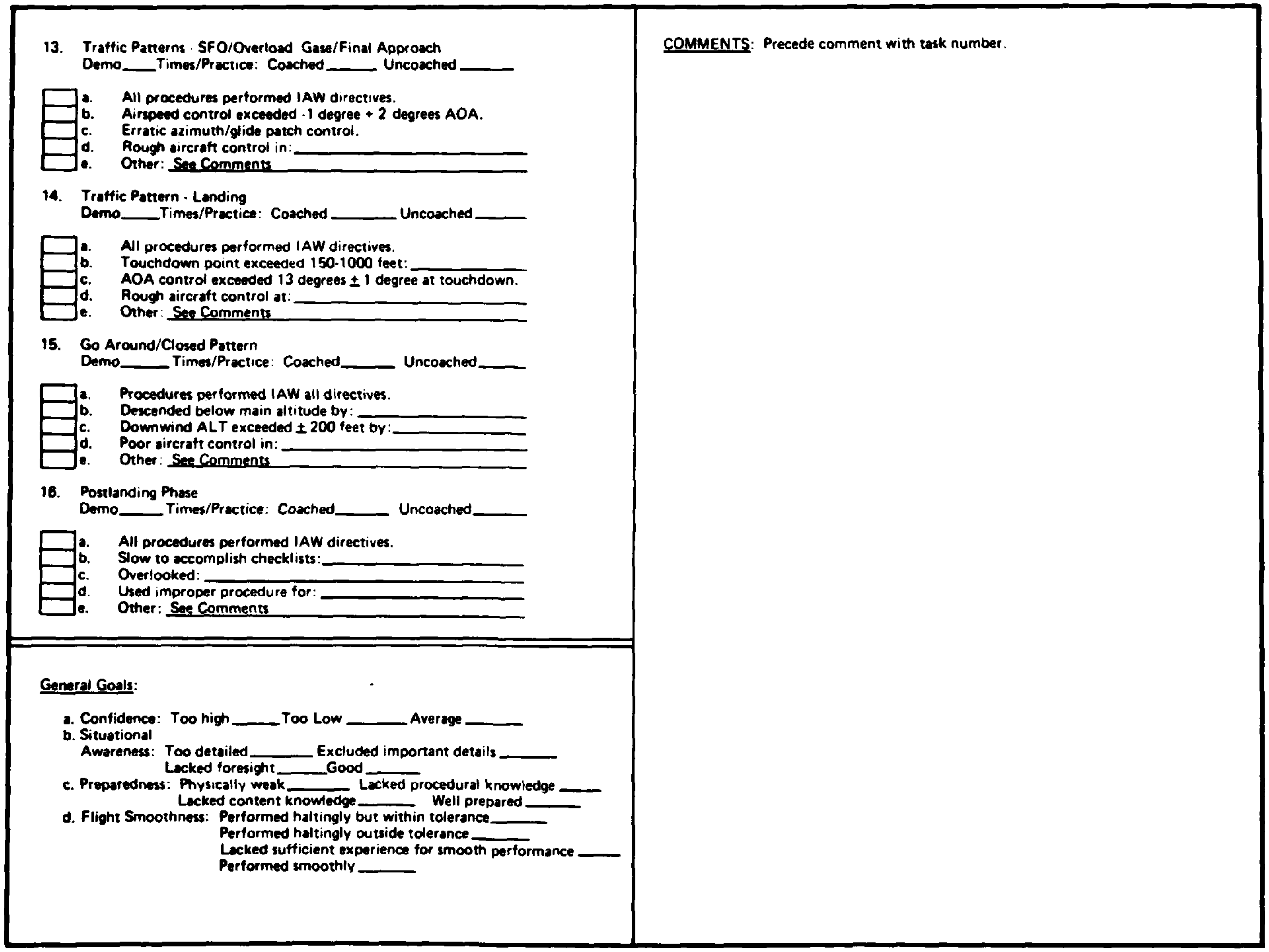
- □ a. All procedures performed IAW directives.
- □ b. ALT exceeded ± 200 feet prior to descent.
- □ c. AOA control exceeded ± .5 degrees.
- □ d. MOA exceeded -50 +100 feet: __________
- □ e. Glide path azimuth/evaluation control poor.
- □ f. Other: See Comments

12. Traffic Pattern - Approach
Demo___Times/Practice: Coached____Uncoached____

- □ a. All procedures performed IAW directives.
- □ b. ALT control exceeded ± 200 feet: __________
- □ c. Airspeed control exceeded ± .5%: __________
- □ d. Poor aircraft control during: __________
- □ e. Performed incorrect procedure for: __________
- □ f. Other: See Comments

AF FORM MAY 78 1363

Figure 1--Sample gradeslip (front).



13. Traffic Patterns · SFO/Overload Gase/Final Approach
Demo____ Times/Practice: Coached ______ Uncoached ______

□ a. All procedures performed IAW directives.
□ b. Airspeed control exceeded -1 degree + 2 degrees AOA.
□ c. Erratic azimuth/glide patch control.
□ d. Rough aircraft control in: ____________________
□ e. Other: See Comments ____________________

14. Traffic Pattern · Landing
Demo_____ Times/Practice: Coached ________ Uncoached ______

□ a. All procedures performed IAW directives.
□ b. Touchdown point exceeded 150-1000 feet: ____________
□ c. AOA control exceeded 13 degrees ± 1 degree at touchdown.
□ d. Rough aircraft control at: ____________________
□ e. Other: See Comments ____________________

15. Go Around/Closed Pattern
Demo______ Times/Practice: Coached______ Uncoached______

□ a. Procedures performed IAW all directives.
□ b. Descended below main altitude by: ______________
□ c. Downwind ALT exceeded ± 200 feet by: ____________
□ d. Poor aircraft control in: ____________________
□ e. Other: See Comments ____________________

16. Postlanding Phase
Demo______ Times/Practice: Coached______ Uncoached______

□ a. All procedures performed IAW directives.
□ b. Slow to accomplish checklists: ______________
□ c. Overlooked: ____________________
□ d. Used improper procedure for: ______________
□ e. Other: See Comments ____________________

General Goals:

a. Confidence: Too high ______ Too Low ________ Average ________
b. Situational
Awareness: Too detailed ________ Excluded important details ________
Lacked foresight ______ Good ______
c. Preparedness: Physically weak ________ Lacked procedural knowledge _____
Lacked content knowledge _______ Well prepared _______
d. Flight Smoothness: Performed haltingly but within tolerance _______
Performed haltingly outside tolerance ________
Lacked sufficient experience for smooth performance _____
Performed smoothly _______

COMMENTS: Precede comment with task number.

Figure 2--Sample gradeslip (reverse side).

exceeded 13 degrees +/- 1 degree of touchdown," (3) "TD point exceeded 150-100 feet," (4) "VVI exceeded +/- (value) degrees," and (5) "rough aircraft control at (point) ".

These descriptors include the standards listed in the CRO for landing, and will constitute the list of the errors most frequently made by the student. More than one descriptor can be checked when applicable. The last item is always left open for other possible occurrences which deviate from "normal" performance of the element, such as dangerous or exceptional execution. (Additional space on the back of the gradeslip will be provided for making further comments and recommendations.) Also included on the gradeslip will be range scores or relevant part-task or simulator scores to be considered for the final recommendation. In this system, the items prompt the IPs to attend to the criteria, reducing the memory load, and giving a guide document for the debrief, without restricting the range of IP response. The "grade" is also diagnostic in that it states specifically the error which occurred. Both the student and the IP for the next mission can benefit from a performance record of this type. Specific attention can be centered on areas of performance where the student is weak. The items are based on within-tolerance aspects of the critical behaviors rather than subjectively comparative behaviors, and are thus more meaningful. Finally, this method eliminates a "response set" where the student gets all threes or just twos and threes. Each item within the specific task must be mastered, and if the student performs outside the criteria at any point, he is below mastery.

Below the heading of each mission task are three other headings: "demo", "practice-coached", and "practice-uncoached". In the case where a task is only being demonstrated, no assignment of a grade is logical, or will be required. Coaching refers to the degree to which the student performs the task independent of external prompts. Coached practice indicates that less than ideal preformance is expected and the student would anticipate receiving some cues or prompting from the instructors.

The concept of a single overall grade has been modified in the F-16 performance measurement system to better suit the ends of instruction. The overall grade will come in the form of an IP recommendation and is listed as follows: "refly the mission", "normal progression", "progress but reaccomplish items in comments", or "proficiency advance". These choices eliminate ambiguity from the grading system, while retaining an overall evaluation or grading quality.

The proposed gradeslip is expected to improve the ability of the instructional system to detect and deal with instructional needs. The performance measurement system must develop guidelines throughout the initial stages of measure for use in defining the rules for post-flight recommendations. Specific recommendations will be suggested when the number of deviations (diagnostic checks) falls within a given range. Standard guidelines will be introduced and utilized.

The performance measurement system also recognizes that every behavior occurs on a continuum and is partially compounded by preceding and succeeding behaviors. The judgment of the IP, based on observation and automated data recording, will remain the final decision source. Extensive measurement and judgment training in the IP training course will help eliminate most avoidable errors and align IP techniques where appropriate.

Gradeslips will contain all the information relevant to the recordkeeping system, such as SP, IP, aircraft names, date, and mission number. The task descriptors will consume the majority of work involved in creating the slips, but will be drawn directly from the CROs, and should therefore not overly burden the system development team. In most cases a given sortie is flown several times with different instructional intent (demo, practice). The proposed design would account for these differences and thus reduce the number of overprints required. The use of an advanced instructional system (AIS) type system would even further simplify the grading procedure, as is discussed below.

Perhaps the most important barrier overcome by this gradeslip format is the discrepancy between criterion-referenced grading and the sliding criteria discussed in earlier sections of this report which causes each student to receive a "2" grade, regardless of absolute performance level. It is felt this change will produce much information for use in progressing students more efficiently and providing more relevant instructional events.

Revision of gradeslips will occur in correspondence with changes in either the syllabus, objectives, or conditions and standards of a given task. The recordkeeping section of the performance measurement system will also keep track of frequently cited student errors not included in the alternatives, and incorporate them into the printed list when justified.

6.1.4.2 Gradeslip--Proficiency Advancement. Although proficiency advancement is not a separate data collecting tool in the performance measurement system, its significance warrants special mention as a subsection of the gradeslip description. The gradeslip recommendation provides for the IP to proficiency advance the student following above-standard performance on the tasks in a mission. As described in Section 3.0, this technique has been abused in the past by using it to eliminate rides, thus reducing the experience and motivation of the student. Although proficiency advancement usually means enabling the student to progress through the program at his own pace maintaining a constant level of challenge, restrictions of management and maintenance prevent this.

As an alternative, the F-16 performance measurement system proposes that the syllabus contain separate instructional units called "alpha rides" which can be attained or earned at specific points in the sequence. For example, if a block consists of three rides, one demo and two practices, the second practice could be replaced by an alpha ride if the student demonstrates mastery on the first practice. The alpha ride would contain the same type of tasks specific to its block, but the complexity and level of tasks would rise according to the abilities of the student.

Such a system would conform to management plans, remove anti-motivational factors implicit in the old system, and generate superior fighter pilots. Proficiency advancement can have an added positive effect on instruction if also used on a limited task-by-task basis. The F-16 syllabus will be created so as to allow for the less capable students to complete each phase without extra rides, thus reducing the overall number of sorties required per student. Therefore, once the less capable student has mastered a given specific mission task, more time and emphasis can be placed on the troublesome tasks in the time remaining. The better students, on the other hand, are accumulating expertise previously unattainable simply by using available time optimally rather than forcing repetition and boredom upon the differential distribution of student abilities. The result is a minimum of standard performance for all students, and above standard performance for those who can attain it, without the use of extra rides. Only extreme problem cases would require extra rides for task-by-task mastery following a phase or block.

6.1.5 Goal Analysis

As alluded to in the gradeslip description, goal analytic measurement will constitute one tool used in the training program. The goal analysis of F-16 training is especially interesting in that it represents the first attempt at integrating affective indicator behaviors in a systematic way into an instructional system. In that little is known about what skills and attributes make a "good" fighter pilot (deLeon, 1977), the validity of the behaviors to be measured is somewhat tentative. The development of behaviors was completed by observing and analyzing the prototypic behavior of experienced and successful fighter pilots. Because the behaviors are derived by concensus, they are also confounded to a certain degree by Air Force tradition. Nevertheless, as a first attempt to collect valid, concrete behaviors aimed at improving pilot training, the F-16 program can take advantage of an excellent development opportunity. On going validity research and training effectiveness studies regarding the measurement and evaluation of goal behaviors can be conducted. These data can then be reapplied to the system and disseminated for present and future training programs. The value of these efforts will ultimately depend upon the ability of the performance measurement system to record the occurrence of indicative behaviors in a systematic and reliable fashion.

For the performance measurement system to collect valid data, F-16 does not envision measurement of goal behaviors outside of the three major instructional arenas: academic, simulator, and inflight. It is very unlikely that information gathered "after hours" will be either valid (there are too many environmental constraints) or available (there exist no inexpensive, practical mechanisms within the system's routine for making the measurements). The F-16 performance measurement system will therefore depend primarily upon self-reports gathered during scheduled syllabus activities and IP comments. As an example, the academic lecture or the segment covering physical inflight demands may require the student to indicate what program he is using to maintain fitness. As a second example, gradeslips, as cited above, would include questions regarding the student's facility at identifying bogies or potential advantage maneuvers, indicating situational awareness and foresight. The weight placed in the gradeslip behaviors will be based on the importance of the skill regarding effectiveness and danger factors. In summary, the F-16 goal analysis is one of the first attempts to objectively define observable behavior reflective of affective states, yet should provide both predictive and diagnostic information which may add significantly to training effectiveness.

6.1.6 Brief/Debrief Guide

The brief/debrief guide will be broken down by missions. Each mission will be described for intent and the standards and conditions of each specific task will be supplied. Below each task, the evaluation item from the gradeslip pertaining to that task will appear. Each item alternative will be described as representing a specified deviation from the established standard.

Recommendations will be included in the guides for facilitating the measurement process (i.e., distinction between the diagnostic and evaluation value of a grade, points of measurement reference, etc.). Attention will also be focused on particular behaviors relevant to the goal analysis which might be observed during a mission. Suggestions will be made for detecting the presence and quality of these behaviors.

6.1.7 Simulator Tools

The extent to which the simulator will be capable of making the proposed measurements depends upon the recording and analysis characteristics engineered into the system. Recommendations from the F-16 project for OFT/WST performance measurement features are contained in project report no. 22, "Recommendations for F-16 Operational Flight Trainer (OFT) Design Improvements". Emphasis should be placed on measures which have direct diagnostic value based on achieving criterion performance of the course objectives. Because the complexity of performance measurement increases exponentially with the increased complexity of the environment, the capabilities of the simulator as a recording and evaluation tool will be reviewed separately.

6.1.7.1 Simulator: Part and Whole Task Measurement. The development of a skill usually occurs in stages involving subtasks, which become integrated into larger units eventually form a single, cohesive skill. Thus, the F-16 simulators and trainers will hopefully be used to measure both part and whole tasks and operate under a variety of conditions to be expected in the aircraft. For example, to acquire proficiency at landing, the "rate of descent" may be the first and most critical behavior for a safe landing and will thus be practiced and measured independent of other factors.

6.1.7.2 Simulator: Criterion Development. In order to derive a valid evaluation standard, provision in the simulator should be made for collecting tolerance data which are considered to be indicative of criterion performance in anticipated inflight conditions. An example will illustrate the thrust of this effort. Although inflight emergency procedures are usually initiated by the malfunction of a single source such as engine overheating, several problems created and compounded by the single malfunction result (e.g., overheating which causes instability anu loss of fuel pressure). Both simple and complex scenarios will be simulated in F-16 instruction and performance measurement to properly prepare the student for real emergencies.

Optimal perfo~mance parameters can be generated by the computer for simulated combat maneuvers. For example, after having recorded and summarized the exact flight path of a barrel roll over hundreds of trials by experienced pilots, and from concensus on flight parameters by subject matter experts, the computer will be able to evaluate the student's performance by his deviation from the acceptable boundaries of this optimal path or tunnel, much like the present Automated Adaptive Flight Training System (see AFTS, Logicon, 1977) being tested. Procedures also should be monitorable by the simulator.

Gradeslips used in simulator evaluation will be the same as those described in this chapter. As much as simulator fidelity will allow, the same criteria will be employed for the use of gradeslips in simulators as for inflight missions. In addition to instructor-produced gradeslips, the simulator should be capable of producing printouts of readouts from the simulator console showing a summary of student performance (e.g., flight paths, critical momentary values). One influence on simulator gradeslip evaluation not available in the airborne environment will be the confirmatory or disconfirmatory information provided by the simulator to be compared with IP ratings. These printouts may be used as backup information either to be given to the student or to be retained for diagnostic use with future IP's.

To facilitate the generation of objective diagnostic and evaluative data, it is recommended that the simulator have stop action and playback capabilities. In this way, the supervisor, the IP, and other trained personnel can point out errors or demonstrate the correct procedure for successful completion of the mission task. Individual tasks will be automatically flagged so that playback occurs only for the relevant segment of training (see Logicon, 1977). Also proposed will be a time-indexed playback capability, such that the IP can review, say, the last 10 seconds of a specific maneuver, a feature which will increase training efficiency and emphasize only errors.

It is hoped that maneuvers can be transitioned from less to more difficult scenarios, a property now part of the AFTS adaptive training system. Initial attempts would be standardized, using predictable external inputs such as a bogie or weather. Later, unpredictable circumstances more like actual combat conditions would be introduced by the IP. Because it is the intent of the F-16 performance measurement system that simulator time not be misused, we will also consider using the simulator for certain lower-order, or part-tasks which can slowly be integrated within the same simulator environment into higher-order tasks.

6.1.7.4 *Simulator evaluation and progress*. During any standardized exorcise in the simulator or trainer, corrective feedback will be provided. If a student reaches the criterion with extra instructional time alotted, F-16 will recommend implementation of alternative sortie plans which will be fully approved in the syllabus and be directly related to the block objective, utilizing the proficiency advancement (alpha ride) concept described above. This format recognizes that some students require less time toward mastery, and that by adapting to his capabilities, the instructional system both creates a better trained pilot and provides a highly motivating opportunity toward which to strive. In addition, a certain degree of normative data can be gleaned from the incident of certain students consistently achieving an alpha ride.

6.1.7.5 *Simulator: Revisions*. Revisions in simulator performance measurement will come from changes in the CROs, the weapon system's mission, and the task syllabus. Rate of acquired proficiency in relation to progress report data and projected learning curves will denote a need to adjust sequencing or difficulty level of instruction.

6.1.8 Inflight Tools

Except for the problems unique to the gathering of data during an inflight exercise, the measurement tools and the production of their results will be the same as those of the simulator.

First described will be the communalities of the measurement tools. Second, the tools employed only during inflight instruction will be proposed.

6.1.8.1 *Inflight tools: General*. As specified under simulator measurement, the system will be capable of measuring both part and whole tasks. The content validity of the various levels of measurement will be established *a priori* by means of the objectives heirarchies and objectives, and *post hoc* via the quality control system. Inflight measurement will necessarily introduce more uncontrolled variation in the data-gathering environment and will thus lose some degree of situational standardization. Nevertheless, the fact that all inflight behaviors are relevant and hold direct correspondence with the instructional objectives, will more than compensate for the decrease in measurement control. Data and their associated grades will, rather, become more meaningful and useful.

The use of such tools as the gradeslips and progress reports create no special needs for inflight measurement. The following section describes two external data collection systems unique to inflight measurement which offset the automated measurement features available with simulators, and assist the IP in completing gradeslips and other evaluation forms.

6.1.8.2 Inflight tools: specific. The F-16 performance measurement system recommends the use of the VTR and ACMR/I systems described in section 5.0. The VTR will provide the IP and student with immediate playback information regarding the straight-on cockpit view and HUD data. Preliminary plans project that one half of the F-16 training models will be fitted with VTR equipment at the onset of training, and that the remaining aircraft will be retrofitted over the subsequent two years. Debrief sessions directed by the debrief guides will include the use of VTR tapes where appropriate. Debriefing rooms will have available playback equipment required for complete implementation of the system's capabilities (e.g., stop action, specific segment playback, etc.). IPs will be trained in the use of these tapes for diagnostic instruction and performance evaluation. Hopefully, time will be scheduled following the sortie for adequate coverage of the data yielded by the system. In all cases, emphasis will be placed on deviation from the CRO parameters, using the data-gathering approach best suited for the instructional task and its level of mastery. Analysis of the tapes will assist the IP in completing the gradeslip in a more objective fashion, as well as encourage diagnostic and evaluative discussion between the IP and student. Student self-evaluation separate from, or in cooperation with, the IP is a proven instructional technique which a dynamic system as the VTR can permit.

The second instructional tool unique to inflight measurement is the ACMR/I system.

Again, all support personnel will be well trained in the optimal use of the system, especially the IPs. Immediate feedback and perhaps ground-based instruction (particularly during single-seat flights) will be used when possible. Visually recorded displays will be used, both separately and in conjunction with VTR data, for completing gradeslips and conducting the diagnostic aspects of debrief.

The same type of quality control and revision procedures discussed above will also be imposed on inflight tools. The usefulness of measures will determine their continued implementation in the F-16 performance measurement system.

6.2 Personnel Involved in Performance Measurement

Academic instructors are primarily responsible for preparing the student for flight. Instruction will include the familiarization or memorization of the position and function of all aircraft instruments, the aircraft behavior corresponding to each instrument and combination of instruments, and all preparatory and post-flight procedures. Although the system design will provide a variety of instructional aids for the dissemination of information (i.e., tapes/slides, CAI, workbooks), evaluation and some diagnostics will be under at least partial control of the instructor. Under the program encouraging instructors to add test items, special training will be included for writing discriminating questions. A complete resource lab will be made available to instructors for creating instructional materials. Space and equipment will also be provided for creative endeavors (e.g., combining video instruction with part-task trainers for adaptive and interactive training). Instructors will have available the personnel and resources for suggesting new ideas, and if approved by the ISD team, will be allowed to participate in its generation and implementation.

As already stated at several points, IP instruction in performance measurement will be extensive. The IP serves the dual purpose of teaching and evaluating. It is the intention of the F-16 system that training in the use of gradeslips (diagnostics, goal analytic behaviors, recommendations), progress reports, and all automated measures allow the IP to use these tools easily and effectively. More time is also suggested for the IP to adequately employ the diagnostic measurement proposed for the F-16 training system, especially during debrief.

The duties and training of other support personnel will be dictated by the system design. The unit DO is responsible for the overall progress and proficiency of the students, as well as the quality of the instruction. The proposed progress reports will vastly simplify his role.

The system design and resulting course syllabus will specify the roles and supervision related to each job. A proposed computer-based recordkeeping system already described for academic testing data processing extended to cover performance testing also would effect roles and assignments. Definitive statements at the present time regarding them may be premature.

6.3 Recordkeeping

The F-16 performance measurement system will depend in large part on the effectiveness and efficiency of its recordkeeping system. As described above, the primary recordkeeping tools for F-16 measurement will be the progress report, gradeslips, academic test and quiz scores, and the AF forms specified in the TACR 50-31 for documentation of each student. Data from these sources will be voluminous, and automation of the summarizing and reporting process on this data base is recommended for computerization. Previous allusions to the desirability of this and a description of some of the possibilities for academic test automation have been made. A full recommendation for the automation of recording and reporting is contained in project report no. 12, "Management System Need and Design Concept Analysis," and report no. 17, "Computer Managed Instruction for the F-16 Training Program".

The general rule to be applied to any and all measurement data is that they be retained as long as is useful and meaningful. The following guidelines attempt to further define the terms "useful" and "meaningful". It is recognized that accurate determination of how long a score or records should be maintained can be ascertained only in an ongoing system. It is anticipated that the system will recommend that the gradeslips be retained following the student's graduation. Although the gradeslip has been designed to function primarily as a diagnostic tool, and is written for RTU use with RTU criteria, much information of a detailed sort will be available in the gradeslips for gaining commanders to use in continuation training. If progress on each task is preserved in the progress report so that acquisition rates are available, that report may be passed on in lieu of gradeslips, but much information will be lost in doing so. A possible recommendation for destroying the gradeslips could arise to prevent future misuse of their content. No reliable conclusions can be reached from gradeslip comments and observations long after they are written. Weight placed on such comments during accident reviews could justify their destruction.

It is proposed that the progress report either supplement or replace the present summary performance record (TAC form 180) as it contains all the information included in form 180 and is far more informative.

REFERENCES

1. Johnson, Steven L. Retention and transfer of training on a procedural task, interaction of training strategy and cognitive style, Calspan Rep. No. DJ-6032-M-1

2. DeVries, P.B.,, Curtin, J. G., Eschesbrenner, A.J. , Rosenow, J.J., and Williams, S.M. Potential applications of the Advanced Instructional System (AIS) to Air Force Navigator Training. McDonnel Douglas Astronautics Co. MDCE1673, May 1977.

3. Advanced Development Technology Plan - Training and Education Innovation. AFHRL, Sept. 1976. Projects 2359 & 2360.

4. Caro, P.W. Some factors Influencing Automated Simulator Training Effectiveness, HumRRO Technical Report 7-7-2 (March 77).

5. Websters New collegiate Dictionary, Q. & C. Merriam Co., Springfield, MA. 1977.

6. Future Undergrad pilot training system study: Final Report. Appendix XIV (Performance Measures). Northrop Corporation, March 1971 see p. 83.

7. Obermayer, Richard W., Vreuls, Donald, Muckler, Frederick A., Conway, Ernest J., Fitzgerald, Joe A., Combat-ready crew perofrmance measurement system: Final Report, AFHRL-TR-74-108 (I), December 1974.

8. Micheli, Gene S., Dr., Analysis of the Transfer of Training, Substitution, and Fidelity of Simulation of Transfer Equipment, Training Analysis and Evaluation Group Naval Training Equipment Center, Final Report--February 1972 - June 1972.

9. Hormer, Walter R., Radinsky, Thomas L., Fitzpatrick, Robert, The development, test and evaluation of three pilot performance reference scales, AFHRL-TR-70-22, August 1970.

10. Stewart, W.A., & Wainstein, E.S. Rand symposium on pilot training and the pilot career: Final report. rand Tech. Rep. R-615-PR, 1970.

11. Automated Adaptive Flight Training System AFTS for F-4E Weapon Systems Training Set (WSTS) Performance Specification--Doc. No. 02-0004, Logicon, San Diego, April 1977.

12. Air combat maneuvering performance measurement, AFHRL Doc No. FY8993-77-01-22, Luke Air Force Base, 1977.

13. Air combat maneuvering range (ACMR) Cubic Corporation Doc. P-75051. May, 1975.

14. Williges, R.C., Automation of Performance Measurement. Paper presented at the NPRDC Productivity enhancement conference, San Diego, October 1977.

15. Caro, P.W., Aircraft simulators and pilot training. Human Factors, 1973, 15, (6), 502-509.

16. Knoop, P.A., and Welde, W.L. Automated pilot performance assessment in the T-37: A feasibility study. (AFHRL TR-72-6). Wright-Patterson AFB, Ohio: Advanced Systems Deivision, Air Force Human Resources Laboratory, April 1973.

17. Development and evaluation of trainee performance measures in an automated instrument flight meaneuvers trainer, Canyon Research Group, Inc., Report NAVTRAEQUIPCEN 74-C-0063-1), May 1976.

18. Tactical Air Command special project to develop and evaluate a simulator air combat training program (Phase I) (TAC ACES I), Tactical Fighter Weapons Center (TAC), TAC project 74T-912F (Phase I), February 1977.

19. Experiments to evaluate advanced flight simulation in air combat pilot training, Development of Automated Performance Measures, Final Report, Volume 2, Northrop, Contract No. N62269-74-C-0314, March 1976.

20. Pilot Performance Measurement System, Advanced Development Program Training and Education Innovations, AFHRL-Project 2359, September 1976.

21. Applied Training Systems, Advanced Development Program, AFHRL-Project 2360, September, 1976.

22. AFHRL - Air Combat Maneuvering Performance Measurement Research Proposal - FY8998-77-01022.

23. TAC Project 76C-071F. Electronic Gunsight Sensor Evaluation. USAF Tactical Fighter Weapons Center, Nellis AFB, March 1977.

DATE
FILMED

- 8