DOCUMENT RESUME

ED 058 320 TM 001 036

AUTHOR	Horner, Walter R.; And Others
TITLE	The Development, Test, and Evaluation of Three Pilot Performance Reference Scales.
INSTITUTION	American Institutes for Research in the Behavioral Sciences, Pittsburgh, Pa.
SPONS AGENCY	Air Force Human Resources Lab., Brooks AFB, Texas.
REPORT NO	AFHRL-TR-70-22
PUB DATE	Aug 70
NOTE	71p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Aircraft Pilots; *Audiovisual Programs; Criterion Referenced Tests; *Evaluation Techniques; Feedback; Observation; Performance Criteria; *Performance Tests; Program Evaluation; Rating Scales; Student Evaluation; *Test Construction; Testing Problems; Test Reliability; Training Techniques
IDENTIFIERS	Audio Video Recording (AVR)

ABSTRACT

A set of pilot performance reference scales was developed based upon airborne Audio-Video Recording (AVR) of student performance in T-37 undergraduate Pilot Training. After selection of the training maneuvers to be studied, video tape recordings of the maneuvers were selected from video tape recordings already available from a previous research effort. Those discriminable performance events which could be observed using the video tapes were defined, and preliminary performance scales were developed to evaluate the video version of student performance. Through assessment and refining of the preliminary scales, the final pilot performance reference scales were developed. These scales were used by experienced instructor pilots to evaluate the performances shown, and results of these evaluations were analyzed. The study indicated that (a) audio-video recordings of in-flight performance can serve as the basis for the efficient development of pilot performance reference scales; and (b) video tapes can provide sufficient information for performance evaluation purposes. (Author/AG)

AFHRL-TR-70-22

AIR FORCE HUMAN RESOURCES LABORATORY

THE DEVELOPMENT, TEST, AND EVALUATION OF THREE PILOT PERFORMANCE REFERENCE SCALES

By

Walter R. Horner
Thomas L. Radinsky
Robert Fitzpatrick

American Institutes for Research

FLYING TRAINING DIVISION
Williams Air Force Base, Arizona

August 1970

Approved for public release; distribution unlimited.

AIR FORCE SYSTEMS COMMAND

BROOKS AIR FORCE BASE, TEXAS

AFPS SA

NOTICE

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

85

AFHRL-TR-70-22

August 1970

THE DEVELOPMENT, TEST, AND EVALUATION OF THREE PILOT PERFORMANCE REFERENCE SCALES

By

Walter R. Horner
Thomas L. Radinsky
Robert Fitzpatrick

American Institutes for Research

Approved for public release; distribution unlimited.

FLYING TRAINING DIVISION
AIR FORCE HUMAN RESOURCES LABORATORY
AIR FORCE SYSTEMS COMMAND
Williams Air Force Base, Arizona

FOREWORD

This study was accomplished under Project 1710, Training for Advanced Air Force Systems; Task 171003, Training Implications of New Military Technology.

The report presents the results of the Development of Pilot Perfomance Reference Scales study conducted by the American Institutes for Research, Pittsburgh, Pennsylvania for the Department of the Air Force, Air Force Human Resources Laboratory, Williams Air Force Base, Arizona, under Contract Number F33615-69-C-1366. The period of performance was from 2 January 1969 to 25 January 1970. The principal investigator for this contract was Dr. Robert Fitzpatrick and the project was directed by Mr. Walter R. Horner.

The initial contract monitor was Dr. H. J. Clark of the Air Force Human Resources Laboratory at Wright-Patterson Air Force Base, Ohio. The final contract monitor was Mr. Milton Wood of the Flying Training Division of the Air Force Human Resources Laboratory, Williams Air Force Base, Arizona. The authors wish to thank these contract monitors for their efforts in support of the study.

The authors also wish to express their appreciation to the twenty-six instructor pilots from Vance Air Force Base, Enid, Oklahoma, for their demonstrated cooperation and professionalism during scale development and the test effort. A special appreciation is expressed to Capt Richard T. Goddard who coordinated the effort at Vance Air Force Base and furnished the requisite technical support and guidance to the American Institutes for Research study team. The contributions made by Dr. Jan Winstad of the Institute of Military Psychology, Stockholm, Sweden; Dr. John T. Cowles of the Department of Psychology, University of Pittsburgh; and Dr. Melvin H. Rudov of American Institutes for Research are also acknowledged.

This technical report has been reviewed and is approved.

John G. Dailey, Colonel, USAF
Commander

ABSTRACT

This report describes the results of a study to develop pilot performance reference scales based upon audio-video recordings of in-flight performances of students undergoing T-37 undergraduate pilot training. The study included scale development as well as the test and evaluation of each scale. All the maneuvers contained on the in-flight recordings were analyzed, and constituent performance elements observable on the video replay were identified. Three maneuvers, Final Turn to Landing, Vertical S "A," and Lazy Eight, were selected for the final scaling effort. Ten performance elements each were identified for the Lazy Eight and Vertical S "A" maneuvers, and twelve elements for the Final Turn to Landing. A performance reference scale was developed for each maneuver. Each scale consisted of a series of subscales for rating performance on each of the elements of the maneuver and an additional subscale for rating the overall performance of the maneuver. Although some elements were common to more than one maneuver, the rating scales for these elements were tailored in each case to the maneuver involved. Each subscale consisted of a ten-point rating line (a row of ten boxes) representing the full range of performance from "unsatisfactory" to "excellent" and, beneath, four graded verbalizations describing different levels of performance. No verbalizations were presented, however, with the subscale used for rating overall performance. Final versions of the scales were subjected to a test and evaluation through their utilization by experienced instructor pilots. These pilots assigned levels of performance based upon what they observed on video replays of selected maneuver examples. The results showed the overall reliability of scales for the three maneuvers was high but that the majority of the individual element scales were of a relatively low to medium degree of reliability. The results are believed to justify more in-depth analysis of the data and continued development efforts to refine and increase the scope of scale application.

5

SUMMARY

Horner, W.R., Radinsky, T.L., & Fitzpatrick, R. *The development, test, and evaluation of three pilot performance reference scales.* AFHRL-TR-70-22. Williams AFB, Ariz: Flying Training Division, Air Force Human Resources Laboratory, August 1970.

Problem

Recent emphasis has been given to the experimental evaluation of airborne Audio-Video Recording (AVR) as a technique for enhancing the training effectiveness of various Air Force flying training programs. A first study showed significant training gains when airborne AVR was used to supplement Undergraduate Pilot Training in the T-37 phase of training. Subsequent studies are now underway to further define the training value of airborne AVR when used for (*a*) gunnery training in Combat Crew Training Schools; (*b*) Pilot Instructor Training; and (*c*) as an aid to gunnery training by the use of AVR through A-7D Head-Up Display.

Because airborne AVR appears to offer significant training advantages through its ability to provide rapid knowledge of results of student performance, it is also appropriate to consider the potential of airborne AVR as a source for student performance evaluation. The possible value of AVR in this regard is twofold: (*a*) as a tool for the initial development of improved performance evaluation scales and (*b*) as the prime source of student performance against which the performance scales are applied. The present study represents a first effort to quantify the value of airborne AVR as a tool for scale development as well as a source of student performance for subsequent evaluation.

Approach

A set of pilot performance reference scales was developed based upon airborne AVR of student performance in T-37 Undergraduate Pilot Training. After selection of the training maneuvers to be studied, video tape recordings of the maneuvers were selected from video tape recordings already available from a previous research effort. Those discriminable performance events which could be observed using the video tapes were defined, and preliminary performance scales were developed to evaluate the video version of student performance. Through assessment and refining of the preliminary scales, the final pilot performance reference scales were developed. These scales were used by experienced instructor pilots to evaluate the performances shown, and results of these evaluations were analyzed.

Results

As a result of the analysis, three UPT Syllabus maneuvers were chosen as the basis for scale development: Final Turn to Landing, Lazy Eight, and Vertical S "A". The subsequent scale development was highlighted by the following results: (*a*) Inconsistent and unpredictable switching between inside and outside video views often eliminated critical performance information. (*b*) Resolution of video-replay was often less than desired. (*c*) The Lazy Eight maneuver (with many outside scenes) was more difficult to score than the basic instrument maneuver of Vertical S "A". (*d*) All intervals of a 10-interval scale were used by instructors with low variability when applied shortly after video replay of performance. (*e*) Instructor use of scales showed high agreement for exemplary performance, but greater variability for poor performance. (*f*) The increased sensitivity of scales identified student problem areas more effectively than operational performance measures. (*g*) Instructors showed high agreement as to which task elements could be measured from VTR. (*h*) Between-group mean reliability was high with experimental scales.

Conclusions

Even though the pilot performance reference scales developed under this program were relatively cumbersome, and not immediately adaptable to operational use, the study did demonstrate that (*a*) audio-video recordings of in-flight performance can serve as the basis for the efficient development of pilot performance reference scales; and (*b*) video tapes can provide sufficient information for performance evaluation purposes. As audio-video recording of various in-flight maneuvers continues to grow as a function of improved AVR equipments and increased utilization as a training aid, efforts should be continued to fully utilize airborne AVR as a performance measurement device. This is particularly true in those instances where the recorded visual field contains most of the information required for evaluation purposes, such as instrument flight, gun-sight, and head-up display.

This summary was prepared by Milton E. Wood, Flying Training Division, Air Force Human Resources Laboratory.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

THE DEVELOPMENT, TEST, AND EVALUATION OF THREE PILOT PERFORMANCE REFERENCE SCALES

I. INTRODUCTION

Background

This report describes the development, test, and evaluation of three pilot performance reference scales based on video-tape recordings of in-flight performance during Air Force Undergraduate Pilot Training (UPT). Each of the scales refers to a specific maneuver and consists of a series of subscales in which varying levels of performance are distinguished for each of the maneuver elements, along with an overall summary scale. A series of video-tape recordings was also produced containing illustrative examples of performance levels for each of the three maneuvers.

This project was made feasible because audio-video recordings were being made of student pilot performance during the UPT program at Vance Air Force Base under a separate contract. This was being done as an experimental evaluation of the usefulness of in-flight performance recordings in certain aspects of the pilot training curriculum (Neese, 1968; Purifoy, 1968; Schumacher, Rudov, & Valverde, 1969).

The availability of video-tape recordings of student pilot performance represented an opportunity for research. Normally, the instructor pilot (IP) is the only observer of a student's performance. He sees the performance fleetingly, while at the same time he is coaching the student and scanning for competing traffic. His evaluation of the performance may not always be highly accurate and reliable because of these other preoccupations, the complexity of the task of flight training, or the use of other unrelated factors to arrive at a recorded grade. There is normally no way to determine the accuracy and reliability of his evaluations. Hence, efforts to carry out research aimed at improving training are hampered by the lack of any opportunity for comparing evaluations.

The availability of video recordings makes it possible for more than one instructor to observe and evaluate a given student performance, and to do this on more than one occasion. Thus, it becomes possible to compare evaluations and to take steps, if necessary, to improve them. The standardization of instructor judgments might then be furthered. Standard video tapes illustrating varying levels of student performance could be used to advantage in the training of instructors and check-pilots. In the long run, through the use and study of video tapes, it should be possible to develop more objective, and perhaps even automatic, methods of evaluating pilot performance.

The video-tape recordings, of course, do not contain all the cues which the instructor pilot may use in actual flight. This is especially true at the present level of development of the recording equipment. A major limitation, for example, is that at any given time the instructor can activate only one of two cameras, one aimed to get a view outside the airplane through the windscreen and the other focused inside on the primary flight instruments. However, even if considerable improvement were made in the video aspects of the system, it could still not represent such sensory information as that gained through kinesthesis. Hence, one of the questions to be answered is whether or not enough cues are represented with enough fidelity on the tapes to support accurate and reliable judgments of performance. This is a matter which can usefully be studied. It will be of particular importance to study if the Air Force determines that a system such as the audio-video recording system should be adopted for regular use in training.

In any case, an essential first step in the study of the video recordings was to develop scale descriptions and evaluation procedures for a sample of maneuvers to establish the feasibility of using this type of recording in further research and training. With such scales and procedures, it could then be determined whether instructor pilots can evaluate performance appropriately and consistently from observation of video recordings.

Purpose and Scope

The purpose of this study was to develop a limited number of pilot performance reference scales, by means of which the performances represented in the video recordings could be judged. No more than six nor less than three of the maneuvers listed on Air Training Command (ATC) Forms 872 and 877 check grade sheets were to be selected for final scaling.

No particular form of the scales was specified in advance. A possible approach to the problem would have been to use the grading system currently specified by Air Training Command. This system is a 4-point scale (U for Unable to accomplish, F for Fair, G for Good, and E for Excellent) in which the points are defined generally, rather than separately and specifically for each maneuver. The instructors are, of course, familiar with this scale. However, with such general definition and broad application, one cannot be absolutely certain that the points on the scale have the same meaning from instructor to instructor and, for a given maneuver, what elements enter into the assignment of the grade. Hence, it was determined that the effort should aim at scales containing more than four points and should involve analysis of each maneuver so that each scale could refer specifically to that maneuver and essential elements in the performance of that maneuver.

Preliminary Considerations

The development of any scale requires that a variety of factors be taken into consideration. Some of these factors relate to practical problems, such as the constraints imposed by the nature of the stimulus material, while others are more theoretical in nature, such as the interval properties desired for the scale or whether to take a multidimensional or unidimensional approach. Two of the preliminary considerations of those factors which influenced the development of the present scales are discussed in this section: (*a*) Whether a multidimensional or unidimensional approach would be best in the present case; and (*b*) the type of interval properties desired in the scales.

Dimensionality of Scales

Fundamental to scaling is an initial consideration as to whether to use the multidimensional or unidimensional approach. In the present case, it was decided to take an approach which is not precisely one or the other, but which is more multidimensional than unidimensional. The rationale for this decision and the apparent advantages and disadvantages of each approach as it relates to the present project are discussed.

The unidimensional approach. In a unidimensional approach it is assumed that a single dimension underlies the set of stimuli to be scaled and that judges are capable of discriminating the stimuli along this dimension. In the present case, such a dimension might be designated "goodness of pilot performance," with very poor performance at one extreme and very good performance at the other.

The primary advantage of the unidimensional approach is that it is easier to conceptualize than the multidimensional approach. Consequently, the experimental design and data analysis could be more easily prepared. Secondly, there is greater efficiency in the use of a rater's time, primarily because a rater need make only one judgment per stimulus with a unidimensional scale but must make several judgments per stimulus with multidimensional scales. Thirdly, most previous scaling work has been unidimensional, thereby providing more reference material.

The major disadvantage in using the unidimensional approach is that it produces little information about the nature of the stimuli and processes of judgment. In the present instance, with a unidimensional approach, little would have been learned about the cues to which instructors actually respond or how they integrate information from several cues to arrive at an overall performance grade.

The multidimensional approach. Judgments about pilot performance are very complex and are made up of more than one dimension. Since it was not feasible to establish, empirically, the number and type of dimensions underlying the set of stimuli to be scaled, the approach taken in this study was to make a priori evaluations concerning these dimensions. Specifically, pilot performances were separated into several performance elements. Conceptually, each performance element was considered to be a dimension.

The principal advantage of the multidimensional approach in this study was that it required a determination and analysis of the components of the complex stimulus dimension of pilot performance. Dividing a pilot performance into performance elements provides a greater opportunity to determine how instructors attend to these elements and how they integrate information about the different elements when making an overall evaluation of a pilot's performance. Another advantage is that multidimensional performance element scales can readily be adapted to new maneuvers since these maneuvers would be made up, at least in large part, of performance elements already identified. A final advantage is one of an applied nature. Separate evaluations of each performance element would permit easy identification of the performance elements with which a student is having difficulty. For example, under the current grading system, two student pilots may be graded as performing a given maneuver in a "fair" manner. Yet they may be committing entirely different errors. There is no means for determining from the grade what these errors are or why the maneuver was graded "fair" (unless, of course, the instructor makes a written comment in the "Remarks" section of the grade sheet). Furthermore, it is not apparent from the grade what specific problem a student is having with a maneuver from day to day or, even, whether it is the same problem. This observation also holds true across maneuvers where there are identical skills being learned (or not learned). For example, "effective[1] use of the power control" is one of the basic skills required of a pilot. This skill is one of those being taught by at least two of the maneuvers selected for this study: Vertical S "A" (VSA) and Final Turn to Landing (FTL). There is no way for an instructor to grade this skill on the current grade sheets. It appears that performance grading should be related to the particular skills being learned. Therefore, the effort at scale development was oriented toward this concept.

The disadvantage to use of the multidimensional approach in this study lay in the difficulty of satisfying the requirements for dimensional independence and dimensional weighting.

Ideally, the dimensions of a set of stimuli should be independent of one another. That is, a high grade on one performance element should not necessitate or be constantly associated with a particular grade level assignment on another dimension. The pilot performance elements which have been identified are not all completely independent; however, every attempt was made to reduce the number and degree of such dependencies to a minimum.

Although not explicitly dealt with in this study, weights should be assigned to every dimension, since each dimension is not necessarily equally important. In the present case where the dimensions consist of pilot performance elements, it was recognized that each performance element is not equally critical to the successful completion of a maneuver nor should they all be graded as equals. It was not obvious, though, how to weigh each performance element in precisely the correct way. Since the issue of validity was not to be tested for the performance scales in this study, the concept of criticality (or relative weight assignments) was used primarily as one of the inputs for decisions as to whether or not to retain an element as one of the group of performance elements to be graded or considered in the scale.

Scale Intervals

Scales may be ordinal, interval, or ratio. It was considered important that the scales to be developed in this project should be at least ordinal in order to have any real value in the practical application and evaluation of student performance by instructors. An interval scale was, or course, considered preferable.

It is pertinent to note here that the current scales, as reflected by ATC Forms 872 and 877, appear, at least in their descriptions, to be an admixture of the three scale types. The U, F, G, and E discriminations appear to be at equal intervals along a linear scale but, according to the numerical values indicated opposite each maneuver, the intervals are not always equal (e.g., the Lazy Eight values are 0 for U, 32 for F, 36 for G, and 40 for E). It is not known whether instructors actually use this scale in an equal or unequal interval fashion. Also, only the upper end of the scale is, to some degree, anchored. This anchor is the "ideal" or "perfect" maneuver performance. The perfect maneuver is described in official Air Force documents, but what constitutes acceptable performance variations to remain within the perfect (or E) envelope is not specified. Variations from the perfect performance for guidance in grading a performance less than E, as a G, F, or U, are not specified either. Grade assignments are made through instructor judgments based upon what he has learned at the instructor's school, his experience as an instructor, and interaction with other instructors and check-pilots.

II. DEVELOPMENT OF SCALES

This section describes the steps taken to develop the scales reported in this study. Three concepts were used as guidelines during scale development.

1. The scales should be usable by instructor pilots in an operational atmosphere. Although it was not the purpose or intent of this study to develop fully an operationally usable set of scales, it was felt that the final product, with modifications, should be adaptable to such an environment.

2. The scales should be as objective as possible. It was believed essential to reduce the number

[1] Definition of *effective* is not material to the discussion.

of subjective judgments currently required of check-pilots and instructor pilots in grading the progress or performance of other pilots (student or rated).

3. The scales should reflect the complexity of the pilot's job. Some simplification was no doubt necessary for practicality, but oversimplification was to be avoided.

Selection of Maneuvers

As stated previously, pilot performance reference scales were to be developed for not more than six nor less than three maneuvers. A maneuver was defined as any one of the line items listed on a T-37 Instrument Check Grade Sheet (ATC Form 877, February 1969) and T-37 Contact Check Grade Sheet (ATC Form 872, July 1968). The maneuvers were selected by the project monitor in consultation with, and on the basis of recommendations by, the study team.

The first step was to select a set of maneuvers which were reasonably representative of the range of flying situations in undergraduate pilot training and which were likely to be amenable to analysis for scaling purposes. Another criterion for selection was that there be a sufficient number of examples of each maneuver contained on video tapes for analysis and the conduct of reliability tests.

The following six maneuvers were selected initially: Normal Pitchout, Final Turn to Landing, Slow Flight, Lazy Eight, Barrel Roll, and Vertical S ("A" and "D" versions only). These maneuvers are defined in ATC Manual 51-4 except for two minor modifications. The entry on ATC Form 872 states "Normal Pattern and Pitchout." The "Normal Pitchout" maneuver used in this study was the pitchout only–from the instant just prior to the point of pitchout through roll-out on to the downwind leg. Also, the study maneuver "Final Turn to Landing" actually is listed as three items on ATC Form 872: Normal Final Turn, Normal Final Approach, and Normal Touchdown. The only modification made in combining these three items was that the study maneuver is completed as the aircraft touches down, the roll-out after touchdown being excluded.

A seventh maneuver was added at a later date. This was the complete Normal Landing Pattern from pitchout to touchdown, as previously defined; i.e., the downwind leg portion of the Normal Landing Pattern was added to make the entire maneuver a logical progression throughout.

A reduction in the number of selected maneuvers was effected during the initial phases of scale development. Video replays of all study maneuvers were thoroughly reviewed in order to establish a development base for the scales. As a result, the Barrel Roll and Slow Flight maneuvers were removed from consideration because neither contained a sufficient number of discriminable performance measures, observable on video replay, upon which to base the development of scales.

A further reduction of the set of maneuvers was effected following a preliminary test and evaluation of a set of interim scales. Results of this testing showed that the operational commitments of the instructor pilots, the limited time during which instructor pilots could realistically be expected to participate in the study, and the requirements for a statistically reliable base for determining the reliability of these scales necessitated the removal of the total Normal Landing Pattern and the Pitchout from the set of maneuvers. The feasibility of deleting these two maneuvers was enhanced by the fact that the three remaining maneuvers provided one example each of the low-level, high-level, and instrument phases of the syllabus. The maneuvers upon which the final performance reference scales were developed thus consisted of the Final Turn to Landing (FTL), the Lazy Eight (L8), and the Verical S "A" (VSA).

Collection of Data Base

The collection of the data base consisted of two primary steps: (*a*) selection and transcription of examples of each of the selected maneuvers from the original audio-video tapes, and (*b*) analysis of performance elements to serve as a framework upon which the scales were to be developed.

Audio-Video Tapes

The original audio-video recordings of in-flight performances were made during Phase III of Contract F33615-68-C-1048 at Vance Air Force Base. These recordings were of actual student performances throughout the T-37 contact and instrument phases of undergraduate pilot training and were contained on ½-inch video tapes. The video playback units associated with the ½-inch system do not have slow-motion or stop-action features. The development and use of pilot performance reference scales made such features mandatory. Therefore, it was necessary to have a recorder which had these features and was fully compatible with the equipment provided under Contract F33615-68-C-1048. The result was the purchase of a SONY EU-210/VTE-4 1-inch recorder and support equipment. Examples of the selected maneuvers contained on the ½-inch tapes were transcribed and organized onto maneuver-specific 1-inch tapes. The audio portion of the ½-inch tape was also recorded onto the 1-inch tape, and the student's name, instructor, date, and overall grade assigned were recorded on a second (special) audio track.

Even though recordings were made of student performances throughout their T-37 training (contact and instrument only), it was felt that students might not achieve the highest levels of proficiency during this time. In order to assure that there would be examples of each of the selected maneuvers representing high levels of performance (i.e., ideal or perfect performance), special ½-inch in-flight recordings flown by expert pilots (i.e., instructor pilots) were also collected. These examples were, as with student performances, re-recorded on the appropriate 1-inch tape and thus became part of the data base. The inventory of the number of examples of each maneuver contained on 1-inch video tape is as follows:

Maneuver	Number of Examples
Pitchout	60
Final Turn to Landing	94
Normal Landing Pattern	17
Lazy Eight	33½
Vertical S ("A" and "D")	70
Slow Flight (dropped)	30 (collection not completed)
Barrel Roll (dropped)	26 (collection not completed)

Each example was further identified in terms of its location on the reel (using counters provided), whether the maneuver was flown by a student or instructor, preliminary remarks as to the pilot performance deficiencies illustrated on the video recording, the quality of the recording, and an estimated grade (U, F, G, or E) for the particular example. The latter grade was assigned by the project staff purely for use as a guideline to indicate relative quality of pilot performance for fast retrieval of examples of maneuvers at different performance levels.

Performance Elements

Another input into the data base from which the pilot performance reference scales were developed was the formulation of performance elements. These elements are descriptions of segments, activities, conditions, or skill requirements which, when totaled, describe a maneuver. As will be seen, these elements were initially developed for each of the study maneuvers and then refined into a set of elements each of which could be applicable to more than one of the selected maneuvers.

Appendix I is a table of the initial performance elements developed so as to gain greater insights into performance requirements of the maneuvers to be scaled. The elements are maneuver-oriented; that is, each maneuver was analyzed to determine the performance elements applicable to that maneuver. A better approach to developing performance elements might have been to start with the skills required of an Air Force pilot based upon task analyses and then to relate the skills to be learned to the maneuvers which have been included in the current undergraduate pilot training syllabus to teach these skills and which are of concern to this study. However, the scope of the study did not permit such an approach.

The sources of information which provided the basis for Appendix I were as follows:

ATC Manual 51-4, 12 June 1967.
RAFB Student Study Guide F1115070-5, June 1967.
Written comments from instructors which included common student errors.
Verbal comments from a panel of instructors from Vance Air Force Base.
(These instructors are not only highly qualified as flight instructors but all have had some degree of experience with the Audio-Video Recording System.)
Instructor comments contained in the "Remarks" section of completed T-37 Contact and Instrument Check Grade Sheets.
Study staff experience with these maneuvers.

Each maneuver is represented by named segments. That is, there is a general or overall segment which presents an overview of the maneuver followed by logical groups within sequential blocks of time which, together, make up the total maneuver. The L8 maneuver, for example, is made up of nine groups: overall and eight logical checkpoints (45°, 90°, 135°, 180°, 225°, 315° and 360°) throughout the maneuver. It is recognized that other breakdowns or groupings of maneuver performance measures could have been made. The column headings are explained as follows:

"Activity or Condition." The items listed under this heading are concerned with what is expected of the pilot during that particular segment of the maneuver and what the initial or ending condition or result should be. For example, there are given conditions which must exist before a pilot starts an L8 maneuver. These are given in the "Overall" section. Then, too, at the 45° point there are certain conditions which are considered ideal for proper performance of the maneuver. Also listed in this column are the activities which contribute to ideal maneuver performance.

"Indicator or Sense." The items under this column include the instrument or sense which is used by an instructor to make judgments about the performance of the activity or condition. One item, "feel," requires some explanation. It is meant to convey that accumulation of all the cues a pilot receives from his environment, instruments, and senses which result in a greater awareness of the performance or judgments as to the degree of "goodness" of performance. "Feel" is also used in a simpler context such as a requirement to touch the landing gear lever during a "gear down and locked" check to ascertain that one condition of the landing gear being in the extended position is met.

"Indications or Stimulus." The items listed in this column reflect the manifestation of the activity or condition.

"Decision Factors." These items are, as the name implies, factors which must be considered, at a minimum, by the instructors who grade the performance of the given activity or condition.

"Performance Criteria." The items in this column were originally intended to convey the performance parameters which would indicate the degree of "goodness" of an activity or condition. As stated previously, there are no such criteria known to exist officially at this time and, therefore, the item is either listed as "none" or "IP judgment." The latter is meant to indicate that the basis for judging the degree of "goodness" of that particular facet of student performance is left to the discretion of the instructor pilot. The item "none" implies just that–there is no criterion to define the degree of "goodness." For example, in the setup for the L8, the manual calls for 200 knots as the starting airspeed. What if a pilot starts the maneuver at some speed other than 200 knots? Under the current grading system, at what airspeed other than 200 knots does his performance become G, F, or U? What are the airspeed limits for an E performance? This is, obviously, a simple example; it becomes much more complex considering a maneuver as a whole or a complicated segment of a total maneuver, such as the final approach to a landing.

"TV System Capability." The items "Yes" and "No" indicate that the performance of a given activity or condition can or cannot be observed to some degree through use of an audio-video recording system.

"Criticality of Performance." The items in this final column were preliminary judgments made by the project staff as to the relative importance or contribution of the performance of a given activity or condition to the overall performance of a perfect maneuver. It is a simple scale of *3* for extremely important, *2* for moderately important, and *1* for minor importance. The information in this column was used as a guideline during the development of the scales as an indication of priority. Appendix I does not include the Barrel Roll or Slow Flight maneuvers which had been, as previously stated, dropped from study consideration.

The content of Table I was prepared from the data contained in Appendix I. The performance elements were simplified and assigned an identification number for future data control. They were also worded such that each element was generalizable across maneuvers with specific definitions reserved for the applicable maneuver. Table I also shows the applicability of a given element to the five maneuvers still under consideration at this point in the study. The "NA" notation indicates that the given element is not applicable or is of minimal importance to the successful completion of an ideal performance or has minimal effect on the overall grade assigned to the maneuver (e.g., element 2B, the control of airspeed during a Pitchout). No notation indicates that the element is applicable. The reason for showing the Lazy Eight as two maneuvers is explained in the following section.

In summary, Appendix I and Table I are the results of establishing identifiable variables associated with the performance of a given maneuver and with the grading of the maneuver. Some identifiable variables are, of course, not included since they could not be judged through utilization of an Audio-Video Recording System (AVRS).

Table 1. Applicability[a] of Performance Elements to Maneuvers

Performance Element	Maneuver					
	Pitch-out	Final Turn to Landing	Normal Landing Pattern	Lazy Eight: First 180°	Lazy Eight: Second 180°	Vertical S "A"
1. Specific check point "book" criteria or setup requirements						
A. Airspeed						
B. Altitude					NA	As Directed
C. Attitude						
D. Heading						As Directed
E. Positioning of Aircraft						NA
F. Aircraft Configuration						NA
G. Trends of:						
1. Airspeed					NA	
2. Altitude					NA	
3. Attitude					NA	
2. Control of:						
A. Power				NA	NA	
B. Airspeed	NA					
C. Altitude				NA	NA	As Directed
D. Heading						
E. Pitch Angle						
F. Rate of Roll		NA				NA
G. Angle of Bank						NA
H. Rate of Turn						NA
I. Rate of Pitch Change	NA	NA	NA			NA
J. Rate of Ascent/Descent	NA			NA	NA	
3. Cross Check						
4. Error Correction						
5. Transitioning	NA	Roundout	Roundout	NA	NA	
6. Use of Trim				NA	NA	
7. Safety (clearing turns or spacing)				NA	NA	NA
8. Cockpit Procedure (Audio only)	NA			NA	NA	NA
9. Use of Ground Reference Points or Lines						NA
10. Aircraft Configuration				NA	NA	NA
11. Aircraft Operation Within Published Limitations				NA	NA	NA
12. Touchdown	NA			NA	NA	NA
13. Radio Procedure (Audio only)	NA			NA	NA	NA

[a]NA = Not applicable, or of minimal importance.

Preliminary Performance Scale Development

As in most scale developments, plans were made to develop and test a preliminary set of scales in order to determine whether the approach being taken was reasonable and merited continued efforts or required a change. An additional incentive for developing a preliminary set of scales was that it could be used to learn more about the maneuvers and about the process through which an instructor relates performance to a scale or a chosen point on the scale. One way to identify those obscure or poorly understood aspects of factors an instructor uses to make judgments on levels of performance is to ask instructors to verbalize what they are seeing or looking for during a performance. An instructor confronted with an incomplete preliminary scale for use as a reference upon which to mark a level of performance would be provided with the requisite stimulus to verbalize.

MANEUVER *VERTICAL S "A"* (page 1)

Performance Element	"Ideal"	Not Obs	Performance Rating			
OVERALL			U			E
1. Specific checkpoint "book" criterion or setup requirements: A. Airspeed	160 kts		Over ±10 kts	Within ± 10 kts	Within ± 5 kts	Within ± 2 kts
1. Specific checkpoint "book" criterion or setup requirements: B. Altitude	Steady as directed		Not settled on altitude			On altitude assigned
1. Specific checkpoint "book" criterion or setup requirements: C. Attitude	level		Started maneuver in an unsettled flight attitude			Wings level and pitch angle set for level flight
1. Specific checkpoint "book" criterion or setup requirements: D. Heading	Steady as directed		Not settled on heading at maneuver start			On heading assigned by IP

Fig. 1. **Sample page of preliminary pilot performance reference scale.**

16

Figure 1 is a representative example of the preliminary scales that were developed. A separate set of scales was developed for each of the five maneuvers, with a double set for the Lazy Eight. The latter maneuver was divided into two parts, the first 180° and the second 180°. This was due to observations from the video tapes and comments by the instructors that a student generally performs that part of an L8 which goes to the right better than the one which goes to the left because of the side-by-side cockpit arrangement in the T-37 aircraft. Therefore, separate evaluations seemed to be appropriate.

To continue with the explanation of Figure 1, the first element for each maneuver was the "Overall" performance. The scale developed for grading the overall performance was a 12-point, equal-interval, unidimensional scale anchored near the two extremes with a U and an E. The primary objective behind this scale was to provide the grader with a greater number of discrimination possibilities than was thought to be required to effectively grade the performance. It was desired to relate the scale to the current 4-point grading system, yet not to restrict the grader to a rigid relationship. Therefore, only the U and E were placed on the scale. As can be easily ascertained, the 12 points can represent U–, U, U+, F–, F, F+, G–, G, G+, E–, E, and E+. Tests of this preliminary scale were to determine how many discriminations were made by the instructors in grading a series of maneuvers. No verbalization was attempted to describe the points along this overall grading scale. Such a task would be forbidding when one considers the number of ways a maneuver could be performed in a "good" manner—especially, complicated maneuvers such as the Normal Landing Pattern. With a verbalized scale, the "good" point, for example, would have to contain a description of each possible way of doing the maneuver which could result in a "good" grade. Such descriptions come within the realm of possibility when maneuvers are broken down into the segments or elements which make up a maneuver. Each element can be more readily verbalized along its dimension.

Again referring to Figure 1, it can be seen that each element contains at least two and sometimes four verbal descriptions across its dimension. The scale used for each of the elements was a 4-point scale with no pretense that it was anything different from the current 4-point system; the only difference was the addition of the descriptions. Parts of the scale were left blank purposely, in order to take advantage of verbalizations solicited during the tests of these preliminary scales.

The final item to be explained on Figure 1 is the "Ideal" column. The purpose of this entry was two-fold. First, it provided a definition of the performance element associated with it and, second, it served as a "reminder" to the instructors as to what the ideal or perfect value or performance requirement was as cited in the ATC Manual 51-4 and other sources.

Before reporting the results of the test and evaluation of the preliminary scales, it is important to comment on an overall consideration which influenced all scale development. It was necessary to use personnel currently assigned as instructor pilots in testing the reliability of any developed scales. Except, perhaps, during unpredictable periods of foul weather, these instructors are extremely occupied in the performance of their primary duty—that of training student pilots. It is axiomatic that any scale developed which was unfamiliar or completely foreign to an instructor could not achieve acceptable reliability without extensive training as to its use. Such training would not have been operationally feasible. The decision to develop a scale which was relatable to the 4-point system and as easy as possible to comprehend and use was based on this consideration. This is not meant to be construed, necessarily, as a study limitation.

The preliminary scales, then, consisted of two parts: a 12-point, apparently equal-interval linear scale for use in grading the overall performance of a study maneuver, and a 4-point verbal scale for each of the performance elements defined for the maneuver or segment.

Test and Evaluation of the Preliminary Scales

Tests of the preliminary scales were carried out at Vance Air Force Base over a period of three days with nine volunteer instructors as participants. The objectives of the tests were as follows:

1. To gain further insights into the performance elements and to obtain additional inputs for possible refinement of the elements.
2. To obtain better verbalizations for each discriminable point on the developing scales.
3. To determine the frequency of use of each of the twelve points provided on the preliminary scales. Note that this objective does not include the determination of whether an instructor could

reliably discriminate between a G– and an F+, for example, but only the use and frequency of such discriminations.

4. To determine, from the examples of maneuvers available to the study, which of the performance elements could be graded on the basis of the video information presented. As in the previous objective, the accuracy of such grades was not an issue.

5. To obtain expert opinions from instructor pilots as to which of the performance elements were most critical to the successful (i.e., ideal) completion of a given maneuver.

6. To obtain data, to the maximum extent possible, upon which to base judgments as to how many instructors would be required to judge how many examples of each maneuver for the formal test and evaluation of the final scales still to be developed, in order to (*a*) provide the data required for statistical treatments, the results of which would be interpretable and applicable to determining the reliability of the performance scales, and (*b*) provide the basis for requesting only as many instructors as necessary for testing scale reliability (i.e., the basis for making a realistic request to the operational command for instructor pilot participation).

It was not an objective of this test to determine whether or not such scales could be effectively utilized in an operational context. The planning, conduct, and results of the test are briefly described in the following paragraphs.

In addition to preparing and developing the preliminary scales, the planning phase consisted of preparing the test video tape, preparing the criticality form, and preparing the guidelines for on-site conduct of the test. In order to obtain the maximum amount of data in the minimum of time and to effectively use the instructor participants, a special video tape was prepared. Six test examples each of the Normal Landing Pattern, Pitchout, Vertical S "A," Final Turn to Landing, and Lazy Eight were selected from the inventory of examples. In addition, one other example of the Normal Landing Pattern, Vertical S "A," and Lazy Eight were selected. The latter examples were used to orient instructors not familiar with using the Audio-Video Recording System to make judgments relative to pilot performance through the medium of a television screen. The test examples for the five maneuvers were selected by considering each of the following factors: (*a*) clarity of recordings; (*b*) appropriateness of the mix of inside and outside views; (*c*) degree to which it was judged that meaningful indications of perfomance could be identified from the recordings; (*d*) extent of variability of each of the performance elements; and (*e*) prejudgments of overall maneuver performance levels which would, hopefully, show different degrees of performance across the scale from "bad" to "good." The examples on the video tape were organized for fast identification and selected retrieval.

The audio portion of the audio-video tape was not transcribed onto the test tape. The reason for this was the number of instructional comments that were made on the tapes when originally recorded "live" which would bias the evaluation of the performance by an instructor other than the one who actually flew the mission. Independent judgments were mandatory in the test situation. This does not connote that the audio is of no value in making judgments on performance levels. In the "live" or operational situation, it would be very valuable in assisting the instructor who flew the mission to recall events for more accurate recording of performance levels on official grade sheets, and in providing a means for recording critical information on the audio portion of the tape when the instructor desires to remain with the outside view camera (for example, the airspeed at the 90° positions in the L8, or the altitude and airspeed halfway around in the FTL).

A "criticality of performance" form was also prepared in order to format instructor response to the question of criticality. This form is illustrated by the example shown in Figure 2. Each of 13 performance elements and their subdivisions (only the first three with their subdivisions are shown in Figure 2) were listed, followed by the "ideal" performance already explained. These elements with appropriate "ideals" were prepared for each of the five maneuvers. The instructors were asked to fill in the third column with a number from 1 to 5 indicating the degree of criticality of that element to successful maneuver performance. Using this scale, a *1* indicated that the element was of minor importance, a *3* indicated that it was of moderate importance, and a *5* indicated that it was of extreme importance. The instructor was also given the opportunity, in the "Remarks" column, to express his opinion as to the relevancy, definition, or whatever of any of the performance elements, or to provide missing performance elements.

MANEUVER: *VERTICAL S "A"*

Performance Element	"Ideal"	Criticality of Performance (1 - 5)	Remarks
1. Specific checkpoint "book" criterion or setup requirements			
A. Airspeed	160 kts		
B. Altitude	steady as directed		
C. Attitude	level		
D. Heading	steady as directed		
E. Positioning of a/c	NA	NA	NA
F. A/C configuration	clean		
G. Trends of			
1) Airspeed	constant		
2) Altitude	holding		
3) Attitude	steady		
2. Control of:			
A. Power	sufficient to control rates of ascent/descent & a/s through transition		
B. Airspeed	constant		
C. Altitude	as directed		
D. Heading	constant		
E. Pitch angle	sufficient to control a/s & reverse direction of vertical movement of a/c through transition		
F. Rate of roll	NA	NA	NA
G. Angle of bank	NA	NA	NA
H. Rate of turn	NA	NA	NA
I. Rate of pitch change	NA	NA	NA
J. Rate of ascent/descent	1000' per min.		
3. Crosscheck	continuous crosscheck of instruments		

NA = Not applicable, or of minimal importance

Fig. 2. Sample page of form for recording criticality of performance.

Table 2. **Distribution of Frequency of Scale Point Usage**

Maneuver	Frequency of Grades Assigned on 12-point Scale												
	U–	U	U+	F–	F	F+	G–	G	G+	E–	E	E+	Total
Lazy Eight (1st 180°)					3	1	5	3	1	2	5		20
Lazy Eight (2d 180° and overall)				3	4	5	9	8	3	4	4		40
Final Turn To Landing	1			3	1		2	2	3		1		13
Vertical S "A"			1		2	2	2	1	2	1	4		15
Normal Landing Pattern				3	2	6	8	8	4		1		32
Pitchout						2	2	3	1	2			10
Total	1	0	1	9	12	16	28	25	14	9	15	0	130
Row A		2			37			67			24		
Row B	1		10		12	44		25	23		15		

The need for an informal atmosphere during the tests in order to obtain maximum flexibility and utilization of available instructor times necessitated the preparation of minimal guidelines. Such guidelines actually took the form of procedures for test conduct.

In order to achieve maximum independence of judgments, each instructor was taken individually. The two-hour time limitation did not allow an instructor to grade examples of all five maneuvers. Therefore, the instructors graded examples of the Normal Landing Pattern and examples of two of the remaining four maneuvers. Each session consisted of four parts: introduction to the test, viewing of each example followed by grading using the preliminary performance scales, a discussion of the example just viewed, and the completion of the criticality of performance forms. During the introduction to the test, each instructor was briefed as to the objectives of the study along with the need for his inputs and expectations therefrom. He was also given an opportunity to view and discuss the three preliminary examples from the video tape. An example of a maneuver was then shown. The instructor was asked to assign a grade to the overall maneuver and then to each of the performance elements by checking the appropriate box. The instructor was allowed to view the example, or parts thereof, as many times as desired since the objectives of the test did not include testing the capability of an instructor to observe all the cues or retain what he had seen after a single viewing. In practice, this opportunity was seldom used. An audio tape recorder was used to record the discussion of each maneuver after it was shown and the grading completed.

Further insights into the performance elements and recommendations for better verbalizations (as well as for filling the gaps) at major points on the scale were, in fact, realized from the test results. The audio tape (containing maneuver discussions) and the modifications to the preliminary scale made by instructors during the actual grading of performances provided the information required. Instructors had been asked to make any changes they desired to the performance elements and scale verbalizations that were contained on the preliminary scales. As a result, instructors changed some values (such as ±5 knots) given under specific performance levels and filled in some blanks in the scale with values or words as to why that particular level of performance was assigned. This information, along with some minor inputs from the completed criticality of performance forms and the marking of the "Not Observed" column on the preliminary scale forms, showed which of the performance elements could and could not be reliably graded on the basis of the video information shown.

Table 2 shows the frequency of use of each of the 12 scale points. For easy reference and transferability to current 4-point scales, the columns have been headed U–, U, U+, F–, F, F+, G–, G, G+, E–, E, E+. There were 130 observations, with the greater number of grades being assigned to the upper half of the scale. It should be noted that the shape of this distribution, and of distributions like it, is readily affected by the extent to which the different performance levels are represented in the examples to be graded. A set of examples selected primarily from early stages of flight training would probably produce a grade distribution skewed in the opposite direction. This table shows that 10 of the 12 points were used in grading the video examples used. The figures in the row labelled "A" show the probable distribution had there been a 4-point scale, and those in row "B" show the probable distribution had there been a 7-point scale. This latter grouping appears most interesting in that the scale point F+/G– received the greatest usage by the instructors for the maneuver examples shown. It is not known whether instructors would have used that point with the same frequency had they been given a 7-point scale. Except on a relative basis the precise position of the point F+/G– on the scale is not known either.

Table 3 is a summary of the average criticality values assigned by the instructors to each of the elements by maneuver. Only the three maneuvers on which the final scales were developed are shown. Each of the values in the cells opposite the elements was obtained by adding the criticality numbers assigned by all the instructors and dividing the result by the number of instructors. An average value for each maneuver was then computed and elements with a criticality value equal to or greater than this average were highlighted. This table was used as an input to decisions made relative to the performance elements in developing final scales. As could be expected, the upper end of the 1-through-5 scale was that primarily used for value assignments, with the numbers 1 or 2 being used rarely.

Finally, the data from the test of the preliminary scales were used to provide the basis for judgments resolving two concerns relative to the test and evaluation of final scales. The first concern involved the number of instructors required to serve as judges in the test and the second involved the number of examples of each maneuver to use in the test. These two concerns are dependent upon the variability to be expected in the data. The data from the test of preliminary scales provided information about variability which was used to estimate variability in the test of the final scales.

Development of Final Pilot Performance Reference Scales

The final development of pilot performance reference scales was based upon the three maneuvers selected, Final Turn to Landing (FTL), Vertical S "A" (VSA), and Lazy Eight (L8), and upon the data collected to support their development. The preliminary pilot performance reference scales were revised by making additional refinements to the performance elements, by completing the verbalizations along the element scales, and by expanding the number of points in each element scale. The complete scales are contained in Appendix II.

The refinements to the performance elements consisted first of improving the wording and definitions (Ideal). Another kind of refinement involved reducing the number of elements to a minimum. The elements retained were those most relevant in contributing to the assignment of an overall grade and which were clearly capable of being individually judged as to performance levels using the video system. The criticality scores and an analysis of the completed preliminary performance scales provided the major inputs into decisions to retain, discard, or revise the wording of an element. A review of the audio tapes obtained during the discussion sessions provided additional insights into the elements. An example of the latter was the elimination of performance element 6, use of trim. It was obvious from the tapes that some of the instructors were grading this element based upon inferences that were being made (and, most probably, reliable inferences) from the way the aircraft was being flown or from the results achieved. Although this capability was interesting to observe, grades which were not assigned as a result of judgments made from direct observations through the video system were not considered acceptable. Table 4 is a comparative summary of performance elements which were used in the preliminary scales and those used in the final scales for the three selected maneuvers. For the Final Turn to Landing, elements 1D (heading set up), 1E (positioning of aircraft at start of final turn), 1G (trends of airspeed, altitude, and attitude), 2A (control of power), 2H (control of rate of turn), 6 (use of trim), 7 (safety), and 12 (touchdown) were removed from the final scales. The reasons for their elimination were as follows:

1D: A critical input into this judgment would be a known wind condition. This could not be ascertained from the video portion of the tape. (Audio would correct this deficiency.)

1E, 1G, 2A, 6, 7: Could not judge from the video system.

2H: Pilot performance is better judged by element 2G. Element 2H also duplicates element 2G.

Table 3. Average Criticality Ratings of Performance Elements

Performance Element	Criticality Value for Maneuver				
			Lazy Eight		
	Final Turn to Landing	Vertical S "A"	1st 180°	2d 180°	Overall Average
1A. Airspeed set up on entry	4.0	4.5 *	3.4	4.0 *	3.7
1B. Altitude set up on entry	4.1 *	4.25*	2.75	NA	2.75
1C. Attitude set up on entry	3.6	4.4 *	4.1 *	3.6	3.9 *
1D. Heading set up on entry	3.6	4.5 *	3.4	3.75	3.6
1E. Positioning of aircraft at entry	3.5	NA	3.9 *	3.75	3.8
1F. Aircraft configuration	4.9 *	3.2	3.75	3.75	3.75
1G. Trends on entry of airspeed, altitude, attitude, and heading	4.0	4.25*	3.1	NA	3.1
2A. Control of power	4.75*	4.4 *	NA	NA	NA
2B. Control of airspeed	4.75*	4.1	3.25	3.75	3.5
2C. Control of altitude	3.5	3.4	NA	NA	NA
2D. Control of heading	3.75	4.0	4.1 *	4.0 *	4.1 *
2E. Control of pitch angle	3.9	4.5 *	4.0 *	4.1 *	4.1 *
2F. Control of rate of roll	NA .	NA	4.1 *	3.9	4.0 *
2G. Control of angle of bank	3.9	NA	3.9 *	3.9	3.9 *
2H. Control of rate of turn	3.3	NA	4.0 *	4.0 *	4.0 *
2I. Control of rate of pitch change	NA	NA	4.1 *	3.9	4.0 *
2J. Control of rate of ascent/descent	4.1 *	4.5 *	NA	NA	NA
3. Crosscheck	4.4 *	4.9 *	4.4 *	4.6 *	4.5 *
4. Error correction	4.25*	4.6 *	4.2 *	4.3 *	4.25*
5. Transitioning	3.75	4.0	NA	NA	NA
6. Use of trim	4.1 *	4.0	NA	NA	NA
7. Safety (clearing or spacing)	4.75*	NA	NA	NA	NA
8. Cockpit procedures	4.25*	NA	NA	NA	NA
9. Use of ground reference points or lines	3.9	NA	4.1 *	4.1 *	4.1 *
10. Aircraft configuration	4.9 *	3.67	4.2 *	4.2 *	4.2 *
11. Aircraft operation within published limitations	4.4 *	4.7 *	4.3 *	4.4 *	4.4 *
12. Touchdown	4.1 *	NA	NA	NA	NA
13. Radio procedure (audio only)	NA	NA	NA	NA	NA
Overall Average	4.1	4.2	3.8	4.0	3.9

Note. — Asterisk indicates the element is considered to be average or above in criticality.

Table 4. **Disposition of Performance Elements in Preliminary and Final Reference Scales**

Performance Element	Element Used to Evaluate Maneuver: Final Turn to Landing	Lazy Eight	Vertical S "A"
1A. Airspeed set up on entry	x o	x o	x o
1B. Altitude set up on entry	x o		x o
1C. Attitude set up on entry	x o	x o	o
1D. Heading set up on entry	o	o	x o
1E. Positioning of aircraft at entry	o	x o	
1F. Aircraft configuration			
1G. Trends on entry of airspeed, altitude, attitude, and heading	o	o	x o
2A. Control of power	o		
2B. Control of airspeed	x o	x o	x o
2C. Control of altitude	x o		o
2D. Control of heading	x o	x o	x o
2E. Control of pitch angle	x o	x o	x o
2F. Control of rate of roll		x o	
2G. Control of angle of bank	x o	x o	
2H. Control of rate of turn	o	o	
2I. Control of rate of pitch change		x o	
2J. Control of rate of ascent/descent	x o		x o
3. Crosscheck			
4. Error correction	x o	x o	x o
5. Transitioning	x o		x o
6. Use of trim	o		o
7. Safety (clearing or spacing)	o		
8. Cockpit procedures			
9. Use of ground reference points or lines	x o	o	
10. Aircraft configuration			
11. Aircraft operation within published limitations			
12. Touchdown	o		
13. Radio procedure (audio only)			

Note. — x indicates element was used in final scale.
o indicates element was used in preliminary scale.

12: Critical inputs to a judgment of performance level are missing on video, such as hardness of landing, airpspeed at touchdown, precise moment aircraft touches ground, and precise place of touchdown.

Elements 1D, 1G, 2H, and 9 (use of ground reference points) were removed from the Lazy Eight scale for the following reasons:

1D: Identical meaning to element 1E in the Lazy Eight context.

1G: Of minimum importance with no set ideal criteria.

2H: This element is a function of elements 2G and 2I and would, therefore, be a repetitious grade.

9: Identical meaning to element 2D in the Lazy Eight context.

Elements 1C (attitude set up), 2C (control of altitude), and 6 were removed from the Vertical S "A" scale for the following reasons:

1C: Identical meaning to element 1G in the Vertical S "A" context.

2C: This element is part of element 5. The altitude, per se, is not important except as a basis to judge the capability of a student to transition.

6: Could not judge from the video system.

MANEUVER VERTICAL S "A" (page 3)

Performance Element	"Ideal"	Not Obs	Performance Rating: U			Performance Rating: E
2. Control of: B. Airspeed	Constant		Over ± 10 kts	Within ± 10 kts	Within ± 5 kts	Within ± 2 kts
2. Control of: C. Altitude	As directed		Over ± 100 feet of altitude assigned by IP			Within ± 50 feet of altitude assigned by IP
2. Control of: D. Heading	Constant		Over ± 10 degrees from assigned heading			Within ± 3 degrees of assigned heading
2. Control of: E. Pitch angle	Sufficient to control rates of ascent/descent & a/s through transition		Too great and too many pitch changes in search for proper angle			Knows proper pitch angles to use, sets pitch, and varies angle as required to maintain a/s and rate of ascent/descent

Fig. 3. **Sample page of a preliminary pilot performance reference scale.**

Maneuver VERTICAL S "A"

Page 2 of 3

Performance Element	"Ideal"	Not Observed	Performance Rating			
2B. Control of airspeed (except during transition, see element #5)	Constant	☐	Over ± 15 kts.	Within ± 15 kts.	Within ± 10 kts.	Within ± 5 kts.
2D. Control of heading	Constant	☐	Over ± 10 degrees from assigned heading.	Within ± 10 degrees from assigned heading.	Within ± 5 degrees of assigned heading.	Within ± 3 degrees of assigned heading.
2E. Control of pitch angle (except during transition, see element #5)	Sufficient to control rates of ascent/descent and a/s	☐	Too great and too many pitch changes in search for proper angle.	Uncertain as to what pitch angle to use but does not vary too excessively from proper angle.	Slightly overcontrolled and within ¼ bar width of knowing proper pitch angle to use.	Knows proper pitch angles to use, sets pitch, and varies angle as required to maintain a/s and rate of ascent/descent.
2J. Control of rate of ascent/descent	1000 feet per min.	☐	Varies more than ± 300 feet/min. from standard or varies excessively within prescribed limits of 700 to 1300 ft/min.	Held steady between 700 and 1300 ft/min.	Held steady between 800 and 1200 ft./min.	Held steady between 900 and 1100 ft/min.

Fig. 4. **Sample page of a pilot performance reference scale.**

A comparison of Figures 3 and 4 illustrates the types of changes in wording or clarity that were made to the performance elements (e.g., see element 2B). A similar inspection exemplifies the additions and modifications made to the verbalizations at the four points along the scale. Also illustrated is the revision of format that was accomplished for the final scales. The format modification consisted primarily of three parts. First, the 12-point scale used in the preliminary performance scale was modified to a 10-point scale. This change was based primarily on the results of the preliminary test shown in Table 2. Except for one occasion, the extreme upper and lower points (i.e., the U– and E+) were not used. Hence, it seemed appropriate to discontinue depicting those points. It was felt that further reductions to the number of scale points, at this stage of scale development, was not warranted from the results of the preliminary tests. The second format change was concerned with the application of the 10-point scale to each of the performance elements in addition to its application for an overall grade (the 12-point scale applied only to the overall grade and a 4-point scale to the performance elements in the preliminary scales). The third change applied to the elimination of the U and E from the scale. Although, conceptually, these values and the F and G values can be superimposed on the scale, it was desired to move as far away from such a depiction as possible while still retaining the trained instructors' familiarity with the UFGE system. The four verbal descriptions are meant to fit the scale, from left to right on Figure 4, as follows:

U	U+	F–	F	F+	G–	G	G+	E–	E

In summary, the final pilot performance reference scales consist of performance elements common across maneuvers (but not necessarily applicable to every maneuver) which describe the maneuver in terms of the skills being learned or a required condition of flight. The level of performance for each element is graded on a 10-point scale. The upper (or Ideal) and lower (or Unsatisfactory) parts of the scale and two additional points are verbally described as to the level of performance required at those approximate points on the scale. These verbalizations are specific to the maneuver and the performance element within the maneuver.

III. TEST AND EVALUATION OF SCALES

This section reports on the conduct and results of the test and evaluation of the final pilot performance reference scales. The objectives of this concluding phase of the study were to determine the number of discriminations that can be made for each of the three selected maneuvers and to determine the reliability of these scales. As has previously been noted, the notion of determining the validity of the scales was abandoned when it was concluded that there was no basis, at this point in scale development and with the current UFGE grading system, upon which to assess what the scale was truly measuring. There were two primary factors which led to this conclusion. First, there is no official documentation which defines the maneuvers in terms of differing degrees of "goodness" of performance. Secondly, the grades which were assigned to the maneuvers by the instructor who actually flew the mission "live" (and which, in the majority of cases, are included in the study data base) could not be used as a basis for establishing validity—the latter instructor had more inputs into the decision as to what to assign as a grade than are provided others by the current AVRS. For example, the element "use of trim," if performed poorly in the final approach, might change an otherwise G performance into an F performance in the "live" situation. The same maneuver would, all else being equal, be graded a G from viewing the video example since the element "use of trim" is not observable from the current AVRS recordings. Thus both the F grade and G grade would represent a valid grade with no way of relating the two. Reliability, on the other hand, can be assessed if based on the assumption that scale reliability relates directly to the capability of instructors to make standard and consistent judgments as to the level of performance of a maneuver. Neither the instructor pilots nor the study team members were oblivious to the possibility that this assumption might not be valid in all instances. However, utilization of the scales to grade a given set of maneuvers by expert and experienced instructor pilots provided the best method available for assessing scale reliability.

The two parts of this section report the planning for and conduct of the test and the analysis of the data obtained.

Test Planning and Conduct

Test Planning

The primary planning efforts consisted of determining the specific number of instructor pilots and maneuver examples required for making the requisite number of judgments for testing scale reliability, preparing the 1-inch test video tape, and selecting the test site and the insturctor pilots who were to participate in the tests.

An estimate of the minimum number of instructor pilots desired for the test effort was made both in terms of Type I and Type II errors. With respect to a Type I error, the question was one of estimating the smallest correlation which one would want to conclude as significant. In the case of the intraclass correlation, for a correlation of .50 to be significant, a minimum sample size of 12 is needed when α is .05. With respect to Type II errors, the question was one of estimating the likelihood of incorrectly accepting the hypothesis that no correlation exists in the population samples. According to Winer's (1962) discussion of the intraclass correlation, this correlation is related to an F ratio, and the power of this ratio can be determined provided the variance can be estimated and the minimum difference between means to be detected can be stated. The variance was estimated to be 2.50 on the basis of pre-study data. The minimal difference was set at one scale unit between a pair of mean scale values for a pair of maneuver examples. Using the technique discussed by Winer (1962, p. 104), it was found that β would be .90 for an α of .05, and .70 for an α of .01, with five maneuver examples and 20 observers. If the value between maneuver example means is reduced to one-half of a scale unit, β would be approximately .70 with α .05. With a smaller variance or more maneuver examples, the power of the tests increases. These considerations appeared to justify the use of a sample of 20 observers.

The principal factor limiting the maximum number of examples to be used per maneuver was the length of time available per observer. However, an estimate of the increase in the correlation coefficient anticipated from increasing the number of examples (homogeneous with the original examples) was made using the preliminary data. In the preliminary data, a correlation of .50 was representative of the intraclass correlation coefficients obtained on the basis of four examples per maneuver. Increasing the number of examples to eight results in an anticipated new correlation of .67 (Guilford, 1954, p. 391). It was, therefore, decided to use a minimum of eight examples per maneuver for the test effort.

As with the test conducted using the preliminary pilot performance reference scales, a special video tape was prepared. This tape contained examples of maneuvers which were to be judged on the degree of "goodness" of pilot performance by experienced instructor pilots using the final scales. The selection of a maneuver example for use during the test was based upon several guidelines. The primary guidelines for selection were those of clarity and completeness of the example and the estimated level of performance depicted. Other factors were the number of examples that could reasonably be judged during a two-hour period and the need for repeating examples (from those in the first test set of examples as well as those within the second set) as an input into statements regarding final scale reliability. Figure 5 summarizes the contents of the video tape developed for the final test and evaluation. As with the first test, non-test examples of the three maneuvers to be judged were included on the test tape so as to provide orientation and practice for the instructor pilots who were not familiar with the video output of the AVRS. Two of the eight examples for the Final Turn, two of the eight examples for the Vertical S "A," and one of the nine examples for the Lazy Eight were repeated, making a total of ten examples per maneuver. Final Turn examples 1 and 3 were repeated as examples 5 and 9; Vertical S "A" examples 2 and 3 were repeated as examples 8 and 9; and Lazy Eight example 3 was repeated as example 9. As can be ascertained from Figure 5, each example was identified as to its location on the tape and the length of time required to view each example and each group of examples with appropriate remarks included for use by the test director. Every attempt was made to select maneuver examples which were representative of major discriminable points across the scale, as well as examples which would illustrate varying degrees of overall performance of the performance elements. This effort was not entirely successful because such a variety of examples was not available from the data bank.

The final major planning effort for the test was to establish the time and place for the test sessions and to obtain the necessary number of judges (i.e., experienced instructor pilots). As previously stated, the video examples of the maneuvers were taped from missions flown from Vance Air Force Base. To obtain accurate judgments on level of performance, its was considered important that the judges be familiar with the terrain over which the maneuvers were flown. For example, in the Final Turn to Landing maneuver, familiarity with the ground track, the immediate surrounding area, and the active runway would allow the instructor pilot to make more accurate performance judgments. To a lesser extent, familiarity with the terrain was also important in judging the Lazy Eight where the section lines, visible on the video recordings, provide familiar cues. Such cues would not be as useful to instructor pilots from Williams Air Force Base, for example, where the terrain

Maneuver	Counter	Time Min.		Sec.	(Min.)	Remarks
NLP	0- 20	2	+	10		Hacker - Right-Hand Pattern
L8	21- 43	2	+	25	(8)	2 examples
VSA	45- 80	3	+	25		1 1/2 examples
FT 1	83- 94	1	+	05		Vance - Right-Hand Pattern
2	95-103	0	+	45		Hacker - Left-Hand Pattern
3	111-120	0	+	55		Hacker - Left-Hand Pattern
4	122-132	0	+	50		Vance - Left-Hand Pattern
5	135-146	1	+	05	(10)	Vance - Right-Hand Pattern
6	147-159	1	+	05		Hacker - Left-Hand Pattern
7	160-172	1	+	05		Vance - Left-Hand Pattern
8	173-185	1	+	05		Vance - Left-Hand Pattern
9	200-210	0	+	55		Hacker - Left-Hand Pattern
10	212-224	1	+	00		Hacker - Left-Hand Pattern
VSA 1	225-253	2	+	20		heading 350
2	255-281	2	+	15		heading 080
3	289-317	2	+	10		heading 350
4	318-347	2	+	15		heading 190
5	348-383	2	+	35	(24)	heading 230
6	384-430	3	+	15		heading 350
7	433-468	2	+	30		heading 110
8	470-506	2	+	25		heading 080
9	507-540	2	+	10		heading 350
10	542-578	2	+	10		heading 170
L8 1	579-610	1	+	50		
2	612-639	1	+	30		
3	641-661	1	+	05		Stop on instr. at end
4	662-694	1	+	40		
5	696-728	1	+	35	(15)	
6	731-764	1	+	30		
7	765-798	1	+	20		
8	800-836	1	+	30		
9	837-868	1	+	05		Stop on instr. at end
10	870-926	1	+	50		

Fig. 5. Final test and evaluation inventory of maneuver examples.

offers no such section lines and other cues must be used to assist in judging performance levels. Therefore, Vance Air Force Base was selected as the location for the tests. A request for 24 instructor pilots was made in order to allow for unpredictable cancellations. As it turned out, 23 of the 24 instructor pilots scheduled were able to participate in the test. Because of the operational commitments of the instructors, the final schedule consisted of eight two-hour sessions over a three-day period, with three instructors participating in each session. It was also requested that as many instructors as possible who participated in the first test on the preliminary scales be included among those participants in the test being described here (six of the nine so responded). It was recognized that one instructor per session would have probably resulted in a more valid test effort, but this was not operationally feasible. However, it is the opinion of the on-site director that independent judgments concerning the level of performance of the maneuver examples shown were, in fact, obtained. This opinion is based primarily upon the outstanding cooperation and interest demonstrated by the participants during the tests. Except for after-the-fact comments or an occasional spontaneous exclamation over some exceptionally poor or outstanding performance being shown, no visual or verbal interaction occurred between the instructor pilots during the test sessions.

Test Conduct

The tests were conducted on October 7, 8, and 9, 1969, with three sessions on the 7th and 9th and two sessions on the 8th. Each session consisted of two primary phases. The first phase was an approximate 20-minute period of orientation. This orientation consisted of a short briefing by the test director covering the purpose and objectives of the study, what was hoped to be accomplished from the inputs to be provided by the instructor pilots, a review of the scales to be used, an explanation of some rules governing the conduct of the test, and a viewing of a sample of each of the three maneuvers. A number of important points were covered during the orientation:

1. The instructor pilots were requested to judge performance levels using the scales provided and to accept the verbal portion of the scales as written.

2. The scales were experimental in nature and were not being developed for operational use by the Air Force during UPT, but were part of a larger program.

3. The instructor pilots were not being tested on how they graded or the accuracy of their grades, nor were their assigned grades to be compared with other instructor pilot assigned grades for any purpose other than those related to making statements about the reliability of the scales being tested.

4. Grade sheets (i.e., the pilot performance reference scales) were being identified by the name of the instructor pilot making the judgments. This was for the purpose of possible use by the project staff in the event analysis of the data suggested that additional useful information relevant to scale reliability could be obtained from instructor pilots who participated in the test.

5. The instructor pilots were advised that they could see an example or portion thereof as many times as they felt necessary, or make use of the slow-motion and stop-action capability of the video playback equipment.

6. The test director explained the importance of receiving independent judgments of performance levels and further requested that there be no verbal exchange of information about the maneuver example until after all participants had finished viewing and grading the maneuver.

7. Each of the pilot performance reference scales for the three maneuvers were reviewed as to content and use of the 10-point scale. In order to achieve maximum understanding of the 10-point scale in the minimum of time, a special depiction of the scales was shown and related to the familiar 4-point grading system. This depiction was as follows:

U		F			G			E	
U	U+	F–	F	F+	G–	G	G+	E–	E

It was also explained that the verbalizations under the scales for each element applied to an area on the scale and did not apply to any specific point on the scale.

8. The orientation phase of each session was concluded with a viewing and discussion of the non-test examples of each of the three maneuvers.

The test was conducted following very simple guidelines. The instructor pilots were given 10 copies of each of the three pilot performance reference scales for use as grade sheets. The 30 maneuver examples were then shown, in order, with each example being graded immediately after its showing. Before each showing of the Final Turn to Landing, information was provided regarding the field (i.e., Hacker or Vance), runway, and pattern (left- or right-hand). For the Vertical S "A" examples, the basic course to be maintained throughout the maneuver was given. The reasons for providing this information are fairly obvious in that they are always known to an instructor pilot in the "live" situation. Primarily because of the short duration of the maneuvers on the video, it was necessary to provide instructors with a situation wherein they would only have to concentrate on evaluating the performance.

Although it was not a formal part of the test, the participants were not discouraged from making whatever observations they desired concerning the use of the video playback to judge level of performances and concerning the associated pilot performance reference scales. These comments are summarized and discussed as follows:

1. *Element 2E of the Final Turn to Landing scale.* One instructor pilot felt that since the T-37 does not have an angle of attack indicator and the scales do not provide for "control of power" as an element, element 2E (and the description provided along the scale), per se, is of minimal value and could even be misleading. In order words, even though the pitch angle does control airspeed (as the present scale so states), it must be correlated with the use of power (which controls the rate of descent) in order to obtain a meaningful evaluation of the level of performance. It is obvious that

this often misunderstood aspect of the final approach must be clarified in any revised version of the scales.

2. *Element 2D of the Final Turn to Landing scale.* One instructor pilot stated that cross-wind correction and an angled approach were two different conditions and recommended that they be shown as separate elements. Although the recommendation is a good one, it is the opinion of the study team that additional insights into the dependencies and independencies of these two elements must be developed, and that the verbal descriptions under the scale opposite the current element 2D should be clarified.

3. *Inside vs. outside views.* Most instructor pilots commented that they would be better able to grade the Final Turn to Landing if the video views presented were of the instruments throughout all (or most of) the maneuver. The only outside view felt to be desirable would be a short period during the final approach. The participants were also of the opinion that more inside views and a more consistent expectant pattern between outside and inside views (or preferably, a split screen showing both) would result in more valid grade assignments. In general, the study team supports this comment. However, additional tradeoff analysis would have to be made before coming to any conclusions since the AVRS is, first of all, a training instrument. Also relevant in the tradeoff analysis would be the development of an effective use of the audio portion of the AVRS. As has previously been stated, no audio was provided during the test.

4. *Element 1B of the Vertical S "A".* One instructor pilot recommended that the ±10 feet verbalized at the ideal, or E, end of the scale be changed to ±20 feet and that the other values be adjusted accordingly. (The other two participants in the session concurred.) The primary reason for this change is that ±10 feet would be extremely difficult to read from the current T-37 altimeter dial, especially from a video playback using the AVRS. Although this reason is not a valid one when determining level of performance requirements for different degrees of "goodness," it becomes a serious consideration in discussions relative to the AVRS and its use as a tool for evaluating performance levels.

Table 5 presents an overall view of the number of judgments made during the conduct of the test at Vance Air Force Base. In summary, a total of 7,148 elements were judged on the basis of video playbacks of 608 maneuver examples of the Final Turn to Landing, Vertical S "A," and Lazy Eight.

Table 5. **Summary of Number of Observations by 23 Instructor Pilots**

Maneuver	N Judges	N Examples Judged	Total Examples Judged	Total Judgments: N Maneuver Examples	Total Judgments: N Judgments per Example[a]	Total Judgments: Total Judgments
FTL	23	10	230	230	13	2,990
VSA	15	10	150			
	3	6[b]	18			
	5	4[c]	20	188	11	2,068
L8	19	10	190			
	4	0[d]	0	190	11	2,090
Total				608		7,148

[a]FTL had 12 elements plus overall grade; VSA and L8 had 10 elements plus overall grade.

[b]Three IPs judged only examples 1 through 6.

[c]Four IPs judged only examples 7 through 10.

[d]Four IPs did not judge L8.

Analysis of Data

This section reports the data which were collected during the test and the results of the analysis of this data. It is organized into two subsections concerned with the number of discriminations that were made for each of the scaled maneuvers and the reliability of the scales.[2]

The data relevant to determining scale discriminatory properties were analyzed in response to the following two objectives:

1. To determine the number of times the judges used each of the 10 points on the scales.

2. To determine the degree to which the judges agreed whether they could or could not grade a performance element based upon observation of the video replay of a given maneuver example.

Since expert and experienced instructor pilots were used to judge the level of performances, scale reliability was related directly to their capability to make similar judgments. Therefore, the data relevant to determining scale reliability were analyzed with respect to answering the broad question of how well the 23 instructor pilots agreed on the grades they assigned to the maneuver examples and their performance elements.

Before proceeding with a detailed report of the data analyses, it is important to emphasize a basic assumption accepted throughout the analysis. As has been stated, the instructor pilots were requested to make their judgements on the level of performance depicted by the video replays using the scales provided and to accept the verbalizations along each of the scales as written. There is no objective evidence available upon which a statement can be made that the participants did or did not use the scales as requested. Therefore, the assumption was made that they did and the analyses conducted accordingly. Although the scope of this effort did not so require, it might have proved interesting to analyze the data under the assumption that the instructor pilots did not really use the scales and compare the results of the two analyses. To clarify what the difference is, the former assumption (the one used for this report) permits the testing of scale reliability whereas the latter assumption tests for the reliability (or standardization) of the grading of the instructor pilots. It is believed that (apparently) different results would be obtained; a detailed comparative analysis would provide additional insights and inputs into further scale development efforts.

Number of Discriminations

Given the 30 examples of the three maneuvers selected for scaling and test and evaluation, Figure 6 shows the number of times each of the ten points were used to indicate a level of performance; also shown are possible distributions of usage of points when the scaling device is collapsed into a 7-point and a 4-point scale. The scale point receiving the greatest usage is at point G for the FTL, L8, and Total, and at point G– for the VSA. Except for the L8, it also shows two breaks from an upward and downward progression at points F+ and G+. The data in Figure 6 show only that instructor pilots, given a 10-point scale, used all points on the scale for grading; whether or not they are able, in fact, to make valid 10-point discriminations, or whether or not such discriminations are important to make during UPT, was not ascertained from the available data. The points F+, G–, G+, and E– on the 7-point scale are of special interest; their relatively high usage suggests, even though projected from the 10-point scale, that there might be some validity to those points and that their possible value to the UPT program should be investigated.

Tables 6, 7, and 8 summarize the data obtained relative to the determination of the degree to which the judges agreed which performance elements could or could not be graded. Table 6 shows the number of times, by performance element and maneuver, the judges marked the "not observed" box on the scales. The term "not observed" had two meanings in the context of these scales. The first connotation was, of course, that the applicable element could not be seen on the video replay (e.g., the airspeed, element 1A, at the start of the Final Turn). The second meaning was that a judge considered that the information presented on the video replay of a given performance element was insufficient for the purpose of grading that element (e.g., if an inside view of the airspeed indicator was shown only once on a Final Turn to Landing at, say, the 90° position, the judge might decline to grade the control of airspeed, element 2B, without knowing what the airspeed was indicating during the final approach). No differentiation was made between the two meanings for this test. It can be seen from Table 6 that 1,569 out of the total of 7,148 (or 22 percent) grading possibilities were marked as "not observed." Performance elements 1A through 2C

[2] A detailed description of various aspects of the reliability studies has been prepared as a separate unpublished appendix which is available to qualified users upon request to AFHRL (FT), Williams Air Force Base, Arizona 85224.

10-POINT SCALE

	U	U+	F–	F	F+	G–	G	G+	E–	E	TOTAL
FTL	25	40	130	244	213	376	426	326	349	148	2,277
VSA	15	62	101	215	228	289	245	209	230	169	1,763
L8	10	37	85	139	169	239	296	238	210	116	1,539

10-POINT SCALE

	U	U+	F–	F	F+	G–	G	G+	E–	E	
TOTAL	50	139	316	598	610	904	967	773	789	433	5,579

7-POINT SCALE

	U	(U+F–)	F	(F+G–)	G	(G+E–)	E
TOTAL	50	455	598	1,514	967	1,562	433

4-POINT SCALE

	(U U+)	(F–F F+)	(G–G G+)	(E–E)
TOTAL	189	1,524	2,644	1,222

Fig. 6. **Usage of points on 10-point, 7-point, and 4-point scales.**

accounted for 1,354 (or 86 percent) of the total "not observed." This was due to the fact that the maneuver examples available for the test did not always begin at the study definition of the start of the maneuver, or an inside view of the airspeed, for example, was not given during the performance of the maneuver. This latter reason is best shown by element 2B. This element was marked 146 times as not observed for the Final Turn to Landing and Lazy Eight maneuvers (showing a nonstandard mixture of inside and outside views), but was always graded (i.e., no "not observed") for the Vertical S "A" (showing only inside views). This illustrates the need for greater coordination and standardization between inside and outside views obtained during a training mission so as to assure capturing a greater percentage of those performance elements pertinent to grading a maneuver. An alternative would be the incorporation of a split-screen capability in an updated version of the AVRS.

Tables 7 and 8 show the degree to which the judges, as a group, agreed which elements could and which could not be graded on the basis of what was shown of a maneuver example on the video playback. A 100 percent agreement would exist if all judges graded an element or if all marked it "not observed." The issue is dichotomous–either a performance element was graded or it was marked "not observed." The grades assigned, when assigned, were not considered in the development of Tables 7 and 8. In order to present a broader view of the agreement figures, the columns in both Tables 7 and 8 are divided into Groups 1, 2, 3, and 4. Group 1 represents 95 to 100 percent of all the instructors who graded any given maneuver or performance element. For example, 23 instructors graded each of the 10

Table 6. **Number of Times Elements Marked "Not Observed"**

Performance Element	Maneuver			Total
	FTL	VSA	L8	
1A	163	79	129	371
1B	162	80	--	242
1C	138	--	94	232
1D	--	73	--	73
1E	--	--	125	125
1G	--	68	--	68
2B	71	0	75	146
2C	97	--	--	97
2D	1	2	33	36
2E	14	1	36	51
2F	--	--	4	4
2G	6	--	29	35
2I	--	--	3	3
2J	12	0	--	12
4	9	0	19	28
5	8	2	--	10
9	32	--	--	32
Overall	0	0	4	4
Total	713	305	551	1,569

examples of the Final Turn to Landing. The number of instructors in Group 1 would then equal 22 to 23. This means that 22 to 23 instructors would have recorded a grade or marked an element as "not observed" to be considered in Group 1. Group 2 represents 90 to 100 percent of the instructors; Group 3, 80 to 100 percent; and Group 4, 70 to 100 percent. To repeat, the percentage is computed on the number of instructors grading a maneuver or performance element (23 for the Final Turn to Landing, 18 for examples 1 through 6 of the Vertical S "A," 20 for examples 7 through 10 of the Vertical S "A," and 19 for the Lazy Eight). As the percentage range broadens, more instructors are allowed into the "area of agreement."

Table 7 shows the degree of agreement across each of the three types of maneuvers and across all maneuvers. To illustrate, the first cell under Group 1 for the Final Turn to Landing (67 percent) will be used. As previously stated, there were 12 performance elements (the element "overall grade" was not included in these computations) to be graded for each Final Turn to Landing shown. This means that there were 12 x 10, or 120, cells to each of which 23 instructors were to either record a grade or mark a "not observed." By actual count, 22 or 23 of the instructors agreed in 80 of the 120 cells (or 67 percent) that they could grade or could not grade a given element in a given example. Conversely, this also means that two or more instructors did not agree with the others as to whether a given performance element could or could not be graded in 40 of the 120 cells (or 33 percent). In the same row under Group 4, 16 to 23 instructors (70 - 100 percent) agreed that they could or could not grade an element in 115 of the cells (or 96 percent). Except for the Group 1 Lazy Eight, this table shows a relatively high rate of agreement among the instructors who participated in the test as to what could or could not be graded. This table also shows that there was a greater degree of agreement between instructors in a greater number of cells in the Vertical S "A" maneuver than the other two maneuvers.

Table 8 is like Table 7, being based on similar computations. However, it is oriented towards the elements across all three maneuvers. Since element 1A is graded in each of the three maneuvers, there are 30 cells (three maneuvers times ten examples each). Element 1B is graded only for the Final Turn to Landing and Vertical S "A"; therefore, there are only 20 cells in that row. The purpose of this table is to show the level of agreement among

Table 7. Degree of Instructor Agreement by Maneuver

Maneuver	Total Number Cells	Group 1 95-100% IPs Included		Group 2 90-100% IPs Included		Group 3 80-100% IPs Included		Group 4 70-100% IPs Included	
		N Cells in Agreement	Percent	N Cells in Agreement	Percent	N Cells in Agreement	Percent	N Cells in Agreement	Percent
FTL	120	80	67	90	75	110	92	115	96
VSA Examples 1 - 6	60	49	82	51	85	53	88	53	88
VSA Examples 7 - 10	40	28	70	29	73	37	93	40	100
L8	100	47	47	64	64	77	77	89	89
All Maneuvers	320	204	64	234	73	277	87	297	93

Table 8. Degree of Instructor Agreement by Element

Element	Total Number Cells	Group 1 95-100% IPs Included		Group 2 90-100% IPs Included		Group 3 80-100% IPs Included		Group 4 70-100% IPs Included	
		N Cells in Agreement	Percent	N Cells in Agreement	Percent	N Cells in Agreement	Percent	N Cells in Agreement	Percent
1A	30	17	57	22	73	26	87	28	93
1B	20	14	70	16	80	18	90	18	90
1C	20	12	60	12	60	14	70	14	70
1D	10	4	40	4	40	7	70	8	80
1E	10	3	30	4	40	4	40	7	70
1G	10	6	60	6	60	8	80	9	90
2B	30	17	57	18	60	24	80	29	97
2C	10	0	0	0	0	5	50	7	70
2D	30	25	83	26	87	27	90	29	97
2E	30	19	63	21	70	25	83	26	87
2F	10	9	90	10	100	10	100	10	100
2G	20	11	55	17	85	17	85	18	90
2I	10	10	100	10	100	10	100	10	100
2J	20	16	80	19	95	20	100	20	100
4	30	22	73	25	83	30	100	30	100
5	20	19	95	20	100	20	100	20	100
9	10	1	10	3	30	9	90	10	100

instructors relevant to the performance elements. Performance element 2C (control of altitude in the Final Turn to Landing) is of special interest in that there is no agreement among the instructors as to their capability to grade or not grade the element until Group 3 (or at least 18 of the 23 instructors) is reached; then they agree only in half the cells. This shows that additional refinements must be made to this element in order to raise the degree of agreement. Similar analyses can be made for other elements in Table 8. Ideally, all the cells within Tables 7 and 8 should contain the 100 percent figure. Since they do not, these tables show that there must be additional refinements made to the AVRS and its application into the UPT program for purposes of both flight training and use of the audio-video tapes for evaluating levels of performance. A high percentage of agreement between instructors upon viewing a given example of a maneuver as to what can or cannot be graded should be a criterion to be met in a redesigned AVRS, if it is to be used for evaluating performance levels. Once this criterion is satisfied, the problems of re-design and utilization of an AVRS for evaluating levels of performance then become associated with the degree to which instructor pilots can agree as to what level of performance (or grade) is to be assigned to a given performance. These problems are addressed in the following paragraphs.

Scale Reliability

Scale reliability was evaluated from two different aspects, not completely unrelated, so as to provide a diversity of insights into the pilot performance reference scales, as developed, and their utilization by experienced instructor pilots. One aspect of the evaluation is concerned with a series of analyses based upon the results of scale usage by combining the judgments of the 23 instructors, as a group, who participated in the test. Included in this evaluation is the use of the intraclass correlation. The second aspect of the evaluation is concerned more with individual instructor judgments using product-moment correlations based upon their use of the scales to grade duplicate examples of maneuvers purposely incorporated in the test efforts. (Details of these analyses are available to qualified requesters.)

Table 9, 10, and 11 contain a summary of the raw data and some computations of the results of the judgments of the 23 pilots from grading examples of the Final Turn to Landing, Lazy Eight, and Vertical S "A," respectively. The data are organized by the three maneuvers, example of the maneuver (ten examples for each maneuver), and performance elements within each example. Using example 1 in Table 9 to illustrate the contents of the three tables, the *N* given for each example gives the total number of instructors who viewed the example during the tests. The performance elements listed in the left-hand column are only those applicable to that maneuver. The five figures opposite each performance element give the mean (the average score of all those recorded by the instructors who judged that element), variance, the highest and lowest grade assigned by the instructors, and the number of instructors who actually graded the element. This latter number is not always equal to the stated *N* since one or more of the instructors may have checked the "not observed" box on the scale. The number of instructors who marked a performance element "not observed" can be computed by subtracting the number who graded the element from the *N* for that example. Also computed and shown in the tables are the mean value of all the element means, the variance of the element means, the mean value of the element variances, and the variance of the element variances. The numerical values were obtained by assigning the numbers 0 through 9 to the 10 scale points from left (U) to right (E). The blank cells within each of the tables occur for two reasons: first, obviously, is that no grades were assigned by any of the instructors who viewed the example; secondly, the data were not included if less than half of the instructors viewing the example recorded a grade for a given element. It was felt that recording data from a group of responses where less than half of those responding agreed that the element could even be graded would not be representative of that group.

Figure 7 illustrates the level of performance assigned by the instructors to each maneuver example. The number used was the mean of the overall grade (e.g., 4.35 for example 1, Table 9). This value was considered the most valid since all instructors who viewed the example (except for four instances–one instance each for four different examples) had an input into determining its value. The four instances just mentioned were inadvertent omissions by the instructors concerned and not discovered in time to obtain their independent judgments. The example numbers which are circled or enclosed by a square in pairs in Figure 7 are those examples which were repeated on the video test tape. This figure also shows that examples 1, 4, and 10 (each different examples of the Vertical S "A") depict, essentially, identical levels of overall performances. An inspection, in Table 11, of the data contained under the three examples (1, 4, and 10) show differing levels of performance of the elements which make up the total maneuver. The data also show that the student in example 1 was experiencing the greatest difficulty in element 2J (control rate of ascent/descent), the student in example 4 element 2D (control of heading), and the student in example 10 not any particular element but, if anything, element 5 (transitioning) was his greatest problem. This analysis is also illustrative of the basic concept used in the developing of the scales: that of scaling the skills that are being taught a student, via performance of defined maneuvers, so that acquisitions (rate and levels) of given skills can be identified. Assignments of level of performance to overall maneuvers are not indicative of the problems in skill acquisition a student may be experiencing, nor are they indicative of similar problems concerning identical skills being taught by more than one maneuver.

Table 9. Summary Data of Final Turn to Landing

Performance Element	Ex. 1 Mean	Ex. 1 Variance	Ex. 1 Highest Grade	Ex. 1 Lowest Grade	Ex. 1 Nr. Grading	Ex. 2 Mean	Ex. 2 Variance	Ex. 2 Highest Grade	Ex. 2 Lowest Grade	Ex. 2 Nr Grading	Ex. 3 Mean	Ex. 3 Variance	Ex. 3 Highest Grade	Ex. 3 Lowest Grade	Ex. 3 Nr. Grading	Ex. 4 Mean	Ex. 4 Variance	Ex. 4 Highest Grade	Ex. 4 Lowest Grade	Ex. 4 Nr. Grading	Ex. 5 Mean	Ex. 5 Variance	Ex. 5 Highest Grade	Ex. 5 Lowest Grade	Ex. 5 Nr. Grading
	Example 1, N = 23					Example 2, N = 23					Example 3, N = 23					Example 4, N = 23					Example 5, N = 23				
1A	7.62	1.95	9	5	21					0					1					0	6.82	2.25	9	3	22
1B	7.81	2.26	9	5	21					0					1					0	7.35	1.42	9	6	23
1C	6.52	5.44	9	2	23					0					1					0	5.04	5.59	9	2	23
2B	5.35	2.87	9	2	20	5.10	2.43	8	2	19	4.50	3.31	9	2	22					1	5.00	3.26	8	0	20
2C	5.86	2.59	8	3	14					9	4.44	2.97	9	2	18					3	5.53	2.12	8	3	15
2D	3.83	5.06	9	0	23	6.30	2.49	9	3	23	6.26	2.56	9	3	23	7.22	2.72	9	3	23	3.68	5.46	8	0	22
2E	5.36	4.05	9	1	22	6.57	1.46	8	3	21	5.18	2.06	8	2	22	6.18	3.40	8	3	17	5.35	1.42	8	3	23
2G	2.09	2.75	6	0	22	6.23	4.56	9	2	22	6.00	2.00	8	3	23	5.95	4.37	9	3	20	2.04	1.22	5	0	23
2J	5.85	3.71	9	2	20	6.87	2.30	9	3	23	4.95	2.71	8	1	22	6.00	3.30	9	3	21	5.48	1.81	8	3	23
4	3.68	3.66	7	1	22	6.55	2.26	9	3	20	4.86	3.46	8	1	22	5.90	2.73	8	3	20	3.30	2.95	7	0	23
5	6.18	4.44	8	3	22	6.39	1.70	8	3	23	4.45	2.45	7	2	22	6.27	3.06	9	2	22	4.65	3.33	8	1	23
9	6.25	2.87	9	3	16	6.78	1.71	9	4	18	5.55	4.36	9	0	20	5.74	5.75	9	2	23	5.53	3.71	9	2	19
Overall	4.35	2.51	7	2	23	6.26	2.38	8	3	23	4.74	2.66	8	0	23	5.91	3.45	9	3	23	3.30	2.31	6	0	23
Mean	5.53	3.40				6.35	2.36				5.13	2.85				6.18	3.60				4.98	2.83			
Variance	2.72	1.20				0.30	0.82				0.45	0.50				0.24	1.03				2.11	2.04			

Performance Element	Ex. 6 Mean	Ex. 6 Variance	Ex. 6 Highest Grade	Ex. 6 Lowest Grade	Ex. 6 Nr. Grading	Ex. 7 Mean	Ex. 7 Variance	Ex. 7 Highest Grade	Ex. 7 Lowest Grade	Ex. 7 Nr Grading	Ex. 8 Mean	Ex. 8 Variance	Ex. 8 Highest Grade	Ex. 8 Lowest Grade	Ex. 8 Nr. Grading	Ex. 9 Mean	Ex. 9 Variance	Ex. 9 Highest Grade	Ex. 9 Lowest Grade	Ex. 9 Nr. Grading	Ex. 10 Mean	Ex. 10 Variance	Ex. 10 Highest Grade	Ex. 10 Lowest Grade	Ex. 10 Nr. Grading
	Example 6, N = 23					Example 7, N = 23					Example 8, N = 23					Example 9, N = 23					Example 10, N = 23				
1A					1					0	7.91	1.71	9	4	22					0					0
1B					1					0	8.41	1.11	9	5	22					0					0
1C					11					10	8.23	1.90	9	3	22					1					1
2B	5.12	2.38	7	2	16					0	7.11	1.40	9	5	18	5.36	1.96	8	2	22	4.71	2.51	7	2	21
2C	4.37	4.25	8	1	16					5	7.22	3.71	9	3	18	5.24	3.82	8	2	17	5.44	3.32	9	2	18
2D	6.74	2.93	9	2	23	4.78	2.54	7	1	23	7.00	1.54	8	4	23	5.87	1.75	8	3	23	5.17	3.15	9	2	23
2E	5.68	2.32	8	3	22	3.90	4.73	8	0	20	7.23	1.80	9	3	22	5.74	2.47	8	2	23	5.13	3.12	9	1	23
2G	5.27	5.16	8	2	22	3.52	5.17	7	0	23	7.26	1.57	9	5	23	6.21	2.45	8	2	23	4.43	5.98	9	0	23
2J	5.24	2.79	8	2	21	4.43	3.16	8	1	21	7.05	2.62	9	4	22	5.35	3.87	8	2	23	4.95	2.90	8	2	22
4	5.09	3.99	8	1	23	3.96	4.13	8	1	23	7.43	1.44	9	5	23	5.45	2.93	7	2	22	4.83	3.70	9	0	23
5	6.26	1.84	9	4	23	4.76	3.49	8	1	21	7.86	1.08	9	5	22	5.14	3.84	8	2	22	4.86	2.98	8	1	22
9	5.35	4.45	9	2	20	5.00	4.63	8	1	20	7.24	1.49	9	5	21	5.50	4.47	9	2	20	5.33	2.53	8	3	21
Overall	4.74	3.38	8	2	23	3.65	3.25	7	0	23	7.35	1.24	9	4	23	5.13	3.21	8	2	23	4.52	3.26	8	0	23
Mean	5.46	3.35				4.33	3.87				7.49	1.74				5.54	3.08				4.98	3.34			
Variance	0.48	1.16				0.30	0.86				0.23	0.51				0.11	0.83				0.10	0.98			

Table 10. **Summary Data of Lazy Eight**

Performance Element	Example 1, N = 19					Example 2, N = 19					Example 3, N = 19					Example 4, N = 19					Example 5, N = 19				
	Mean	Variance	Highest Grade	Lowest Grade	Nr. Grading	Mean	Variance	Highest Grade	Lowest Grade	Nr. Grading	Mean	Variance	Highest Grade	Lowest Grade	Nr. Grading	Mean	Variance	Highest Grade	Lowest Grade	Nr. Grading	Mean	Variance	Highest Grade	Lowest Grade	Nr. Grading
1A					5					0					3	7.15	2.47	9	4	13	7.41	2.38	9	4	17
1C					7	7.33	2.42	9	4	12					8	6.50	5.47	9	3	16	7.67	2.24	9	4	18
1E					1					5					5	7.35	1.94	9	5	14					8
2B	6.27	2.21	9	4	15					0	6.08	1.58	8	3	13	6.08	2.58	8	3	13	6.19	2.83	9	2	16
2D	5.88	1.99	9	4	17	4.93	3.21	8	2	15	5.50	1.50	7	3	14	6.06	4.17	9	2	18					7
2E	6.33	3.09	9	3	15	4.91	4.09	7	1	11	5.60	4.54	8	1	15	5.87	4.65	9	3	16	7.00	3.14	9	2	15
2F	6.21	3.40	9	3	19	4.56	5.56	8	1	18	5.35	2.62	9	3	17	5.42	3.48	9	3	19	6.74	2.09	9	3	19
2G	6.11	1.40	8	4	18					9	5.65	3.74	8	2	17	5.94	2.81	9	3	17	6.94	1.93	9	5	17
2I	6.26	3.43	9	3	19	5.00	2.56	8	2	19	5.28	3.62	8	1	18	5.84	4.92	9	3	19	6.47	2.71	9	5	19
4	6.19	2.43	9	4	16	4.69	3.03	8	2	16	5.53	2.27	8	3	15	5.83	3.56	9	3	18	6.28	2.09	9	4	18
Overall	6.00	2.56	9	4	19	4.67	2.82	8	3	18	5.11	2.93	7	2	18	5.74	3.76	9	3	19	6.26	2.32	9	3	19
Mean	6.15	2.48				5.24	3.38				5.57	2.85				6.20	3.62				6.84	2.41			
Variance	0.02	0.54				1.08	1.22				0.07	1.15				0.38	1.24				0.28	0.16			

Performance Element	Example 6, N = 19					Example 7, N = 19					Example 8, N = 19					Example 9, N = 19					Example 10, N = 19				
	Mean	Variance	Highest Grade	Lowest Grade	Nr. Grading	Mean	Variance	Highest Grade	Lowest Grade	Nr. Grading	Mean	Variance	Highest Grade	Lowest Grade	Nr. Grading	Mean	Variance	Highest Grade	Lowest Grade	Nr. Grading	Mean	Variance	Highest Grade	Lowest Grade	Nr. Grading
1A					2					1					0					2	7.94	1.23	9	6	18
1C	7.33	2.24	9	4	12					1					0					3	8.11	1.32	9	5	19
1E	8.00	0.89	9	7	10					1					0					2	8.00	1.67	9	5	19
2B	6.50	2.42	9	5	14	5.57	1.34	8	4	14					0	6.33	2.79	9	3	12	4.72	3.15	9	2	18
2D	6.33	2.24	9	3	18	4.89	2.22	8	2	18	5.00	3.00	8	1	13	5.11	2.46	8	2	18	5.37	4.69	9	2	19
2E	6.65	1.74	9	4	17	3.94	3.81	7	1	17	3.83	3.79	8	1	12	4.56	3.91	7	0	18	4.78	3.01	8	2	18
2F	6.47	1.71	9	3	19	4.32	5.56	8	1	19	2.95	3.83	5	1	19	4.83	3.56	9	2	18	4.74	4.43	8	1	19
2G	6.06	2.56	9	3	17	4.82	3.65	7	1	17	3.38	2.59	6	0	13	4.65	3.74	8	2	17	4.44	5.56	8	1	18
2I	6.42	1.37	8	4	19	4.15	4.81	8	0	19	2.94	3.58	7	0	18	4.28	4.21	8	1	18	5.21	3.95	8	1	19
4	6.29	2.47	9	2	17	4.12	3.99	7	1	17	2.83	3.21	7	0	18	4.88	4.49	8	1	17	5.16	4.36	8	1	19
Overall	6.21	1.84	9	4	19	4.00	4.24	8	1	18	2.89	2.81	6	0	18	4.58	3.04	8	2	19	5.00	4.33	8	1	19
Mean	6.68	1.95				4.54	3.70				3.49	3.26				4.95	3.25				5.85	3.43			
Variance	0.37	0.29				0.33	1.83				0.69	0.24				0.44	0.50				2.32	2.17			

Table 11. Summary Data of Vertical S "A"

Performance Element	Mean	Variance	Highest Grade	Lowest Grade	Nr. Grading	Mean	Variance	Highest Grade	Lowest Grade	Nr. Grading	Mean	Variance	Highest Grade	Lowest Grade	Nr. Grading	Mean	Variance	Highest Grade	Lowest Grade	Nr. Grading	Mean	Variance	Highest Grade	Lowest Grade	Nr. Grading
	Example 1, N = 18					Example 2, N = 18					Example 3, N = 18					Example 4, N = 18					Example 5, N = 18				
1A	8.11	1.16	9	6	18					6					2	6.76	1.44	8	4	17	8.39	0.37	9	7	18
1B	7.83	1.21	9	6	18					6					0	7.06	1.56	9	5	17	8.39	0.25	9	8	18
1D	7.89	1.87	9	6	18	7.89	1.36	9	6	9					3	6.59	3.38	9	2	17	8.61	0.25	9	8	18
1G	7.28	1.62	9	5	18	5.56	5.53	8	3	9					4	5.12	4.24	8	1	17	8.44	0.50	9	7	18
2B	7.06	1.47	9	5	18	4.22	2.89	8	2	18	4.61	1.19	7	2	18	5.61	2.49	7	1	18	8.50	0.38	9	7	18
2D	4.59	4.01	7	1	18	2.44	1.79	6	0	18	4.59	2.38	7	1	17	3.47	2.89	6	0	17	8.28	0.80	9	7	18
2E	5.33	2.47	8	2	18	4.67	2.47	7	2	18	4.59	1.51	7	3	17	5.39	4.02	8	1	18	8.56	0.26	9	8	18
2J	3.89	3.75	8	1	18	3.56	2.97	6	1	18	4.00	2.71	7	1	18	5.44	4.03	8	1	18	8.56	0.38	9	7	18
4	4.72	3.39	8	2	18	3.33	3.52	6	0	18	3.78	1.83	6	0	18	4.00	3.88	7	0	18	8.33	0.35	9	7	18
5	4.72	3.27	7	2	18	3.22	1.48	5	0	18	2.94	1.23	6	1	18	4.94	2.43	7	1	17	6.78	2.65	9	4	18
Overall	4.50	2.15	8	3	18	3.28	2.09	5	0	18	3.56	1.20	5	1	18	4.56	2.85	7	1	18	8.17	0.50	9	6	18
Mean	6.14	2.51				4.36	2.68				3.98	1.72				5.44	3.02				8.28	0.61			
Variance	2.68	1.04				2.96	1.66				0.57	0.38				1.33	0.97				0.29	0.48			
	Example 6, N = 18					Example 7, N = 20					Example 8, N = 20					Example 9, N = 20					Example 10, N = 20				
1A					6	7.58	1.37	9	6	19	7.24	1.69	9	5	17					3					3
1B					6	6.74	3.54	9	3	19	5.50	4.26	9	1	18					3					3
1D					6	7.58	1.70	9	5	19	7.75	1.13	9	6	16					5					4
1G					7	6.80	2.06	9	4	20	6.00	4.13	8	2	16					5					6
2B	5.56	1.32	8	5	18	7.10	0.73	8	5	20	5.00	2.21	8	3	20	5.15	1.40	8	3	20	5.10	0.62	6	4	20
2D	4.22	2.42	7	1	18	6.20	3.54	9	3	20	2.35	2.13	5	0	20	5.05	2.05	7	3	20	5.40	1.83	8	3	20
2E	4.56	2.38	7	1	18	6.60	1.20	8	5	20	4.50	1.74	6	2	20	4.70	3.06	8	1	20	4.95	1.21	7	3	20
2J	3.44	2.61	7	1	18	5.85	1.61	8	3	20	3.20	2.80	6	1	20	3.60	2.36	7	1	20	4.55	3.21	7	2	20
4	3.56	2.03	6	0	18	6.00	1.79	8	2	20	2.95	2.68	5	0	20	3.60	1.83	6	1	20	4.65	3.40	8	1	20
5	3.06	1.82	6	1	18	5.68	2.23	8	3	19	3.65	3.29	7	0	20	3.60	2.46	6	1	20	4.35	2.98	7	1	20
Overall	3.28	1.98	6	1	18	6.10	1.04	8	5	20	2.90	2.09	5	0	20	3.80	2.48	6	1	20	4.50	1.84	6	2	20
Mean	4.07	2.08				6.61	1.89				4.81	2.55				4.28	2.23				4.83	2.15			
Variance	0.83	0.19				0.46	0.85				3.33	1.00				0.58	0.28				0.15	1.13			

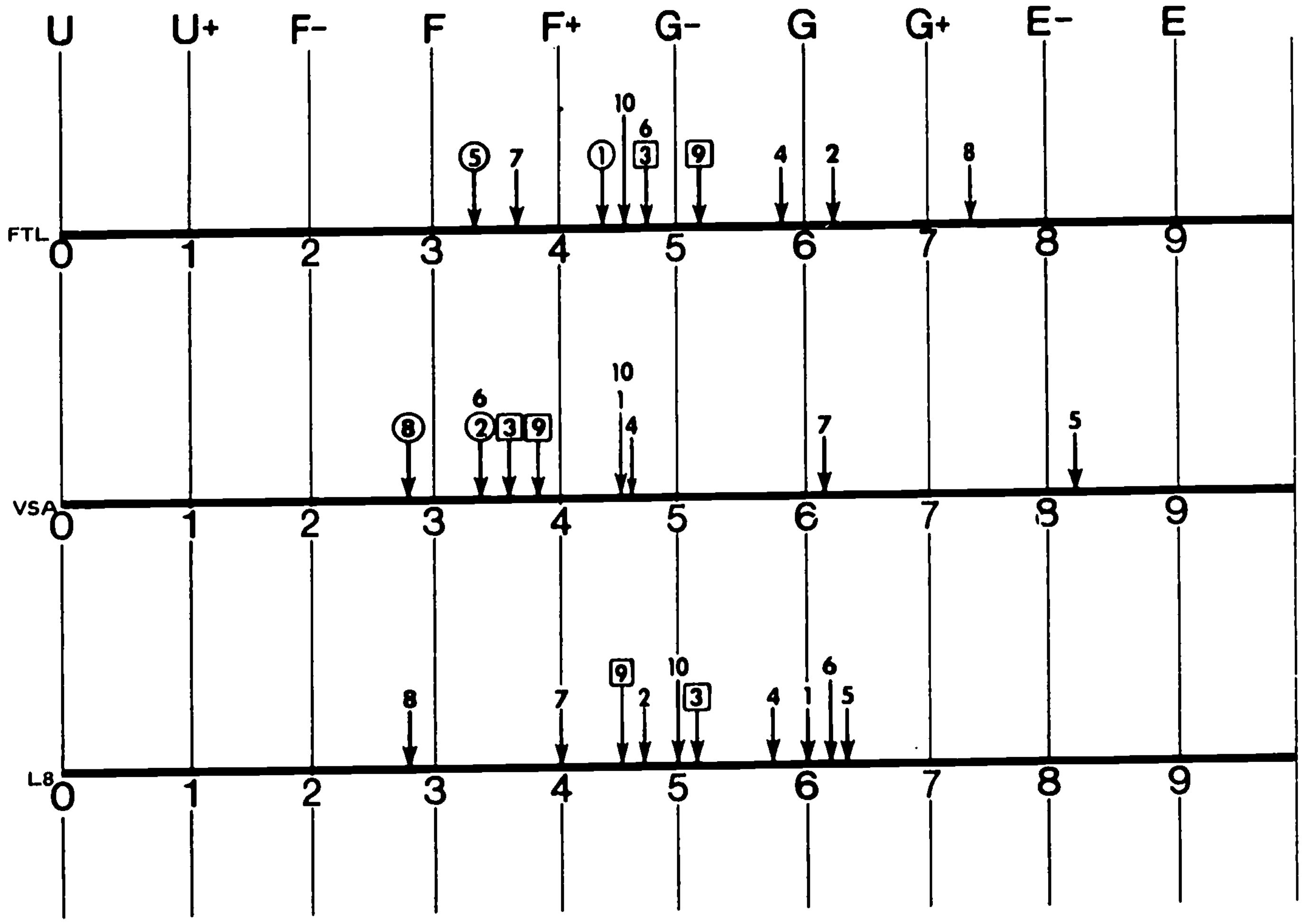

Fig. 7. **Distribution of mean overall grade of maneuver examples. (Circled numbers indicate one repeated example and numbers in squares indicate the other repeated example.)**

Table 12 presents a comparison of the mean overall grade, computed from individual assignments by all the instructors, to the mean grade computed from the performance element means. As can be ascertained from inspection, in every case the mean of the element means is greater than the mean of the overall grade. In addition, the difference between the two means is less than one in all but two of the 25 cases (duplicate examples cannot be considered different examples). The two cases (example 1 of the FTL and example 2 of the VSA) are repeated in their identical counterparts (example 5 of the FTL and example 8 of the VSA). As can be anticipated from inspection, the correlation coefficient for each of the three maneuvers is, in fact, extremely high.

It can, therefore, be stated that the pilot performance reference scales in their current state of development are highly reliable *providing* they are used and evaluated under conditions similar to those described. The most important condition is considered to be that of using a level of performance obtained by computing the mean of scores assigned by an experienced group of at least 12 instructor pilots. Obviously, such conditions are impossible to accept in an operational context and the reliability of scales must be demonstrable to the level of single instructor usage.

The concensus of those who evaluate and record level of performances is that not all of the factors which affect the final grade are weighted equally, nor do all instructors agree as to which factors are the most important (or have the greatest impact) to the establishment of a final grade. For example, one instructor may consider airspeed control to be of greater importance to a Vertical S "A" than control of the ascent/descent, and another may be of the reverse opinion. This is not only common knowledge, but the results from the initial test and evaluation of the preliminary scales, shown in Table 3, are evidence of this knowledge. Table 12 represents additional support

Table 12. **Comparison of the Mean of the Element Means and the Mean Overall Grade**

Maneuver	FTL	L8	VSA
Example 1			
Overall	4.35	6.00	4.50
Mean	5.53 o	6.15	6.14
Difference	1.18	.15	.64
Example 2			
Overall	6.26	4.67	3.28
Mean	6.35	5.24	4.36 x
Difference	.09	.57	1.08
Example 3			
Overall	4.74	5.11	3.56
Mean	5.13 x	5.57 x	3.98 o
Difference	.39	.46	.42
Example 4			
Overall	5.91	5.74	4.56
Mean	6.18	6.20	5.44
Difference	.27	.46	.88
Example 5			
Overall	3.30	6.26	8.17
Mean	4.98 o	6.84	8.28
Difference	1.68	.58	.11
Example 6			
Overall	4.74	6.2!	3.28
Mean	5.46	6.68	4.07
Difference	.72	.47	.79
Example 7			
Overall	3.65	4.00	6.10
Mean	4.33	4.54	6.61
Difference	.68	.54	.51
Example 8			
Overall	7.35	2.89	2.90
Mean	7.49	3.49	4.81 x
Difference	.14	.60	1.91
Example 9			
Overall	5.13	4.58	3.80
Mean	5.54 x	4.95 x	4.28 o
Difference	.41	.37	.48
Example 10			
Overall	4.52	5.00	4.50
Mean	4.98	5.85	4.83
Difference	.46	.85	.33
r	.939*	.986*	.934*

Note.— o = identical examples
x = identical examples
$^*p < .001$

in that in all cases there is a difference between the mean overall grade and the mean of the element means. This indicates that there are some performance element or elements being weighted more (or perhaps some a great deal more and some much less) than others. Example 8 (and its identical example 2) of the Vertical S "A" and example 5 (with its identical example 1) of the Final Turn to Landing appear to be the most likely candidates for obtaining initial data as to what weights instructors actually applied to the elements within those examples to arrive at a final overall grade so different from the straight mean of the element grade. Such an analysis would not ignore all other examples, but those mentioned would be a logical start point. This observation demonstrates the necessity for obtaining further insights into what elements are most important to a maneuver, in terms of a weighting factor as well as the important skills (or essential elements of performance) to be learned, in future development efforts to revise and refine the pilot performance reference scales presented here.

The analysis of data to this point has been concerned mainly with scale reliability based upon considerations of their overall use. Since the major development thrust of the pilot performance reference scales was oriented towards perfomance elements and a scale specific to each element and maneuver, an analysis of these scales was made. The primary measure of scale reliability is in terms of the variance of instructor pilot judgments. Under the basic assumption specified at the beginning of Section III, a high variance would indicate a low degree of scale reliability, and a low variance a high degree of scale reliability. The basic problem, however, is defining what constitutes "high" variance and "low" variance. Although it appears logical to select a variance of 2.25 (derived from assuming a normal distribution of scores with at least two-thirds of the scores falling within ±1 unit on a 10-point scale) for purposes of defining "high" and "low" reliability relative to the scales, it was considered premature, at this stage in scale development, to do so and that the data ought to suggest what the factors of scale reliability determinations should be. Therefore, variances (or scale reliability) were analyzed only on a relative basis. Results of these analyses, which are detailed in a supplementary appendix (available to qualified requesters), are reflected in the conclusions presented in the following section.

IV. CONCLUSIONS AND RECOMMENDATIONS

The following conclusions and recommendations are made based upon the result of the test and evaluation, and analysis of the data therefrom, of the pilot performance reference scales developed during this study. The scales, as given in Appendix II, are specific to each of three maneuvers included in the flight training syllabus of the Air Force UPT program—the Final Turn to Landing, Lazy Eight, and Vertical S "A." Each maneuver consists of 10 or 12 (depending upon the maneuver) performance elements individually scaled with verbalizations describing four levels of performance equally spaced along a dimension consisting of 10 possible points for discriminations.

Conclusions

The following conclusions are made relative to the scales, number of discriminations, and scale reliability.

Pilot Performance Reference Scales

1. There are no bases, from official documentation upon which to make judgments as to what a scale is truly measuring (i.e., determining scale validity).

2. Although a single instructor per session would have guaranteed independence of judgments, such independent judgments were obtained during the test effort with multiple judges.

3. The inconsistent and unpredictable switch between the outside and inside views from the video replays (which was not only disturbing to the viewer but did not depict, at times, the information needed at a particular moment) and the somewhat less than satisfactory quality of some of the video replays, affected the judgments made by the instructors to some undeterminable extent.

4. There was a high rate of agreement between the instructors as to what performance element could or could not be graded for any given maneuver example. The greatest area of agreement was with those elements of the Vertical S "A." Performance element 2C (control of altitude) in the Final Turn to Landing provided the greatest area of disagreement.

5. The pilot performance reference scales, as developed, identify specific problem areas (and the extent thereof) being experienced by a student during his training not obtainable from a maneuver-oriented overall grade.

6. The grade obtained from the mean grade assigned to each of the performance elements of a given maneuver is highly predictive of the mean of the overall grades assigned.

7. The results of use of the scale by experienced instructors showed that these instructors were in high agreement as to what constitutes exemplary performance but that the grades resulting from performances to varying degrees less than exemplary were more variable.

8. The Lazy Eight maneuver was the most difficult to grade (and the most variable) and the Vertical S "A" the easiest (and the least variable).

9. When duplicate maneuver examples were shown during a test session, the instructors were highly consistent (90.3 percent) in their judgments as to whether an element could or could not be observed from the video replay.

10. When the same maneuver example was shown on two separate occasions, about one month apart, judgments as to whether an element could or could not be observed from the video replay was a relatively stable one, with the degree of stability influenced by the nature of the maneuver, the particular maneuver example, and the instructor making the judgment.

11. The reliability of overall scale usage and some of the individual performance element scales, the probability of a high intelligence gain from a more detailed in-depth analysis of the available data, and the concepts and principles upon which the scales were developed, suggest that continued development of the pilot performance reference scales by expansion, revisions, and refinements would provide a product useful to the flying training program of the Air Force.

Discriminations

1. All points on the 10-point scale were used during the evaluation and such usage was reasonably normally distributed with scale points 6 (for the Final Turn to Landing and Lazy Eight) and 5 (for the Vertical S "A") being those representing the greatest usage.

2. Results from grading identical examples after a period of time has elapsed suggest that scales which require a relatively small number of discriminations (four in this case) are fairly reliable.

Reliability

1. The pilot performance reference scales as developed are highly reliable when used with the AVRS and the mean scores of a group of at least 12 experienced instructors are used to record levels of performance.

2. The individual performance element reference scales are considered to be of relatively high, medium, and low reliability as follows:

Final Turn to Landing

High: Altitude at start (1B); Control of airspeed (2B).

Medium: Airspeed at start (1A); Control of pitch angle (2E); Control of angle of bank (2G); Control of rate of ascent/descent (2J); Error correction (4); Transitioning (5).

Low: Attitude at start (1C); Control of altitude (2C); Control of heading (2D); Use of ground reference points or lines (9).

Vertical S"A"

High: Airspeed set up on entry (1A); Heading set up on entry (1D); Control of airspeed (2B); Control of heading (2D); Error correction (4); Transitioning (5).

Medium: Altitude set up on entry (1B); Trends on entry (1G); Control of pitch angel (2E); Control of rate of ascent/descent (2J).

Low: None.

Lazy Eight

High: Airspeed at entry (1A); Position of aircraft at entry (1E); Control of airspeed (2B).

Medium: Attitude at entry (1C); Control of angle of bank (2G); Control of rate of pitch change (2I).

Low: Control of heading (2D); Control of pitch angle (2E); Control of rate of roll (2F); Error correction (4).

3. The mean results from another sample of experienced instructors, who viewed and graded the same maneuver examples, would correlate highly with those obtained in the present study.

4. Based upon use of the scales in grading identical maneuver examples of the Final Turn to Landing and Vertical S "A" after a time lapse of approximately one month, the relative reliability of the performance element scales is as follows:

Final Turn to Landing

High: Control of pitch angle (2E); Control of rate of ascent/descent (2J); Error correction (4); Overall grade.

Low: Control of airspeed (2B); Control of heading (2D); Control of angle of bank (2G); Transitioning (5).

Inconclusive Data: Control of altitude (2C); Use of ground reference points or lines (9).

Vertical S "A"

High: Airspeed set up on entry (1A); Altitude set up on entry (1B); Heading set up on entry (1D); Control of airspeed (2B); Control of pitch angle (2E); Control of rate of ascent/descent (2J); Error correction (4); Overall grade.

Low: Trends on entry (1G); Control of heading (2D); Transitioning (5).

5. Based upon use of the scales in grading identical maneuver examples of the three study maneuvers when viewings were separated by viewing other examples and a time lapse of about 30 minutes, the relative reliability of the performance element scales are as follows:

Final Turn to Landing

High: Airspeed at start (1A); Use of ground reference points or lines (9).

Medium: Altitude at start (1B); Control of heading (2D); Control of pitch angle (2E); Control of rate of ascent/descent (2J); Control of angle of bank (2G); Overall grade.

Low: Attitude at start (1C); Control of airspeed (2B); Error correction (4); Transitioning (5); Control of altitude (2C).

Vertical S "A"

High: Control of airspeed (2B).

Medium: Control of heading (2D); Control of pitch angle (2E); Overall grade.

Low: Control of rate of ascent/descent (2J); Error correction (4); Transitioning (5).

Incomplete Data: Airspeed set up on entry (1A); Altitude set up on entry (1B); Heading set up on entry (1D); Trends on entry (1G).

Lazy Eight

High: Control of heading (2D); Error correction (4).

Medium: Control of rate of roll (2F); Control of rate of pitch change (2I).

Low: Control of pitch angle (2E); Control of angle of bank (2G); Overall grade.

Incomplete Data: Airspeed at entry (1A); Attitude at entry (1C); Positioning of aircraft at entry (1E); Control of airspeed (2B).

Recommendations

On the basis of the conclusions and the overall report of the development, test, and evaluation of the pilot performance reference scales, the following recommendations are made:

1. That the suggested additional analyses as outlined in Appendix III of the report be accomplished.

2. That the scales developed during this study be refined and revised on the basis of the results of their evaluation as given in this report (and supplemented by the additional recommended analysis) and the scope of the scales be expanded to include all the maneuvers (or pilot skills) taught during UPT. This recommendation does not necessarily encompass a requirement that the scales be (performance element be) maneuver-oriented.

3. That reliable (and valid to the greatest degree possible) pilot performance reference scales be developed prior to additional efforts to evaluate the usefulness of the AVRS as a tool for determining levels of performance.

4. That the results of this study (and the recommended additional analysis) be used as inputs into the design specifications for an updated AVRS responsive to studies for projected utilizations (or studies of different possibilities) within the flying training programs of the Air Force.

REFERENCES

Guilford, J.P. *Psychometric methods.* (2nd ed.) New York: McGraw-Hill, 1954.

Neese, J.A. *Description of audio/visual recording equipment and method of installation for pilot training.* AMRL-TR-68-73. Wright-Patterson AFB, Ohio: 6570th Aerospace Medical Research Laboratories, October 1968.

Purifoy, G.R., Jr. *Instructional methodology and experimental design for evaluating audio/video support to undergraduate pilot training.* AFHRL-TR-68-5. Wright-Patterson AFB, Ohio: Training Research Division, Air Force Human Resources Laboratory, October 1968.

Schumacher, S.P., Rudov, M.H., & Valverde, H.H. *Evaluation of a low cost in-flight audio/video recording system for pilot training.* AFHRL-TR-69-31. Wright-Patterson AFB, Ohio: Training Research Division, Air Force Human Resources Laboratory, January 1970.

Winer, B.M. *Statistical principles in experimental design.* New York: McGraw-Hill, 1962.

44

APPENDIX L. PRELIMINARY PERFORMANCE ELEMENTS OF SELECTED MANEUVERS

45

APPENDIX I. PRELIMINARY PERFORMANCE ELEMENTS OF SELECTED MANEUVERS

Activity or Condition	Indicator or Sense	Indications or Stimulus	Decision Factors	Performance Criteria	TV System Capability	Criticality of Performance
	Maneuver: Normal Landing Pattern		Maneuver Segment: General			
Ground references (a) Pitch out (3000′ up the runway) (b) Downwind (road, fence line, etc.) (c) Final turn (abeam a point) (d) Final approach (½ mile from end of runway) (e) Aim point for touchdown (dependent upon wind)	Visual	Position relative to reference points	On, off, distance off	IP judgment	Yes	2
Use of checklists	Audio Visual "Feel"	Item call out, pointing to instrument, touching control or ind.	Use, all items covered, items missed	IP judgment	Yes	1
Use of power control	"Feel," RPM ind. EGT ind. sound Fuel flow ind.	Movement, change	Increasing, decreasing, smooth, rough, steady	IP judgment	No	3
Flying aircraft onto ground at correct A/S (75 to 80 kts.), attitude (nose wheel off ground and in TO attitude), direction (centerline of aircraft pointing straight down runway), and at desired point of touchdown (within first 1000′)	Visual A/S ind. Att. ind.	Aircraft on runway and staying there	Airspeed, attitude, direction, and point at touchdown	None	Yes	3
Non-interference with other aircraft (3000′ spacing between aircraft)	Visual	Position of other aircraft	At a safe distance	IP judgment	No	1
Effective use of trim (Note: can see use of trim and how used, but not how effective its use was)	Visual Trim Light (neutral only) "Feel"	Movement of trim button Control pressure decreasing	Being used with resultant minimal control pressure	None IP judgment	Yes	2
	Maneuver: Normal Landing Pattern		Maneuver Segment: Pitchout			
Altimeter 1000′ above terrain	Altimeter	Pointer	At correct attitude (2300′ for Vance)	None	Yes	1
A/S 200 kts. at start	A/S ind.	Pointer	At "2" on indicator	None	Yes	1

Appendix I *(Continued)*

Activity or Condition	Indicator or Sense	Indications or Stimulus	Decision Factors	Performance Criteria	TV System Capability	Criticality of Performance
Throttles to 50-60%	RPM ind.	Pointer	As indicated	None	Yes	1
	Audio	Horn blowing	Known to be below 70%	None	Yes	1
Smoothly by firmly bank to approx. 60° (dependent on wind)	Att. ind. Visual	Moving marker matched approx. with 60° stationary marker on instrument	At appropriate position	None IP Judgment	Yes	2
Level turn	VV ind. Att. ind. Visual	Pointer at "0" and "airplane" attitude gradually increasing	Remaining on "0", moving up or down pitch changing smoothly	None	Yes	2
Airspeed decreasing	A/S ind.	Pointer moving counter-clockwise	Moving or steady	None	Yes	2
Roll out to level flight on heading (approx. 180° from start) with proper wind correction	Visual Att. ind. Altimeter CRSE ind. V.V. ind.	Aircraft wings level, at pattern altitude	Wind correction made, in level flight, smoothly accomplished without excessive corrections	None IP judgment	Yes	2
Spacing with other aircraft in pattern (3000' separation in pattern)	Visual	Position of other aircraft	At required separation	IP judgment	No	1
Effective use of trim (see previous note)	Visual Trimlight "Feel"	Movement of trim button decreasing	Trim being used with resultant minimal control pressures	None IP judgment	Yes	2
Constant turn rate	Visual	Progress of nose around horizon	Constant or irregular		Yes	3
	Maneuver:	Normal Landing Pattern	Maneuver Segment: Downwind Leg			
Extend speed brakes	"Feel" Visual Audio	Aircraft buffeting and slowing switch activated	Up or down	Full out	No	1
Extend landing gear below 150 kts.	"Feel", Visual A/S ind. Audio	Gear indicator "unsafe," aircraft buffeting and slowing, lights, pointer, horn on	Up, down, or partially extended	Handle down	Yes	2

Appendix I *(Continued)*

Activity or Condition	Indicator or Sense	Indications or Stimulus	Decision Factors	Performance Criteria	TV System Capability	Criticality of Performance
Five point gear check	Visual Audio Gear ind. Lights Hydr. press gage	Items called out horn off gear indication "safe" lights in handle hydr. press. up	Checks made properly and completely, lights on and off as appropriate	IP judgment Gear down and locked	Yes	3
Maintain constant altitude	Altimeter V.V. ind. Att. ind.	Pointers	At pattern altitude Pointers at "0"	None IP judgment	Yes	2
Use of power to control airspeed and altitude	RPM ind. Altimeter A/S ind. V.V. ind.	Pointers	Use of power as altitude and airspeed speed varies above or below standard	IP judgment	No	3
Minimum airspeed 120 kts.	A/S ind.	Pointer	Not below 1.2 on indicator	None	Yes	2
Correct distance abeam of r/w and holding (de; endent on wind)	Visual	Reference line on ground (for "normal" pattern)	OK, too close, too far abeam	IP judgment	No	1
Wing flaps full down (just prior to commencement of final turn - A/S less than 135 kts.)	Flap ind. A/S ind. Visual "Feel" Audio	Pointer movement of flaps handle aircraft buffeting and slowing	Up, down, or at some position in between flap handle activated at A/S less than 1.35 on ind.	Full down	Yes	3
	Maneuver: Normal Landing Pattern		Maneuver Segment: Final Turn			
Turn to final started at right point	Visual Att. ind.	Banking started movement of ind.	Started too soon too late, OK	IP judgment	No	2
Final turn radio call	Audio	Radio button activated	Made or not made correctly	Made	Yes	1
A/S decreasing to 110 kts. and holding	A/S ind.	Pointer	At 1.1 on indicator	None	Yes	2
Aircraft descending at rates to meet 90° and final approach altitude requirements	Altimeter V.V. ind.	Pointers	Altimeter pointer moving counter-clockwise VVI pointer indicating down	IP jugdgment	Yes	2
Angle of bank fairly constant (max. 45°)	Att. ind. Visual	Indicating bank	Too much, too little, or OK angle of bank	IP judgment	Yes	2

Appendix I *(Continued)*

Activity or Condition	Indicator or Sense	Indications or Stimulus	Decision Factors	Performance Criteria	TV System Capability	Criticality of Performance
Maintenance of appropriate flight track over ground	Visual	Aircraft movement along flight track (some clues visible for a "normal" track)	Outside, inside, or on flight track	IP judgment	Yes	3
No abrupt movements or large corrections	Att. ind. Visual "Feel"	Large or fast change to pitch angle of bank, power control	Too rough, OK, too much of a change	IP judgment	Yes	2
Altimeter 700′ above ground level after 90° of turn	Altimeter Visual	Pointer passing through, at check point	Too high, too slow, OK	None	Yes	2
Roll-out to wings level with aircraft heading for runway	Att. ind. Visual	Wings returning to level position runway directly ahead	Degree wings level and heading of aircraft in relation to runway	None IP judgment	Yes	2
	Maneuver: Normal Landing Pattern		Maneuver Segment: Final Approach			
Set up: (a) Altitude 300′ above terrain	Altimeter	Pointer	Set up or not and	None IP judgment	Yes	3
(b) ½ mile from approach end of runway	Visual	At reference point	amount of variability	IP judgment	Yes	3
(c) A/S 110 kts.	A/S ind.	Pointer	from "ideal"	IP judgment	Yes	3
(d) Crosswind correction established	Visual	No drift	Safe or unsafe	IP judgment	Yes	3
(e) Point of aim established	Visual	Pointer	Descending to reference point	IP judgment	Yes	3
(f) Power as required	RPM ind. V.V. ind.	Pointer		IP judgment	No	3
(g) At least 3000′ spacing on aircraft ahead	Visual	Position of other aircraft		IP judgment	No	1
Airspeed reducing to 100 kts. and holding	A/S ind.	Pointer	At "1" on indicator	None	Yes	3
Smooth power control (minimum 50%)	RPM ind. Visual Sound "Feel"	Pointers	No large and sudden increases, decreases of power	IP judgment	No	2

Appendix I *(Continued)*

Activity or Condition	Indicator or Sense	Indications or Stimulus	Decision Factors	Performance Criteria	TV System Capability	Criticality of Performance
Glide slope constant	Visual VVI	Pointer	Constant and controlled rate of descent	IP judgment	Yes	2
Longitudinal axis of aircraft aligned with runway	Visual	Pointing straight down the runway	Straight or cocked off to one side	IP judgment	Yes	2
	CRSE ind.	Pointer	On runway heading			
Roundout:						
(a) Started at right altitude (height above ground variable)	Visual	Pitch angle starting to increase	Safe or unsafe at right point or too late or too soon	IP judgment	Yes	3
(b) Pitch angle increased to landing altitude	Att. ind.	Pitch angle increasing	Increased to proper angle or not	IP judgment	Yes	3
(c) A/S decreasing from 100 kts. to touchdown A/S	A/S ind.	Pointer moving counter-clockwise	Properly executed or not	IP judgment	Yes	3
(d) Rate of descent decreased	V.V. ind.	Pointer moving up up towards "0"	Moving in right direction or steady	IP judgment	Yes	3
(e) Aircraft aiming directly down runway	Visual	Not moving to one side of runway or the other	Aligned or amount not aligned with runway	IP judgment	Yes	3
(f) Power decreasing	RPM ind.	Pointer moving counter-clockwise	Holding steady, moving OK, too fast, or too slowly	IP judgment	No	3

Maneuver: Normal Landing Pattern · Maneuver Segment: Touchdown and Roll-Out

Activity or Condition	Indicator or Sense	Indications or Stimulus	Decision Factors	Performance Criteria	TV System Capability	Criticality of Performance
Landing attitude	Visual Att. ind.	Pitch angle same as for TO	Too high, too low, OK	IP judgment	Yes	3
A/S 75-80 kts. and decreasing to taxi speed	A/S ind.	Pointer continuing to move counter-clockwise	At count A/S	None	Yes	2
Power to idle	RPM ind.	Pointer moving to idle %	At idle	None	No	1
At point of touchdown	V.V. ind.	Pointer moving to "0"	At "0"	At "0"	Yes	1
After touchdown nose wheel lowered to runway	Att. ind. Visual	Aircraft moving from above bar to below bar	Lowered gently and started at right moment	IP judgment	Yes	2

Appendix I *(Continued)*

Activity or Condition	Indicator or Sense	Indications or Stimulus	Decision Factors	Performance Criteria	TV System Capability	Criticality of Performance
Heading directly down runway	Visual	Aircraft moving straight down runway	Drifting to one side, or steady, or cocked	IP judgment	Yes	2
Aircraft touches runway	"Feel"	Aircraft touching runway	Too hard, OK, safe or unsafe	No accident IP judgment	No	3
	Visual	TV picture jiggles	Picture jiggles and at what moment	IP judgment	Yes	3
	V.V. ind.	Moves to "0"	At "0" or not	IP judgment	Yes	2
	Maneuver: Lazy Eight		Maneuver Segment: Overall			
Slow coordinated maneuver without hesitations	Visual	Rate of roll and nose track	Rate of accomplishment	IP judgment	Yes	3
Constant G	"Feel"	Constant seat pressure	Heavy, light, or negative light	IP judgment	No	2
	"G" meter	Positive or negative	Position of needle	IP judgment	No	2
180° change in direction, reverse, and complete on same start heading	CRSE ind.	Specific start heading noted in degrees	No. of degrees off heading at check points	None IP judgment	Yes	2
	Visual	Section lines or identified point	No. of degrees off heading at check points	IP judgment	Yes	1
Continuous change of pitch and bank	Att. ind.	Continuously moving	In right direction	IP judgment	Yes	2
	Visual	Continuously moving	Rate of movement	IP judgment	Yes	2
Constantly changing control pressures	"Feel"	Presure or exertion of pilot	Not measureable	IP judgment	No	2
Selection of reference point on horizon	Visual	Verbal acknowledgment and identification	90° to selected start CRSE	Indentifiable and stationary	No	3
Set up:						
(a) No clearing turn required	None	No clearing turn made	Made or not made	IP judgment	Yes	1
(b) Throttle at 90%	RPM ind.	Pointer at 90%	Position of pointer	None	No	3
(c) A/S at 200 kts.	A/S ind.	Pointer at "2"	Position of pointer	None	Yes	3
(d) Reference point off wingtip	Visual	Reference point selected	Proper selection of reference point	IP judgment	No	2

Appendix I *(Continued)*

Activity or Condition	Indicator or Sense	Indications or Stimulus	Decision Factors	Performance Criteria	TV System Capability	Criticality of Performance
	Maneuver: Lazy Eight		Maneuver Segment: First 45° of Turn			
Initial bank shallow and increasing	Visual	Wing relation to horizon	Rate of turn	IP judgment Very slow starting	Yes	3
	Att. ind.	Angle to bank slowly indicating a bank	No. of degrees of bank	Initially less than 10° angle of bank	Yes	3
Pitch being increased slowly	Att. ind.	Point of nose to horizon and continuing upward	Rate of pitch increase	IP judgment	Yes	3
	Visual	Point of nose to horizon and continuing upward	Rate of pitch increase	IP judgment	Yes	3
	A/S ind.	Decreasing	Pointer moving counter-clockwise	IP judgment	Yes	3
At 45° point:						
(a) Maximum pitch attitude approx. 45° up (internal only)	Att. Ind/ Visual	No increase being shown at top of arc	At highest point	IP judgment	Yes	3
(b) Angle of bank - approx. 45° (internal only)	Att. Ind./ Visual	½ way to 90° position	Degree of bank	IP judgment	Yes	3
(c) A/S approx. 150 kts. (internal only) (technique)	A/S ind.	Pointer at 1.5	Fast, slow, on speed	None	Yes	3
(d) CRSE 45° from original start course	Visual	Halfway around from start point to 90° reference	How close to being halfway	IP judgment	No	2
	Maneuver: Lazy Eight		Maneuver Segment: Second 45° of Turn			
Angle of bank increasing	Att. ind.	Wing bank increasing	Smooth, rough, too fast, too slow, OK	IP judgment	Yes	3
A/S decreasing	A/S ind.	Pointer moving counter-clockwise	Too fast, too slow, OK	IP judgment	Yes	2
Nose approaching horizon	Att. ind. Visual	Moving downward	Smooth, rate	IP judgment	Yes	2
Rate of turn continuing	Visual	Continuing to turn	Stopped, OK, too fast, too slow, smoothly	IP judgment.	No	2

Appendix I *(Continued)*

Activity or Condition	Indicator or Sense	Indications or Stimulus	Decision Factors	Performance Criteria	TV System Capability	Criticality of Performance
At 90° point:						
(a) A/S - 100 kts.	A/S ind.	Pointer at "1"	Low, high, OK	None	Yes	3
(b) Angle of bank 80°-90°	Att. ind., visual	As stated	Between 80-90°, or not	None	Yes	2
(c) Nose position at reference point	Visual	As stated	Short, at, or over	IP judgment	Yes	2
	Maneuver: Lazy Eight		Maneuver Segment: Third 45° of Turn			
Angle of bank decreasing	Att. ind. Visual	Wing bank decreasing	Smoothly, too fast, too slow, OK	IP judgment	Yes	3
A/S increasing	A/S ind.	Pointer moving clockwise	Too fast, too slow, OK	IP judgment	Yes	2
Nose being lowered to maximum pitchdown point	Att. ind. Visual	Moving downward into "black" part of ball	Smoothness, rate	IP judgment	Yes	2
Rate of turn continuing	Visual	Continuing to turn	Stopped, OK, too fast, too slow	IP judgment	No	2
At 135° point:						
(a) A/S at 150 kts. (technique)	A/S ind.	Pointer at 1.5	Fast, slow, on speed	None	Yes	3
(b) Nose at lowest point- approx. 45° down	Att. ind/ Visual	No further lowering	At lowest point	IP judgment	Yes	3
(c) Angle bank at approx. 45°	Att. ind/ Visual	½ way back to level	Degree of bank	IP judgment	Yes	3
(d) CRSE 135° from original start course	Visual	½ way from 90° point to 180° point	Short, at, over	IP judgment	Yes	3
	Maneuver: Lazy Eight		Maneuver Segment: Fourth 45° of Turn			
Angle of bank decreasing	Att. ind/ Visual	Wing bank decreasing to level	Smooth, rough, too fast, too slow, OK	IP judgment	Yes	2
Nose approaching horizon	Att. ind/ Visual	Moving upward	Smoothness, rate	IP judgment	Yes	2
A/S increasing	A/S/ Visual	Pointer moving clockwise	Too fast, too slow OK	IP judgment	Yes	3
Rate of turn continuing	Visual	Continuing to turn	Stopped, too fast, too slow, OK, smoothly	IP judgment	No	2
At 180° point:						
(a) A/S at 200 kts.	A/S ind.	Pointer at "2"	Too fast, too slow, OK	None	Yes	3
(b) Nose on horizon	Att. ind/ Visual	As stated	High, low, on	IP judgment	Yes	3

Activity or Condition	Indicator or Sense	Indications or Stimulus	Decision Factors	Performance Criteria	TV System Capability	Criticality of Performance
(c) CRSE 180° from original start of course	Visual	Opposite to start point	How close	IP judgment	Yes	3
(d) Angle of bank level	Att. ind/ Visual	As stated	Right wing down, left wing down, level	IP judgment	Yes	3
	Maneuver: Lazy Eight		**Maneuver Segment: Final 180° of Turn**			
Repeat of 1st, 2nd, 3rd, and 4th "45° of turn" given on pages B-13, B-14, B-15, and B-16 respectively. Direction of turn opposite to initial direction taken.					Yes	3
	Maneuver: Vertical "S" A		**Maneuver Segment: General**			
Instrument cross-checks	Visual	Late corrections of errors	Corrective action being taken, late, not enough or OK	IP judgment	No	3
Airspeed constant (160 kts.)	A/S ind.	Pointer	Pointer at 1.6	None	Yes	2
Climb (and descent) 1000 ft./min.	V.V. ind.	Pointer	Pointer at "1"	None	Yes	2
Transition from climb to descent, descent to climb, and to level-off	Att. ind.	Pitch angle changing	Smoothly accomplished	IP judgment	Yes	2
	V.V. ind.	Change of position	Moved to appropriate new position	None	Yes	2
Heading constant (relative to start heading)	CRSE ind.	Heading not changing	On start heading or drifting to one side or the other	None	Yes	2
Wing position (level when not correcting for drift in heading)	Att. ind.	Bar remains level	Wings level or banking	IP judgment	Yes	2
Coordination of pitch and power control	V.V. ind. Att. ind. RPM ind. A/S ind.	Power holding constant or changing to maintain A/S as pitch angle or rate of ascent (or descent) changes	A/S constant and rate of climb or descent holding constant except through transitions	IP judgment	No	2
Lead points for transitions	V.V. ind Altimeter Att. ind.	Starts to move and change position	Right time, too late, or too early	None IP judgment	Yes	2

Appendix I *(Continued)*

Activity or Condition	Indicator or Sense	Indications or Stimulus	Decision Factors	Performance Criteria	TV System Capability	Criticality of Performance
Instrument cross-checks	Visual	Late corrections of errors	Corrective action being taken, late, not enough, or OK	IP judgment	No	3
Airspeed constant (160 kts.)	A/S ind.	Pointer	Pointer at 1.6	None	Yes	2
Climb (and descent) 1000 ft./min.	V.V. Ind.	Pointer	Pointer at 1	None	Yes	2
Transition from climb to descent, descent to climb, and to level-off	Att. ind. V.V. ind.	Pitch angle changing Change of position	Smoothly accomplished Moved to appropriate new position	IP judgment None	Yes Yes	2 2
Trainsition from one direction of turn to another direction	CRSE ind. Att. ind.	Changing direction of movement from one direction to another or to level flight altitude	Smoothly accomplished at same time transition to climb or descent is being done	IP judgment	Yes	2
Aircraft heading - changing during maneuver	CRSE ind. Att. ind.	Card rotating	Heading changing	None	Yes	1
Angle of bank (30°)	Att. ind.	Pointer on bar matched with 30° marker	Short, on, or over	None	Yes	2
Coordination of pitch and power control	V.V. ind. Att. ind. RPM ind. A/S ind.	Power holding constant or changing to maintain A/S as pitch angle or rate of ascent (or descent) changes	A/S constant and rate of climb or descent holding constant except through transitions	IP judgment	No	3
Lead points for transitions	V.V. ind. Altimeter Att. ind.	Starts to move and change positions	Right time, too early, or too late	None IP judgment	Yes	2

APPENDIX II. FINAL PILOT PERFORMANCE RATING SCALES

APPENDIX II. FINAL PILOT PERFORMANCE RATING SCALES

PILOT PERFORMANCE REFERENCE SCALE

IP Name ______________________

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Maneuver ______ U F G E
Example ______

Appendix II *(Continued)*

PILOT PERFORMANCE REFERENCE SCALE

IP Name ______________________ Maneuver *FINAL TURN TO LANDING*

Example ______________ Page *1* of *4*

Performance Element	"Ideal"	Not Observed	Performance Rating			
1A. Airspeed at start	Minimum of 120 kts. prior to turn	☐	Less than 110 kts. or over 135 kts.	No less than 110 kts. or over 135 kts.	No less than 115 kts. or over 135 kts.	No less than 118 kts. or over 135 kts.
1B. Altitude at start	1000 feet above terrain	☐	Over ± 200 feet from ideal.	Within ± 100 - 200 feet from ideal.	Within ± 50 - 100 feet from ideal.	Within ± 50 feet of ideal.
1C. Attitude at start	Level	☐	Abnormal attitude at F.T. point.	Neither settled nor making correction.	Not settled but correcting nicely.	Settled in level flight attitude.
2B. Control of airspeed	110 kts. final turn, 100 kts. final approach, 75 - 80 kts. at touchdown	☐	Over ± 10 kts. from proper pattern airspeeds.	Within ± 10 kts. of proper pattern airspeeds.	Within ± 5 kts. of proper pattern airspeeds.	Within ± 2 kts. of proper pattern airspeeds.
2C. Control of altitude	300 feet above field level at final approach point	☐	Over ± 100 feet of 300 feet above field level.	Within ± 100 feet of 300 feet above field level.	Within ± 50 feet of 300 feet above field level.	Within ± 25 feet of 300 feet above field level.

Appendix II *(Continued)*

Maneuver *FINAL TURN TO LANDING*

Page *2* of *4*

Performance Element	"Ideal"	Not Observed	Performance Rating			
2D. Control of heading	Heading for runway on roll out onto final approach with x-wind correction made	☐	Acute angling final approach or x-wind correction (or lack of) creates unsafe conditions. Used wrong x-wind landing technique. Overshot excessively.	Excessive angled (but correctible) approach. Some x-wind correction made but not adequate to effect safe landing without applying other types of corrective actions.	Slight angling approach or overshot slightly and came back. Should have made slightly more (or less) of a x-wind correction.	Centerline of a/c headed directly down runway with x-wind correction made.
2E. Control of pitch angle	Approximate angle to correlate with airspeed control	☐	Does not understand use of pitch control during maneuver or allowed a/c to attain unsafe pitch attitude.	Excessive changes to pitch made. Pitch control erratic.	Used pitch to control airspeed but over-corrected slightly.	Used pitch effectively to control airspeed.
2G. Control of angle of bank	Maximum of 45° and established such that turn be completed with minimal change	☐	Exceeded 45° angle of bank during a correction maneuver. Wide on excessive number of unnecessary changes. Angle bank varied by more than 10°.	45° angle of bank exceeded unintentionally but corrected promptly. Made changes that indicated indecision. Angle of bank too shallow (approx. 10°). Angle bank varied 5 - 10°.	Introduced a small number of incorrect bank changes in turn which were safe but not necessary. Angle bank varied less than ±5°.	Maximum angle of bank throughout maneuver was no more than 45° with small number of correct changes to angle of bank.

Appendix II *(Continued)*

Maneuver *FINAL TURN TO LANDING*

Page *3* of *4*

Performance Element	"Ideal"	Not Observed	Performance Rating			
2J. Control of rate of ascent/descent	Constant and controlled	☐	Excessive rate of descent. Large correction required to control rate of descent.	Erratic rate of descent. Low or too steep final approach.	Better control over rate of descent could have been exercised. Flat or steep final approach.	Constant, smooth, and controlled rate of descent. On glide slope.
4. Error correction	Prompt and positive	☐	Errors ignored or overcorrected for. Failed to remove corrective action. Very abrupt corrections made.	Slow to recognize errors or initial response improper. Late to remove corrective action. Corrective action not made too smoothly.	Errors noted but slow to respond properly within an acceptable time frame. Could have made corrections more smoothly.	Errors noted and proper corrections made promptly and smoothly.
5. Transitioning	Smooth roundout (a flare)	☐	Erratic pitch changes through transition, "fishing" for ground, or loss of control of rate of descent or airspeed.	Abrupt transition. Ballooned. Incomplete pitch change. Excessive use of power control.	Slightly late (or early) starting transition. Small number of unnecessary changes made to pitch angle. Unnecessary changes made to power control.	Smooth pitch changes started at right distance above ground level with a gradual decrease in a/s to touchdown.

Maneuver *FINAL TURN TO LANDING*

Page 4 of 4

Performance Element	"Ideal"	Not Observed	Performance Rating			
9. Use of ground reference points or lines	Final approach started, ½ mile from end of runway flying a/c toward runway aimpoint	☐	Did not use reference point or lines available. Does not understand use of aimpoint. Cannot visualize where ½ mile point for start of final approach is.	Aimpoint changed several times or pointed too far down (or short of) runway.	Slightly long (or short) final approach. Point of aim slightly long (or short).	Used correct reference point on roll-out to final approach and aimpoint.
OTHER						
OTHER						
OVERALL						

Appendix II *(Continued)*

PILOT PERFORMANCE REFERENCE SCALE

IP Name ____________________

Example ____________________

Maneuver *VERTICAL S "A"*

Page *1* of *3*

Performance Element	"Ideal"	Not Observed	Performance Rating			
1A. Airspeed set up on entry	160 kts.	☐	Over ± 10 kts.	Within ± 10 kts.	Within ± 5 kts.	Within ± 2 kts.
1B. Altitude set up on entry	On altitude assigned	☐	Over ± 50 feet of assigned altitude.	Within ± 50 feet of assigned altitude.	Within ± 25 feet of assigned altitude.	Within ± 10 feet of assigned altitude.
1D. Heading set up on entry	On assigned heading	☐	Over ± 10° from assigned heading.	Within ± 10° of assigned heading.	Within ± 5° of assigned heading.	Within ± 2° of heading assigned by IP.
1G. Trends on entry of: 1) Airspeed 2) Altitude 3) Attitude 4) Heading	 Constant Holding Steady Steady	☐	Unacceptable a/s, altitude, attitude, or heading changes at start of maneuver.	Two or more of sub-elements (airspeed, altitude, attitude, or heading) changing slightly.	One of four sub-elements (airspeed, altitude, attitude, or heading) changing slightly.	No adverse trend of a/s, altitude, or heading and pitch angle set for level flight.

Appendix II *(Continued)*

Maneuver *VERTICAL S "A"*

Page 2 of 3

Performance Element	"Ideal"	Not Observed	Performance Rating			
2B. Control of airspeed (except during transition, see element #5)	Constant	☐	Over ± 15 kts.	Within ± 15 kts.	Within ± 10 kts.	Within ± 5 kts.
2D. Control of heading	Constant	☐	Over ± 10 degrees from assigned heading.	Within ± 10 degrees from assigned heading.	Within ± 5 degrees of assigned heading.	Within ± 3 degrees of assigned heading.
2E. Control of pitch angle (except during transition, see element #5)	Sufficient to control rates of ascent/descent and a/s	☐	Too great and too many pitch changes in search for proper angle.	Uncertain as to what pitch angle to use but does not vary too excessively from proper angle.	Slightly overcontrolled and within ¼ bar width of knowing proper pitch angle to use.	Knows proper pitch angles to use, sets pitch, and varies angle as required to maintain a/s and rate of ascent/descent.
2J. Control of rate of ascent/descent	1000 feet per min.	☐	Varies more than ± 300 feet/min. from standard or varies excessively within prescribed limits of 700 to 1300 ft/min.	Held steady between 700 and 1300 ft/min.	Held steady between 800 and 1200 ft./min.	Held steady between 900 and 1100 ft/min.

Appendix II *(Continued)*

Maneuver *VERTICAL S "A"*

Page 3 of 3

Performance Element	"Ideal"	Not Observed	Performance Rating			
4. Error corrections	Prompt and positive	☐	Errors ignored or over-corrected for. Failed to remove corrective action. Very abrupt corrections made.	Slow to recognize errors and initial response improper. Late to remove corrective action. Not made too smoothly.	Errors noted but slow to respond properly within an acceptable time frame. Could have made corrections more smoothly.	Errors noted and proper corrections made promptly and smoothly.
5. Transitioning	Smooth and at proper a/s and altitude	☐	Rough (or sudden) transition with loss or gain of a/s or altitude control (over ± 10 kts. or ± 100 feet).	Early (or late) reversal. Altitude within ± 100 feet through reversal. Airspeed within ± 10 kts. Reversal too abrupt.	Airspeed with ± 5 kts. through transition. Altitude within ± 50 feet through reversal. Slightly late in changing pitch attitude.	Smooth with a/s holding constant with bottom (or top) points of "S" at altitudes directed by IP (within ± 25 feet).
OTHER						
OVERALL						

PILOT PERFORMANCE REFERENCE SCALE

IP Name ____________ Maneuver *LAZY 8*

Example ____________ Page *1* of *3*

Performance Element	"Ideal"	Not Observed	Performance Rating			
1A. Airspeed at entry	200 kts.	☐	Over ± 10 kts from ideal.	Within ± 10 kts.	Within ± 5 kts.	Within ± 3 kts. from ideal.
1C. Attitude at entry	Level	☐	Started maneuver in an abnormal flight attitude.	Started maneuver in an unsettled flight attitude.	Slightly out of level flight attitude.	In level flight attitude.
1E. Positioning of a/c at entry	Aligned with 90° reference point	☐	90° reference point neither identified nor used to align a/c.	90° reference point established but a/c not properly aligned.	90° reference point established but a/c slightly misaligned.	90° reference point established and a/c properly aligned to it.
2B. Control of airspeed	Approx. 150 kts. @ 45 and 225°; Approx. 100 kts. @ 90 and 270°; Approx. 150 kts. @ 135 and 315°; 200 kts. @ 360°	☐	A/S off more than ± 15 kts.	A/S within ± 15 kts.	A/S within ± 10 kts.	A/S within ± 5 kts.

Maneuver *LAZY 8*

Performance Element	"Ideal"	Not Observed	Performance Rating			
2D. Control of heading	Nose position on or passing through horizon at 90°, 180°, 270°, and 360° positions from known start reference point	☐	Over ±.15° from the desired heading at checkpoints.	Within ± 15° of desired heading at checkpoints.	Within ± 10° of desired heading at checkpoints.	Within ± 5° of desired heading at checkpoints.
2E. Control of pitch angle	Nose at highest point: 45° and 222° Nose at lowest point: 135° and 315° Nose at horizon: 90°, 180°, 270°, 360°	☐	Nose position excessively high/low at any one of the specified checkpoints.	Nose position low or high at more than two of the checkpoints.	Nose position slightly low or high at no more than two of the checkpoints.	Nose position as required at intermediate checkpoints.
2F. Control of rate of roll	Slow and constant	☐	Too fast or with frequent starts and stops throughout.	Erratic rates at times throughout performance.	Erratic no more than twice during performance.	Slow and constant throughout performance.

Appendix II *(Continued)*

Maneuver *LAZY 8*

Page 3 of 3

Performance Element	"Ideal"	Not Observed	Performance Rating			
2G. Control of angle of bank	45° @ 45° and 135, 225, and 315° position; 90-90° @ 90 and 270° position	☐	Over ±15° from proper angle of bank at given checkpoints.	Within ± 15° of ideal.	Within ± 10° of ideal.	Within ± 5° of ideal.
2I. Control of rate of pitch change	Slow and at a constant rate	☐	Too fast or with frequent starts and stops throughout changes too jerky or abrupt.	Erratic changes at times throughout maneuver.	Erratic changes no more than twice during performance.	Slow and at a constant rate coordinated well with angle of bank to control rate of turn.
4. Error correction	Prompt and positive	☐	Errors ignored or over corrected for. Failed to remove correction action. Very abrupt corrections made.	Slow to recognize errors or initial response improper. Late to remove corrective action. Corrective action not made too smoothly.	Errors noted but slow to respond properly within an acceptable time frame. Could have made corrections more smoothly.	Errors noted and proper corrections made promptly and smoothly.
OVERALL						

APPENDIX III. SUGGESTED ADDITIONAL ANALYSES

This appendix is concerned with the identification of areas of data analysis considered pertinent to continued efforts in the development of pilot performance reference scales. As with most studies of this type where there is a great deal of raw data to be processed, there are many statistical approaches which can be pursued. In addition, the results of one analysis usually suggest several other potential analyses. Of course, not all possible analyses are useful; in this section every attempt has been made to include only those analyses thought to be the most useful to further developments of pilot performance reference scales or utilization of an AVRS during pilot training or retraining. The areas of data analyses are concerned primarily with a greater in-depth analysis consistent with a determination as to why certain inconsistencies exist in the data already presented in the report and with additional data analyses thought to be relevant to the overall objectives of the Air Force's flying training programs. Specific objectives of the analyses are (*a*) to provide additional insights and guidance into pilot performance scales prior to further scale development efforts; and (*b*) to provide a greater understanding of the factors which impact on decisions or judgments made by instructor pilots relevant to the assignment of level of performances.

There were many inconsistencies in the data presented in the main body of the data analysis. An in-depth analysis of these inconsistencies would provide a greater understanding of the actual usage of the pilot performance reference scales developed in this study and greater insights into the reasons why those scales were not as reliable as envisioned in order to develop rationales for their improvements (ergo, reliability). Such analyses must be accomplished in conjunction with a study of the video replay of the maneuver examples involved in the inconsistency. The following suggested analyses are considered to be minimal:

1. Elements marked observed-not observed: The data suggest that the difference in opinions as to what elements could and could not be graded should be investigated in greater depth than has already been accomplished (and reported). For example, in a particular series of observations, only 9 of the 23 instructors who viewed the maneuver graded an element. Why? Or why did 14 instructors *not* grade the element? All such differences should be considered in the additional data analysis.

2. Elements with high variances or a high spread of assigned performance levels should be analyzed in depth in order to ascertain what the primary problems might be. There are 19 such instances of high element variance which represent the initial concentration of effort.

3. Several inconsistencies require a greater in-depth analysis in consonance with the stated objectives of this section. Investigations should be undertaken to determine why certain maneuver examples contain both high and low variability elements (example 2 of the Final Turn to Landing, examples 2, 4, and 8 of the Vertical S "A," and example 7 of the Lazy Eight).

Another area for study concerns examples 1 and 5 (identical performances), which appear to present a unique problem to the scales. The only consistency appears to be with elements 1C and 2D which show high variabilities in both examples. Element 2E is most unique in that it is highly variable in example 1 and is among those with the lowest variability in example 5.

The following elements are listed among those with the highest relative variance *and* the lowest relative variance: elements 2E, 2G, and 4 of the Final Turn to Landing; elements 1C, 2G, and 2I of the Lazy Eight; and element 2E of the Vertical S "A". The fact that the levels of performance assigned to these elements were determined from their demonstration within different examples of maneuvers (i.e., within different contexts) enhances the expected intelligence gain from a greater in-depth analysis.

4. Examples 5 and 8 of the Final Turn to Landing are on opposite ends of the scale, and these two examples have the lowest variability of all the Final Turn to Landing examples. This situation exemplifies an ideal result if the scales were found to have been reliable. (In fact, such a situation is a measure of scale reliability.) Unfortunately, it is known that the scales are not that reliable. In conjunction with an analysis of the apparent reliability of the scales in judging the two Final Turn to Landing examples, should be an analysis of two examples of the Vertical S "A" which are on opposite ends of the scale (examples 5 and 8), further apart, but with greater degree of variability. Examples 5, 6, and 1 (taken as a unit) and example 8 of the Lazy Eight should also be included in the analysis, but to a lesser extent since the difference between the two performance levels are less than the difference between the extremes in the other two maneuvers.

5. In the Vertical S "A", a difference of 1.28 was shown in the variances of the overall grades assigned to examples 3 and 9–identical performances. This is the largest difference between all duplicate examples. Yet the Vertical S "A" has been found not only to be the easiest maneuver to make judgments about, but that, in general, the variabilities of these judgments are the smallest. It was also noted that examples 2 and 8 of the Vertical S "A" had zero difference in variability. Again, an in-depth analysis should provide necessary pertinent inputs into the overall development of reliable scales.

6. It is felt that the performance elements concerned with the set-up of a maneuver prior to maneuver performance (elements 1A through 1G) received too great an emphasis during the data analysis. There is no doubt that proper set-up is a prerequisite to good maneuver performance and a skill to be learned. However, scaling each part of the overall set-up requirements (e.g., airspeed, altitude) and treating each part as an equal to elements concerned with maneuver performance is not considered to be appropriate. Therefore, it is suggested that the set-up elements and associated data be combined as a single element and that selected analyses similar to these already reported be repeated. The results from the two sets of analysis should then be compared as to impact on scale reliability and the scales themselves.

7. The basic data bank contains information as to what actual syllabus mission each of the 25 (30 minus duplicates) maneuver examples represented. The name of the student is also available. It is suggested that the levels of performance assigned by individual instructors and the instructors as a group be analyzed in conjunction with the syllabus mission flown relative to the overall syllabus. In the event the data analysis so suggests, the name of the student (and, thus, his flight training record) will provide additional data to the data analysis effort.

8. The scale developed in this study was a 10-point scale. The data analysis was, therefore, oriented towards establishing results relevant to that scale's discriminatory properties as well as its reliability. Although some analysis was accomplished, it is suggested that the raw data be translated into 4-point and 7-point scales (method for so doing previously explained). Selected analyses similar to those which used the 10-point scale should then be repeated and the results compared and analyzed as they affect the number of discriminations that can reliably be made and the scale reliability.

9. Further insights into the number of discriminations an instructor can make for each maneuver may be available from the basic data. At least one additional analysis seems to be indicated–and that is determining the number of discriminations that were made by the test instructors for each performance element both within the three types of maneuvers and across maneuvers. The same analysis should be accomplished, if the data are translated into 7- and 4-point scales, and a comparative analysis of the results made.

10. The data obtained during this study contain insights into the process wherein evaluations of performance elements are integrated into a composite, overall evaluation of a maneuver example. At present, there is little doubt that this process, or formula, used by an instructor to arrive at an overall grade for a maneuver remains a mystery–and that there are probably as many formulas as there are instructors. Nonetheless, it is reasonable to assume that such formulas exist and that such formulas (including those which have different applications, since, for instance, the needs of an instructor differ in detail from the needs of upper management) are basic to the future requirements of automated systems. It is suggested that this area of analysis should be accomplished as an initial step in the development of a formal mathematical model of how instructors grade, or should grade, maneuvers.

These suggestions do not constitute the entire range of data analysis which could be conducted with the available data, nor are they meant to be restrictive. However, they do reflect the conviction of the study group that the concepts and principles used to develop the scales in this study were valid and that additional efforts to gain further insights which would broaden the base upon which scale revisions, refinements, and expansions could be effected are warranted.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* American Institutes for Research 135 North Bellefield Avenue Pittsburgh, Pennsylvania 15213	2a. REPORT SECURITY CLASSIFICATION 2b. GROUP

3. REPORT TITLE

THE DEVELOPMENT. TEST, AND EVALUATION OF THREE PILOT PERFORMANCE REFERENCE SCALES

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

5. AUTHOR(S) *(First name, middle initial, last name)*

Walter R. Horner
Thomas L. Radinsky
Robert Fitzpatrick

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
August 1970	62	4

8a. CONTRACT OR GRANT NO. F33615-69-C-1366	9a. ORIGINATOR'S REPORT NUMBER(S) AFHRL-TR-70-22
b. PROJECT NO. 1710	
c. Task No. 171003	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d. Work Unit No. 171003024	

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Flying Training Division Williams Air Force Base, Arizona

13. ABSTRACT

This report describes the results of a study to develop pilot performance reference scales based upon audio-video recordings of in-flight performances of students undergoing T-37 undergraduate pilot training. The study included scale development as well as the test and evaluation of each scale. All the maneuvers contained on the in-flight recordings were analyzed, and constituent performance elements observable on the video replay were identified. Three maneuvers, Final Turn to Landing, Vertical S "A," and Lazy Eight, were selected for the final scaling effort. Ten performance elements each were identified for the Lazy Eight and Vertical S "A" maneuvers, and twelve elements for the Final Turn to Landing. A performance reference scale was developed for each maneuver. Each scale consisted of a series of subscales for rating performance on each of the elements of the maneuver and an additional subscale for rating the overall performance of the maneuver. Although some elements were common to more than one maneuver, the rating scales for these elements were tailored in each case to the maneuver involved. Each subscale consisted of a ten-point rating line (a row of ten boxes) representing the full range of performance from "unsatisfactory" to "excellent" and, beneath, four graded verbalizations describing different levels of performance. No verbalizations were presented, however, with the subscale used for rating overall performance. Final versions of the scales were subjected to a test and evaluation through their utilization by experienced instructor pilots. These pilots assigned levels of performance based upon what they observed on video replays of selected maneuver examples. The results showed the overall reliability of scales for the three maneuvers was high but that the majority of the individual element scales were of a relatively low to medium degree of reliability. The results are believed to justify more in-depth analysis of the data and continued development efforts to refine and increase the scope of scale application.

DD FORM 1 NOV 65 1473

Unclassified

Security Classification

Unclassified
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
pilot performance reference scales						
pilot performance scales test and evaluation						
pilot performance measures						
pilot performance scale development						
Undergraduate Pilot Training						
Audio-Video Recording System						

Unclassified
Security Classification