NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. 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.
Improving Flight Proficiency Evaluation in Army Helicopter Pilot Training

by

George D. Greer, Jr., Wayne D. Smith, and Capt Jimmy L. Hatfield

U.S. Army Aviation Human Research Unit
Fort Rucker, Alabama

Under the Technical Supervision of

The George Washington University
HUMAN RESOURCES RESEARCH OFFICE
operating under contract with
THE DEPARTMENT OF THE ARMY
CRD/J

SUBJECT: HumRRO Technical Report, "Improving Flight Proficiency Evaluation in Army Helicopter Pilot Training" (LIFT II)

TO: COLONEL
ARMY CIVIL TECH INF AGENCY
ARLINGTON HALL STATION
ARLINGTON, VA 22203
ATTN: TIPR

1. The attached report is for your information and retention.

2. The objective of this study was to develop a more reliable system of evaluating helicopter pilots' flight performance by putting emphasis on standardized and objective measures, which would also provide a diagnostic record of student performance.

3. This report is considered to be of primary interest to those organizations and agencies concerned with helicopter pilot training. The Pilot Performance Description Records (PPDR) have proved useful in administering check rides in primary helicopter training. The system provides a means of diagnosing specific sources of a student's end-of-phase deficiencies, by the detailed recording of his flight performance.

4. The report serves to standardize pilot proficiency evaluation through reducing subjective differences in scoring procedures.

FOR THE CHIEF OF RESEARCH AND DEVELOPMENT:

[Signature]

GEOFFREY BATEMAN, JR
Colonel, USA
Chief, Human Factors Research Div
IMPROVING FLIGHT PROFICIENCY EVALUATION
IN ARMY HELICOPTER PILOT TRAINING

by

George D. Graer, Jr., Wayne D. Smith,
and Capt Jimmy L. Hatfield

Approved:

J. DANIEL LYONS
Director of Research
U.S. Army Aviation Human Research Unit
Fort Rucker, Alabama

ARNE H. ELIASSEN
Lt Col, Inf, Unit Chief
U.S. Army Aviation Human Research Unit
Fort Rucker, Alabama

Meredith P. Crawford
Director
Human Resources Research Office

The George Washington University
HUMAN RESOURCES RESEARCH OFFICE
operating under contract with
THE DEPARTMENT OF THE ARMY

Technical Report 77
May 1962
COMPOSITION OF RESEARCH TEAM

Subtask LIFT II was initiated while Dr. George D. Greer, Jr., was Director of Research at the U.S. Army Aviation Human Research Unit. Dr. J. Daniel Lyons was Director of Research during the preparation of this report.

Dr. Greer was Task Leader during the planning, data gathering, data analysis, and early report writing stages. Dr. Carroll M. Colgan was Task Leader during the final report writing stage.

Dr. Greer, Mr. J. Albert Southern, Capt. Jimmy L. Hatfield, and Mr. Wayne D. Smith planned the study and collected the data. Mr. John O. Duffy and Captain Hatfield prepared the report in the final stages.

The Human Resources Research Office is a nongovernmental agency of The George Washington University, operating under contract with the Department of the Army (DA 44-186-ARO-2). HumRRO's mission, stated by AR 70-8, is to conduct studies and research in the fields of training, motivation, leadership, and man-weapons system analysis.

Research is reported by HumRRO in publications of several types.
1. Technical Reports are prepared at the completion of a research Task or major portion thereof. They are designed specifically for a military audience and convey recommendations for Army action.
2. Research Reports may be prepared at any time during a Task. They are designed primarily for a research audience but may be of interest to a military audience. They report research findings of interest and value to the scientific community and do not recommend Army action.
3. Research Memoranda may be prepared at any time and need not be directly associated with a particular research Task. They report findings that may be of interest to a research or military audience or to both. They do not recommend Army action.
4. Consulting Reports are prepared following completion of a specifically requested consulting action under HumRRO's Technical Advisory Services. They are designed for a specific military audience and usually convey recommendations for Army action.
5. Research Bulletins are prepared as nontechnical summaries of one or more research Tasks or as reports of other HumRRO activities. They are intended primarily for a military audience and do not present recommendations for Army action. Their distribution usually includes agencies and individuals conducting research, and the general public.

Technical Reports and Research Bulletins may be requested from the Director's Office, which also issues a complete bibliography. Other publications may be obtained from the Director of Research of the originating Unit or Division.
PROBLEM

Improvement in the efficiency of the Army's primary helicopter training program depends to a large degree on the reliability of flight training evaluation. The traditional flight check has consisted of an evaluation of the flight by the check pilot not on the basis of a uniform series of maneuvers and measures, but on the basis of his personal specifications. It seemed probable that the unreliability of the traditional method of evaluation, which had been repeatedly demonstrated, was due primarily to this lack of standardization. This study was initiated to develop a more reliable system of evaluating helicopter pilots' flight performance, by emphasizing standardized and objective measures which also provide a diagnostic record of student performance.

PROCEDURE

Training grades and check flight grades were analyzed for Army helicopter pilots at both the U.S. Army Aviation School (USAAVNS), Fort Rucker, Ala., in 1956-57 and at the U.S. Army Primary Helicopter School (USAPHS), Camp Wolters, Tex., in 1957. In general, the relationships between the training grades and the corresponding test scores proved to be little better than zero. In another analysis, it was found that ratings of students' flight performance reflected the standards of evaluation applied by individual check pilots more than they did the students' flying skill.

The first step in the development of a more effective method of flight evaluation was an analysis of the light helicopter training program content into fundamental training maneuvers and maneuver components. Simple scales of several types were developed for use by the check pilot in recording the students' performance on each of these components. Where it was possible, direct instrument observations were recorded. However, many evaluations are necessarily based on individual judgment, to a lesser or greater degree; where judgments were required, the performance being evaluated was defined as specifically as possible at each point on the scale in order to narrow the range of personal interpretation in assigning ratings.

The next step was the development of a format for an Intermediate and an Advanced Pilot Performance Description Record (PPDR). Each PPDR was based on a standard ride, that is, the same maneuvers flown in the same sequence. The scales included as PPDR items were those judged to be most critical to successful performance in each maneuver. The number of scales that an expert check pilot could safely observe and record during a check ride was used as the basis for setting the total number of PPDR items (most items were recorded as the operation was being accomplished, but on operations that are considered hazardous, recording was delayed until completion of the dangerous portion).

The PPDR's were then tested by administering check rides to 40 Intermediate and 35 Advanced students at the Primary Helicopter School (Camp Wolters) in 1957. Each
student was administered one ride by a LIFT research staff pilot and one ride by a military check pilot assigned to USAPHS.

The PPDR's were revised on the basis of experience in the first administration, and the revised PPDR's were evaluated in 1958. Check pilots were given one week of training in the use of the PPDR system, with emphasis placed on identification and reduction of check pilot differences in scoring standards. Two successive rides, each with a different USAPHS check pilot, were given to 50 Intermediate and 50 Advanced students.

Several approaches to summarizing the data on student performance which the PPDR check rides provided were explored. One was simply to total the number of errors recorded on the PPDR in a check flight. A second weighted items according to difficulty. In another approach ("error pattern-weighted") the pilot rated the student's over-all performance on a maneuver segment, taking into consideration not only errors but their sequence and combination; these segment ratings were weighted according to difficulty and importance of the maneuver. Finally, the check pilot assigned an over-all judgmental rating, based upon a review of the detailed PPDR record of the student's performance, and comparable to the "traditional" score.

FINDINGS

1. Improved reliability of flight proficiency evaluation resulted from the use of the PPDR system.
2. The PPDR system provided a means of diagnosing specific sources of a student's end-of-phase deficiencies, by recording, in detail, the student's performance on his flight check rides.
3. Check pilots who were completely familiar with the PPDR were reliably more similar in their evaluation of proficiency than were check pilots who were only oriented to the PPDR.

CONCLUSIONS

1. The PPDR flight evaluation system can provide an evaluation of helicopter students' flight performance that is at an acceptable level of reliability. The resulting diagnostic data provide the basis for determining flight deficiencies of individual students and for maintaining uniform standards for both instruction and evaluation.
2. To maximize the effectiveness of the PPDR system, it is necessary that personnel serving as check pilots be trained in the concepts, objectives, and techniques of the system, and in administering and scoring the PPDR's.
RECOMMENDATION

It is recommended that the PPDR system be adopted in primary helicopter training and further developed. Special emphasis should be given to (a) training check pilots thoroughly in the PPDR system, especially on scoring PPDR's, and (b) developing a system for processing the diagnostic data both for debriefing students and for maintaining standards of instruction and check pilots' evaluation.

A quality control program based upon a revised version of the Pilot Performance Description Record system has been devised and is being implemented at the U.S. Army Primary Helicopter School at Camp Wolters. Experience has been obtained in using the PPDR check ride at Camp Wolters over the past two years, both in research and in operation. No safety problems have arisen during administration of the program, and it has been generally well accepted by check pilots. A report on this program, designated as Subtask LIFT IV, is in preparation.
CONTENTS

SUMMARY AND RECOMMENDATIONS

Problem.................................................................................. iii
Procedure.............................................................................. iii
Findings.................................................................................. iv
Conclusions............................................................................ iv
Recommendation....................................................................... v

DESCRIPTION OF THE RESEARCH

Chapter

1 The Flight Proficiency Evaluation System
   Introduction........................................................................... 3
   The Traditional Flight Check System.................................. 4
   The Flight Check..................................................................... 4
   The Training Grade.............................................................. 7
   Reliability of the Traditional Flight
   Check System........................................................................ 7
   The Flight Check System Used in Army
   Helicopter Training............................................................. 8
   Reliability of the Army Flight Check System....................... 8
   Variation in Army Check Pilots' Evaluation Standards............. 9
   The Flight Proficiency Evaluation Process......................... 10

2 HumRRO Research on Flight Proficiency Evaluation
   Development of the Pilot Performance
   Description Records (PPDR's)............................................ 14
   Tryout of the 1957 Version of the PPDR.............................. 15
   Procedure.............................................................................. 15
   Scoring of the PPDR's......................................................... 18
   Results of the 1957 PPDR Tryout........................................ 18
   Tryout of the 1958 Version of the PPDR.............................. 20
   Procedure.............................................................................. 20
   Results of the 1958 PPDR Tryout........................................ 21
Chapter

3 Application of the PPDR System

Characteristics of the PPDR System ........................................ 27
Scoring Standard PPDR's as Part of the 1958 Training Program .......... 27
Classroom Scoring Agreement and Ride/Ride Relationships ............... 28
Development of the PPDR System for Operational Use .................... 29

References ................................................................................. 33

Appendices

A Prior Research on the Use of Objective Measures in Flight Performance Evaluation .................................................. 35
B Relationships Between Check Grades and Training Grades in the Army's Fixed Wing Training Program ......................... 38
C Attempts to Study Interobserver Relationships ......................... 40
D Variation Among Check Pilots in Scoring the 1958 PPDR's ............. 42

Figures

1 Sample Training Grade Slip .................................................. 5
2 Sample Performance Record Used in H-23 Primary Helicopter Training ................................................................. 6
3 Sample Record Sheet From 1957 Version of the Intermediate (Primary) Pilot Performance Description Record ............. 16
4 Sample Record Sheet From 1957 Version of the Advanced (Basic) Pilot Performance Description Record .................. 17
5 Sample Record Sheet From 1958 Version of the Intermediate PPDR ................................................................. 22
6 Sample Record Sheet From 1958 Version of the Advanced PPDR ................................................................. 23

Tables

1 Correlations for Primary Rotary Wing Flight Training Grades and Check Grades at Fort Rucker, 1956-1957 ...................... 8
2 Means and Ranges of Rotary Wing Check Grades Given by Check Pilots, Fort Rucker, 1956-1957 ................................. 9
3 Estimates of the Percentage Contribution Intermediate Maneuvers Should Make to Total Score .................................... 12
4 Estimates of the Percentage Contribution Intermediate Maneuvers Should Make to Total Score .................................... 13
5 Correlations Between Rides, 1957 ........................................... 18
Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Mean Percentage of Errors Recorded for Selected Items and Mean Over-All Scores for the Intermediate PPDR's, 1957.</td>
<td>19</td>
</tr>
<tr>
<td>7</td>
<td>Mean Percentage of Errors Recorded for Selected Items and Mean Over-All Scores for the Advanced PPDR's, 1957</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>Correlations Between Rides, 1958</td>
<td>24</td>
</tr>
<tr>
<td>9</td>
<td>Correlations Between Training Grades and Check Grades (1957) and Training Grades and PPDR Scores (1958)</td>
<td>25</td>
</tr>
<tr>
<td>10</td>
<td>PPDR Flight Check Scoring Agreement for Check Pilot Pairs Compared With Classroom Standard PPDR Scoring Agreement</td>
<td>29</td>
</tr>
<tr>
<td>B-1</td>
<td>Correlations of Fixed Wing Check and Training Grades, Camp Gary, 1957-1958</td>
<td>38</td>
</tr>
<tr>
<td>B-2</td>
<td>Correlations of Fixed Wing Check and Training Grades, Fort Rucker, 1957-1958</td>
<td>39</td>
</tr>
<tr>
<td>D-1</td>
<td>Means and Standard Deviations of Percentages of Errors Scored by Check Pilots on Selected PPDR Items and of PPDR-Derived Scores, 1955</td>
<td>42</td>
</tr>
</tbody>
</table>
DESCRIPTION
OF THE RESEARCH

IMPROVING FLIGHT PROFICIENCY EVALUATION
IN ARMY HELICOPTER PILOT TRAINING
Chapter 1

THE FLIGHT PROFICIENCY EVALUATION SYSTEM

INTRODUCTION

How reliable and how analytic is the traditional flight check as a measure of flight proficiency? This question has been of particular significance to flight training administrators charged with the responsibility for continually trying to improve the flight check system used in both fixed and rotary wing flight training in the Army.

The flight check is a measure of student progress given at the end of each phase of pilot training, administered by an experienced check pilot who acts as observer and safety pilot. Under the traditional procedure, which was studied during early stages of LIFT research (1956-57), the student flies a sample, selected by the check pilot, of the flight maneuvers taught in the preceding phase. The check pilot grades the student on the basis of his personal evaluation of the student's performance, both on circumscribed aspects of the flight and on the over-all performance.

The nonstandardized nature of the traditional flight check, in which each check pilot follows personal standards in grading, has been criticized in the past as a source of unreliability in evaluating flight performance. The nature of in-flight proficiency evaluation makes it impossible to eliminate variations in the test due to different aircraft, shifting weather conditions, and transient check pilot or student moods; the evaluation process itself is, of necessity, complex. However, many of the causes of variability—those resulting from different test components and different check pilot standards—are subject to control. To the extent that the traditional grading system is unnecessarily unreliable, flight proficiency measurement will be less valuable as a means of identifying the weaknesses and strengths of students and instructors and for pinpointing shortcomings in the Program of Instruction (POI). As a result, flight training will be less efficient than it can be (in terms of amount of training per dollar spent).

The argument between advocates of "subjectivism" and of "objectivism" in flight proficiency measurement has been going on in research and training circles for years. Lt. J.M. Brown of the Royal Canadian Air Force (4) says:

To understand the origin of this controversy one has to go back about twenty-five years in the short history of aviation to the time when flying itself was subjective. At that time the success of a pilot depended upon how well he could fly his aircraft by feel, or quite literally, "by the
seat of his pants." Flight instruments and radio information was meager
and, with the exception of a crude compass, navigation aids consisted
largely of railroad tracks and grain elevators.

Essentially there was no difference between civilian and military flying.
All that was required of a student was that he solo the aircraft safely;
after that, he was on his own; and if he became lost or caught away
from base in unfavorable weather he simply landed in the nearest field.
RCAF "Wings" standard at the time was reached in less than 100 hours
of flying, three-quarters of which was solo practice. With such a
leisurely program, pilot training could be operated quite success-fully
by experienced personnel without an elaborate system of flying assess-
ments. Indeed, flying was an art and it was considered that assessments
of proficiency could be made only by expert pilots on intuitive bases.

Despite the misgivings of researchers and a few flight training
administrators as to the reliability of the traditional evaluation system,
there had been little change in military flight training evaluation over the
years. Substitute measures developed through research were difficult,
and sometimes unsafe, to administer. There has also been the usual
human "resistance to change"; in fact, many flight training personnel
have not viewed the shortcomings of the system as being serious enough
to indicate real need for change.

The study described in this report is, in effect, a continuation of
earlier flight proficiency measurement work. 1 It has been carried
out in the Army Aviation training context with the aim of answering
these questions:

(1) How reliable is the traditional flight check system?
(2) Can standardized, objective, practicable measures of
flight proficiency be developed that will increase both the
reliability and the general diagnostic capacity of flight
training evaluation?

THE TRADITIONAL FLIGHT CHECK SYSTEM

The Flight Check

The flight check is a test of student progress given at the end of
each phase of training. Under the traditional system, the student is
required to fly a sample, or perhaps all, of the flight maneuvers he
has been taught in the preceding phase.

The check pilot usually records his judgments of the student's
performance on a check grade slip after the check ride is completed.
Generally, check pilots do not take notes during the flight. Examples
of two check grade slips, representing two levels of evaluation
specificity, are presented in Figures 1 and 2. After each maneuver,
maneuver part, or specific aspect of flight performance listed on the
check grade slip, the check pilot records a grade which represents his
judgment of the student's performance. Finally, an over-all judgment

1 Much of the research that provided the starting point for this study is summarized in
Appendix A.
### Sample Training Grade Slip

**UNITED STATES ARMY AVIATION SCHOOL**  
**DEPARTMENT OF ROTARY WING TRAINING**  
**GRADE SLIP**

<table>
<thead>
<tr>
<th>Pre-Solo</th>
<th>Intermediate</th>
<th>Advanced</th>
<th>Night</th>
<th>Special</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Grade)</td>
<td>(Rank)</td>
<td>(Class)</td>
<td>(Date)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instructor</th>
<th>(Type Aircraft)</th>
<th>(Grade)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Type Aircraft)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DESCRIPTION</th>
<th>GRADE</th>
<th>DESCRIPTION</th>
<th>GRADE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preflight Inspection</td>
<td>Running Landings</td>
<td>Decelerations</td>
<td></td>
</tr>
<tr>
<td>Cockpit Procedure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Takeoff to Hover</td>
<td>Confined Area Operations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reversing Flight</td>
<td>Pinnacle Operations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal Take-Off</td>
<td>Road Operations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Airwork</td>
<td>Slope Operations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal Approach</td>
<td>Flight w/Servo Off</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Landing From Hover</td>
<td>Use of Lights</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traffic Patterns</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autoglide From Hover</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taxiing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autoglide Approach</td>
<td>Coordination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orientation</td>
<td>Control Touch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPM Control</td>
<td>Planning &amp; Judgment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum Performance Take-Off</td>
<td>Apatitude</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Steep Approach</td>
<td>Composer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross Wind Approaches &amp; Landing</td>
<td>Attitude</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Running Take-Off</td>
<td>Personal Appearance</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TODAY**

<table>
<thead>
<tr>
<th>DUAL</th>
<th>SOLO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Instructor**

---

Figure 1

of the check ride is recorded. This grade is computed as an average of the grades for individual maneuvers or is determined subjectively, without computations, by the check pilot. The check pilot usually explains low grades on the back of the check grade slip.

Check pilots may belong to a special check section. In some units, to become a member of the check section one must be a highly experienced instructor, exceptionally competent and familiar with the training
# Sample Performance Record Used in H-23 Primary Helicopter Training

## PRIMARY PERFORMANCE RECORD

### H-23 PILOT TRAINING

<table>
<thead>
<tr>
<th>Pre-Flight &amp; Cockpit Procedure</th>
<th>UBA AA</th>
<th>Shallow App &amp; Landing: Did He:</th>
<th>UBA AA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hovering: Did He:</td>
<td></td>
<td>Execute Proper Hovering Technique</td>
<td></td>
</tr>
<tr>
<td>Normal Take-Off: Did He:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enter Climb Properly</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintain Proper Power Setting &amp; A/S</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintain Proper Pedal &amp; Cyclic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintain Ground Track</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintain Proper RPM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traffic Patterns: Did He:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintain A/S, Alt. &amp; RPM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coordinate Climbs, Turns &amp; Decelerate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintain Proper Ground Track</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enter and Exit Properly</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal Approach: Did He:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Start at Correct Alt. &amp; A/S</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintain Proper Glide Path &amp; RPM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintain Proper Rate of Closure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Terminate Properly at Hover</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum Performance Take-Off: Did He:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use Correct Pitch &amp; Throttle</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Establish Proper Climb</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintain Directional Control</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Return to Normal Climb</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Steep Approach: Did He:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Start at Correct Alt. &amp; A/S</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintain Proper Glide Path &amp; RPM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintain Proper Rate of Closure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Terminate Properly at Hover</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Running Take-Off: Did He:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use Correct Pitch &amp; Throttle</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Establish A/S Properly</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintain Directional Control</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Return to Normal Climb</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintain Proper Ground Track</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>STUDENT'S NAME</strong></td>
<td></td>
<td><strong>CLASS &amp; FLIGHT</strong></td>
<td></td>
</tr>
<tr>
<td><strong>INSTRUCTOR</strong></td>
<td></td>
<td><strong>MEASUREMENT</strong></td>
<td></td>
</tr>
<tr>
<td><strong>SUPERVISOR'S INITIALS</strong></td>
<td></td>
<td><strong>STUDENT'S INITIALS</strong></td>
<td></td>
</tr>
<tr>
<td><strong>SARP FORM 2</strong></td>
<td></td>
<td><strong>FLYING TIME</strong></td>
<td></td>
</tr>
<tr>
<td><strong>DATE</strong></td>
<td></td>
<td><strong>OVERALL EVALUATION</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2**
program; in other units, requirements are not stipulated. In many flight training organizations there is no formal check section; check pilots may simply be instructors who happen to be available when a check ride is due. In a few training programs all but final check rides are administered by the instructor and are not, in the strictest sense, formal check rides.

The Training Grade

Students are graded by their instructors on their daily performance throughout training. These daily flight training grades and the instructor's written comments are recorded on grade slips which are identical in format with the check grade slips in Figures 1 and 2.

Training grades would be expected to be relatively trustworthy because they are based on many observations by the instructor of the student's performance. However, when the same instructor administers to a student all phases of training in a training stage, a substantial amount of "halo effect" can result. That is, a grade given after a training ride in the latter part of the training phase is likely to be influenced not only by present but also by past performance as remembered by the instructor or as reflected in past training grades. Thus, a reliable check ride, administered by an expert evaluator (other than the instructor), applying a uniform set of standards, is needed as a final independent judgment of the student's proficiency.

Reliability of the Traditional Flight Check System

There is a reasonable basis for the view that students generally may be classified as good, average, or poor throughout training—from stage to stage and, even more clearly, within a stage (i.e., from training to test grades in a given level of training). Perfect consistency in the individual student's performance is not to be expected since various stages of training require different kinds, as well as levels, of skill from the pilot. However, certain perceptual, psychomotor, and mental skills are basic to all flying, whether it is primary, instrument, or tactical. If the evaluation system is adequate there should be an appreciable relationship between a student's training and check grades at different levels of training. Such relationships would not be evident if unreliable measures were used.

Twelve years of flight training research, conducted primarily on Air Force pilot training, and summarized by Ericksen (7) and Ben-Avi (1), indicate that the correlations between check grades at the completion of training and earlier check or training grades are rarely greater than .30.1

1Studies conducted by the Air Force and other research personnel are briefly summarized in Appendix A.
THE FLIGHT CHECK SYSTEM USED IN ARMY HELICOPTER TRAINING

Reliability of the Army Flight Check System

To make preliminary tests of check system reliability, the interrelationships among flight training grades and check grades in primary helicopter training were analyzed by research personnel of the U.S. Army Aviation Human Research Unit in 1956 and 1957 at the U.S. Army Aviation School at Fort Rucker, Ala. At that time primary helicopter training was accomplished in three phases: Pre-Solo, Intermediate, and Advanced. A phase check ride was given at the end of each phase.

The interrelationships among the training grades and check grades for a hundred students are presented in Table 1. Training grades were obtained by averaging the daily dual ride grades for each student for each training phase. The check grades were those recorded in the grade books. The relationships between the training grades and the corresponding check grades were .35, .08, and .09 at the Pre-Solo, Intermediate, and Advanced stages respectively. Thus, the average training-test relationship was little better than zero. In fact, the average of all interrelationships between training and check grades in Table 1 is of the same order of magnitude.

Throughout this report, a value is considered to be reliable (reflecting a true value) if it is significant at the .05 level of confidence or less. For example, if the true correlation were zero, an obtained correlation as large or larger than one marked significant would be expected to occur five or fewer times in a hundred. However, a correlation of small magnitude, even though reliable, is not generally useful for prediction.
A correlation analysis was made of the grades of 55 students at the U.S. Army Primary Helicopter School (USAPHS) at Camp Wolters, Tex. in 1957. The relationships among training grades were generally about the same as those for the U.S. Army Aviation School; relationships between training grades and check grades were slightly higher.

These analyses indicated that results from the proficiency check system used in Army flight evaluation were little more consistent in evaluating student performance than was the traditional system in previous flight training programs. If the assumption of reasonable consistency in individual student performance is correct, there should be an appreciable relationship between a student's training and check grades. Since this was not the case, an examination of the system, with emphasis on check pilot standards, is in order to determine the causes of low reliability.

Variation in Army Check Pilots' Evaluation Standards

To determine the extent of variability in the evaluation standards of check pilots in Army aviation flight training, grades were analyzed for rides administered by the check section in the Department of Rotary Wing Training at Fort Rucker during 1956 and 1957. Ten Intermediate check scores were selected at random from those given by each of eight check pilots, and 10 Advanced check scores were selected for each of eight other check pilots. In Table 2, the mean check grade and the range of the means of the check pilots in each group are presented, as well as the mean variability and the range of variability.

Table 2

Mean and Ranges of Rotary Wing Check Grades Given by Check Pilots, Fort Rucker, 1956-1957

<table>
<thead>
<tr>
<th>Training Phase</th>
<th>Mean^a</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermediate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Check Grade</td>
<td>74.3^a</td>
<td>70.6 to 79.0</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>5.2^a</td>
<td>2.2 to 7.3</td>
</tr>
<tr>
<td>Advanced</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Check Grade</td>
<td>74.9</td>
<td>72.2 to 79.4</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>6.5^a</td>
<td>3.7 to 9.9</td>
</tr>
</tbody>
</table>

^aTen grades are represented in each check pilot mean; eight check pilots are represented in each value given in the table. Check pilots are not the same in the analyses of Intermediate and Advanced grades.

^bThe symbol ~ indicates significance at the .05 level of confidence. Analysis of variance was used to test differences among means. The L test was used to test differences among standard deviations (see Palmer O. Johnson, Statistical Methods in Research, Prentice-Hall, Inc., New York, 1949, pp. 82-86).

^cAs the letter grade system was used at Camp Wolters (AA, above average; A, average; BA, below average; U, unsatisfactory), the Pearson product-moment r was computed by assigning successive integers to letter grades.

^dThe relationships between training grades and check grades are shown in Table 9, p. 25.

^eAnalysis of data from 100 students in the Army, a fixed wing training program in 1957-58 showed interrelationships among the training grades and check grades of about the same magnitude as the rotary wing interrelationships. (These data, from the fixed wing training program at Camp Gary, Tex. and Fort Rucker, Ala., are presented in Appendix B.)
The check pilots differed considerably in the mean ratings they assigned students. Some check pilots seldom fail a student and some fail about half of their students. The differences are statistically reliable for the Intermediate check scores (i.e., they were larger than would result from chance differences in the proficiency of students assigned to these pilots), but not the Advanced check scores.

Since students were not assigned to check pilots on the basis of prior student performance, student assignment is considered to have been random and the results are interpreted to reflect differences in individual check pilot standards. There is a tendency for pilots whose average ratings are high to vary less in their ratings—that is, they rate within a narrow range. This indicates that “easier” check pilots seem to be less willing or less able to discriminate among student performances.

THE FLIGHT PROFICIENCY EVALUATION PROCESS

There are several ways in which individual check pilot standards can be introduced into the evaluation process:

(1) Flight Performance Sample. A check ride should be based on a standardized sample of the student's performance in all the critical maneuvers taught in the preceding phase. However, under the traditional flight check system, the student flight performance actually sampled on a check ride is determined to a large extent by the individual check pilot. Each check pilot tends to have his own set of “favorite” maneuvers which he believes best shows a student’s capability. Then, too, such factors as weather conditions, availability of a particular stage field, and shortage of time may further influence the check pilot’s decision as to what he will include in the flight performance sample on a particular check ride. The check pilot may require the student to repeat a maneuver when he performs it poorly on the first attempt, thus reducing the variety of maneuvers sampled in the time available for the check ride.

To the extent that variations do occur in the flight performance sample from one check ride to the next, different students are faced with different “tests” that usually vary in degree of difficulty. This is particularly true when one student is required to repeat a difficult maneuver several times and another student is not. The first student has a greater opportunity to err than the second student; consequently, he will probably present a poorer over-all picture of his performance, but not necessarily because he is less proficient over-all.

Thus, the test situation is not uniform. Nevertheless, the grade for a check ride is considered to reflect uniformly the level of competence of students at a particular level of training, whether the check ride consists of all the phase maneuvers or a selection of more difficult or easier ones.

Often, if a student performs dangerously on his first or second maneuver, or perhaps halfway through the check, the check pilot terminates the ride and gives the student a failing grade.
he relinquishes the opportunity to analyze the student's difficulties in all maneuvers in which the student has been trained in the preceding phase. When this is the case, subsequent additions, flight time used for the purpose of correcting the student's deficiencies is less likely to be well directed.

Variation in test content, from one check to another, violates what is probably the most fundamental principle of sound evaluation: The sample of knowledge, performance, or behavior which is to be measured must be uniform. Every deviation from the rule, "Every student must be faced with the same set of requirements, under the same conditions," leads to unreliable evaluation.

(2) Observation and Perception. There are many aspects of a student's flight performance toward which the check pilot might direct his attention, such as attitude, altitude, directional control, and power control. Because he cannot observe all these things at once, the check pilot must settle for only a few observations at any one time. From those which he chooses to view at a particular time, his perceptual process may eliminate more.

For example, at a certain point the check pilot may choose to look at the instrument panel. What he actually sees on the instrument panel, however, might be the air speed indicator to the exclusion of the altimeter, tachometer, and needle ball; thus he would notice only certain air speed deviations out of all the many elements he might have observed at that moment. Undoubtedly, in such an instance check pilot bias may play a significant role. Since the check pilot cannot see everything, he looks at what he thinks is most significant.

This problem can be reduced by objectively determining the important indices of flight performance, and from these selecting and standardizing the items that can, practically, be observed.\(^1\)

(3) Memory. A check pilot must observe many details during a check ride. Unless he records descriptions or evaluations of performance at the time it is observed or very shortly thereafter, memory will become a factor in determining the final grade. Indeed, if he completes more than two check flights before recording his judgments on either, he is apt to forget on which ride a particular event occurred. Check pilots with good memories will probably base the grade on more complete details of the student's performance than will check pilots with short memories. More critical, probably, is the problem of selective recall; the check pilot may remember what was most dramatic or most important from his own point of view, which may differ from what another check pilot recalls. Thus, selective bias in observing performance may be compounded by bias in recall of what was observed. To the extent possible, standardized observations should be recorded during or immediately after the actual student performance.

\(^1\) A method frequently used in research for pinpointing interobserver differences is to have two observers evaluate the same performance at the same time. HamRNO researchers' unsuccessful attempts to do this are summarized in Appendix C.
Relative Importance of Maneuvers. Because a single grade must result from a check ride, a weighting method is implicit in the grading process. For example, in helicopter flying, a well-executed forced landing will usually be considered more important than a well-executed normal take-off. Unfortunately, there is less than perfect agreement even among experienced flight instructors as to the relative importance of each training maneuver.

For example, 12 experienced check pilots comprising the entire check section of the U.S. Army Primary Helicopter School were asked to judge the contribution that each of 12 Intermediate and 7 Advanced maneuvers should make to a total flight evaluation score. The means and ranges of the values assigned to Intermediate maneuvers are shown in Table 3, and those for Advanced maneuvers are presented in Table 4. Although the individual check pilots generally agreed on the order of importance of maneuvers, there was substantial variation in ranges of values given each maneuver. For example, one check pilot judged all Intermediate maneuvers as equally important while another check pilot assigned values to the most important maneuvers that were 20 times larger than those he allotted the least important maneuvers. Such variation among check pilots in itself can lead to marked scoring variability.

Table 3

<table>
<thead>
<tr>
<th>Maneuver</th>
<th>Percentage Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>Normal Take-Offs</td>
<td>7</td>
</tr>
<tr>
<td>Normal Approaches</td>
<td>9</td>
</tr>
<tr>
<td>Maximum Performance</td>
<td></td>
</tr>
<tr>
<td>Take-Offs</td>
<td>8</td>
</tr>
<tr>
<td>Steep Approaches</td>
<td>9</td>
</tr>
<tr>
<td>Running Take-Offs</td>
<td>6</td>
</tr>
<tr>
<td>Running Landings</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Basic Autorotations</td>
</tr>
<tr>
<td></td>
<td>180° Autorotations</td>
</tr>
<tr>
<td></td>
<td>Traffic Patterns</td>
</tr>
<tr>
<td></td>
<td>Hovering</td>
</tr>
<tr>
<td></td>
<td>Hovering</td>
</tr>
<tr>
<td></td>
<td>Autorotations</td>
</tr>
<tr>
<td></td>
<td>Forced Landings</td>
</tr>
</tbody>
</table>

The Pilot's Expectations of Student Performance. In his final evaluation, the check pilot must make a judgment about the level of performance that can reasonably be expected from a student for a given phase of training. In the Air Force training programs in the late 1940's and early 1950's, such judgments were found to be substantially

Each check pilot had a minimum of four years' experience in flight training and evaluation programs. Six of the pilots had at least two years' experience as civilian supervisors and check pilots in the Army Helicopter Flight Training program. The six military check pilots had from four to eight years' experience in helicopter flight training. All 12 check pilots had attended the same standardization program for primary helicopter instructors and had worked together for two years.
Table 4
Estimates of the Percentage Contribution
7 Intermediate Maneuvers Should Make to Total Score*

<table>
<thead>
<tr>
<th>Maneuver</th>
<th>Percentage Contribution</th>
<th>Manoeuver</th>
<th>Percentage Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Range</td>
<td></td>
</tr>
<tr>
<td>Take-Offs</td>
<td>16</td>
<td>5 to 25</td>
<td>Slope Operations</td>
</tr>
<tr>
<td>Approaches</td>
<td>17</td>
<td>4 to 20</td>
<td>Hovering</td>
</tr>
<tr>
<td>Flaring Stems</td>
<td>23</td>
<td>15 to 40</td>
<td>Autorotations</td>
</tr>
<tr>
<td>Aircraft Control and Patterns</td>
<td>11</td>
<td>5 to 20</td>
<td>Forced Landings</td>
</tr>
</tbody>
</table>

*One of the 12 USAPHS check pilots did not rate the Advanced maneuvers.

affected by the proficiency of the students whom the instructors used as a basis of comparison. In 1953, Boyle and Hagin (2) demonstrated in a primary pilot training program that 70 per cent of the students with no previous flying experience passed when they were grouped together under the same instructors, and only 49 per cent of these students passed when they were considered with students who had had prior light plane training.

In 1957, Krumboltz and Christal (13) reported data that demonstrated the variation among Air Force instructors in the level of proficiency they expected of their students. The study analyzed the grades for a sample of 216 Air Force aviation cadets from one primary training base during a six-year period. It revealed that a cadet had a better chance of success if he was grouped with cadets of relatively lower aptitude. This was true within several aptitude levels.1

To summarize, the variation in check pilots' standards can be manifested in a number of ways in the flight performance evaluation process. These standards can influence the selection of the flight performance sample, the direction of attention during the flight to certain aspects of performance, perceptual selection from the information to which attention is directed, what is remembered about the performance at the time of recording, the relative importance given to the various maneuvers, and expectations of what student performance should be. In a process as complex and as important as the check flight, it is mandatory that the check pilot's standards for the evaluation process be as uniform as possible.

For these reasons, work was initiated on the development of a flight check system designed to reduce variations in check pilot standards, standardize the sample of flight performance on which scoring is based, and reduce the effects of the check pilots' observation and memory bias on overall score reliability.

1Aptitude was measured with the pilot selection predictors used by the Air Force for selection of air cadets.
Chapter 2
HumRRO RESEARCH
ON FLIGHT PROFICIENCY EVALUATION

DEVELOPMENT OF THE PILOT PERFORMANCE
DESCRIPTION RECORDS (PPDR's)

As part of Subtask LIFT II, development of standardized, relatively
objective measures of primary and basic light helicopter pilot profi-
ciency was undertaken. Initial guidelines for the format of the measures
were provided by Air Force research (18, 19).

An analysis of the training program, including study of grade books
and interviews with instructors, was the basis of determining the areas
of flight training in which students have the most difficulty (17). Each
primary and basic training maneuver was analyzed into its components.
For each component, simplified scales were developed on which the
check pilot could quickly record his observations or judgments as the
student performed. For some components (such as pedal usage, approach
path to confined areas, and ground track on downwind legs) on which the
check pilot had to make more complex judgments, rating scales of a
more subjective type had to be developed. In such items the points on
the scales were defined as precisely as possible to minimize personal
interpretation by the check pilot. Where possible, however, scales
were developed on which the check pilot could immediately record direct
observations on instruments or outside cues (such as RPM, air speed,
altitude, and approach termination points).

The original list of item components for each maneuver was thor-
oughly tested in simulated check rides by LIFT II research personnel.
It was found that the number of items was more than a check pilot could
safely evaluate in the allotted time. Therefore, experienced flight
training personnel were asked to select only the items which would
adequately describe the most critical components from each maneuver
segment. In subsequent tryouts a descriptive record of student perform-
ance was produced which could be administered safely and accurately
by a trained check pilot.

The measures—the Pilot Performance Description Records
(PPDR's)—were based on standard rides; that is, the same number of
maneuvers were to be flown in the same sequence on each ride. The
check pilots were instructed to immediately record their observations
or judgments of each maneuver component, except for those maneuvers
in which safety considerations dictated against this procedure. For
autorotations, the latter half of approaches, the initial phase of takeoffs, and forced landings, recording was postponed until completion of the maneuver.

The Pilot Performance Description Records were administered on a trial basis in 1957, and in revised form in 1958. The 1958 versions of the Intermediate and Advanced PPDR's are described in the Manual of Instruction for use of the PPDR's (10).

TRYOUT OF THE 1957 VERSION OF THE PPDR

Procedure

In 1957, the PPDR was used in check rides administered to 75 students (40 Intermediate; 35 Advanced) at the U.S. Primary Helicopter School at Camp Wolters. Examples of maneuver record sheets from the 1957 versions of the Intermediate and Advanced PPDR's are presented in Figures 3 and 4.

Each student flew two check rides, each with a different check pilot. The student's first ride was flown by one of the two check pilots on the LIFT staff, and the second by one of four USAPHS military check pilots. This procedure made it possible to estimate the agreement (ride/ride relationship) between repeated evaluations of the same students. The student did no flying between these check rides.

The assignment of students to the LIFT check pilots was on a random or "chance" basis. For the second ride, each military check pilot was alternately assigned a student checked by the first LIFT check pilot and one checked by the second LIFT check pilot. The initial random assignment of students to the two LIFT check pilots ensured that there was no selective bias throughout the checking procedure.

The LIFT check pilots were intimately familiar with the system, having been part of the team responsible for its development. The military check pilots had received only a brief training program from the LIFT research staff pilots. This training consisted of (1) approximately four hours of lectures, during which the rationale of the system was presented and each type of scale was described and interpreted; (2) in-flight demonstration by the LIFT staff check pilots of the recording system, including safety training (e.g., the check pilot is to stop recording during certain maneuvers or parts of maneuvers for reasons of safety); (3) a complete check ride with a LIFT check pilot acting as the student and the military pilot recording the flight; and (4) at least one practice ride with a student pilot.

Readers familiar with the Camp Wolters training program will note that "Intermediate" is substituted for "Primary" and "Advanced" for "Basic" to reduce confusion that might result from referring to the primary phase of primary training. "Intermediate" and "Advanced" had previously been used to refer to the same training phases at Fort Rocker when primary helicopter training was conducted there.
Sample Record Sheet From 1957 Version of the Intermediate (Primary) Pilot Performance Description Record

**PANEL**
- RPM
  - 29 20 22 23
- End/Disc
  - Short
- Alt
  - 1 2 3
- Rev/Cl
  - Slow

**Descent**
- Line of Descent
- Flight Path
- Rev/Cl
  - Slow
- Pedals
- RPM
  - 29 20 22 23

**Final**
- Sigh/Pto
  - Shallow
- A/S
  - -30 -10 -3 5 15 -30 50
- Alt
  - -300 -100 -30 30 100 -300

**Final Turn**
- RPM
  - 29 30 22 23
- Pedals

*Normal Approach*
Sample Record Sheet From 1957 Version of the Advanced (Basic)
Pilot Performance Description Record

CONFINED AREA OPERATION

CHECK PILOT: Select a confined area to which a touchdown is possible; point it out to the student.

I. High Reconnaissance
   A. Pattern flown with respect to:
      1. Wind and forced landing areas
         U    Poor    Adeq    Best
         U    Poor    Adeq    Best
      2. Observation/angle of sight
         U    Poor    Adeq    Best
   B. Aircraft Control:
      1. Airspeed (40-50 K)
         Ent    Low    High    Proper
      2. RPM (3050-3150)
         Ent    Low    High    Proper
      3. Altitude
         Ent    Low    High    Proper
      4. Pedals
         U    Poor    ........    Proper

II. Approach and Low Reconnaissance
   A. Down to barrier:
      1. Line of descent
         ........    ....    ....    ....    ....
      2. Approach angle
         ........    Shallow    Steep    Proper
      B. Rate of closure
         Ent    Slow    Fast    Proper
      3. Rate of closure
         U    Poor    ........    Proper
      4. Pedals
         U    Poor    ........    Proper
   B. Over barrier to ground:
      1. Clearance of barrier
         ........    Too Close    Too High    Proper
      2. Line of descent
         ........    ....    ....    ....    ....
      3. Rate of closure
         Ent    Slow    Fast    Proper
      B. Altitude/visibility approach
         Ent    Low    High    Proper
      5. RPM (3050-3150)
         Ent    Low    High    Proper
      6. Unnecessary lowering
         Yes    ........    No
      7. Pedals
         U    Poor    ........    Proper

Confined Area

Figure 4
Scoring of the PPDR's

Four types of scores were used in scoring the check rides: total error score, item-weighted score, error pattern-weighted score, and traditional score. The scores are defined as follows:

The total error score—the number of item errors recorded on the PPDR.

The item-weighted score—the sum of item errors weighted by item importance and difficulty, converted to a percentage of the total possible score. Item weights are the average of values (ranging from 1 to 5) assigned by experienced check pilots judging the difficulty and importance of each item.

The error pattern-weighted score—the sum of check pilot ratings on maneuver segments, weighted by maneuver importance and converted to a scale ranging from 0 to 100. Check pilots rate the performance on each maneuver segment on a scale ranging from 0 to 10 (0, dangerous; 1-2, unsatisfactory; 3-5, below average; 6-8, average; 9-10, above average). Maneuver weights reflecting the difficulty and importance of each maneuver are the average of values assigned by experienced check pilots.

A traditional score—an over-all score for the check ride in terms of a letter grade (AA, above average; A, average; BA, below average; U, unsatisfactory). Check pilots were asked to assign this score on the basis of their own judgments (i.e., not to take the PPDR's into account).

The HumRRO check pilots scored all of the Intermediate check rides to obtain the error pattern-weighted scores.

Results of the 1957 PPDR Tryout

The ride/ride relationships for the PPDR check rides administered in the 1957 tryout are presented in Table 5. These data indicate an increase in reliability over the traditional system, particularly in the item-weighted and error pattern-weighted scores.

### Table 5
Correlations Between Rides, 1957
(Camp Walters)

<table>
<thead>
<tr>
<th>Score</th>
<th>Intermediate PPDR (N = 40)</th>
<th>Advanced PPDR (N = 35)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPDR Score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item-Weighted</td>
<td>.42*</td>
<td>.37*</td>
</tr>
<tr>
<td>Error Pattern-Weighted b</td>
<td>.51*</td>
<td>-</td>
</tr>
<tr>
<td>Total Error</td>
<td>.17</td>
<td>.28*</td>
</tr>
<tr>
<td>Traditional Grade</td>
<td>.25*</td>
<td>.10</td>
</tr>
</tbody>
</table>

*The symbol* *indicates significance at the .05 level of confidence.

bThe error pattern-weighted scores were obtained only for the Intermediate PPDR's.
The diagnostic capacity of the 1957 PPDR's was clearly demonstrated. The PPDR's made it possible to count not only total errors but also errors on specific elements in the performance (such as pedals, RPM, air speed, altitude, and ground track).

An analysis was made of the errors recorded by the check pilots on selected PPDR scales (those accounting for over half of the PPDR items) and two over-all scores. The difference between the LIFT staff pilots (who were thoroughly familiar with the PPDR) was subtracted from the average difference between the USAPHS pilots (who were only oriented in the use of the PPDR) to obtain a "similarity index" for each PPDR item and score listed in Table 6. The scoring similarity of the LIFT pilots was reliably greater for the items "pedals" and "RPM" and for the traditional grades. The difference between the two groups on the remaining items analyzed was negligible.

Table 6

<table>
<thead>
<tr>
<th>Item</th>
<th>Check Pilot</th>
<th>PPDR Experts</th>
<th>PPDR-Oriented</th>
<th>Similarity Index*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Rides</td>
<td>20</td>
<td>20</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Mean Percentage of All Errors Possible for:</td>
<td>20</td>
<td>20</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Pedals</td>
<td>21</td>
<td>19</td>
<td>5</td>
<td>11*</td>
</tr>
<tr>
<td>RPM</td>
<td>37</td>
<td>37</td>
<td>10</td>
<td>16*</td>
</tr>
<tr>
<td>Air Speed</td>
<td>59</td>
<td>47</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>Altitude</td>
<td>26</td>
<td>22</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>Ground Track</td>
<td>15</td>
<td>13</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Mean PPDR Item-Weighted Score</td>
<td>75</td>
<td>74</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>Traditional Grade†</td>
<td>40</td>
<td>30</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

*The t test was used to determine the significance of the amount of difference between the two groups of check pilots. The one-tailed test was used for the null hypothesis that the check pilots who were experts in the PPDR system were no similar or less similar to each other as were the check pilots who were only oriented in the PPDR system. The symbol * indicates a difference that is significant at the .05 level of confidence.

†These scales constituted over half of the items on the PPDR.

†Based on the percentage of "average" and "above average" grades given. The PPDR was not referred to in assigning the traditional grades in 1957.

The analysis of the Advanced PPDR data in Table 7 indicates there is no systematic difference between the similarity of PPDR experts and that of the PPDR-oriented check pilots. The only significant difference between the two groups is for the traditional score, on which the PPDR experts were reliably more alike. For the item-weighted score and for "air speed," there is no difference in the amount of similarity.
of the evaluations made by the two groups of check pilots. On the remaining items, the PPDR-oriented check pilots were more similar than the PPDR experts, but not to a degree that is statistically reliable. Since the LIFT pilots had extensive experience with the Intermediate PPDR only (their work with the Advanced PPDR was in the early stages), their familiarity with the Advanced PPDR was little greater than that of the PPDR-oriented check pilots. Over-all, the value of experience in the use of the PPDR is strongly indicated.

Table 7
Mean Percentage of Errors Recorded for Selected Items and Mean Over-All Scores for the Advanced PPDR's, 1957

<table>
<thead>
<tr>
<th>Item</th>
<th>PPDR Experts</th>
<th>PPDR-Oriented</th>
<th>Similarity Index*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Rides</td>
<td>15 20 10 10 6 9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Percentage of All Errors Possible for:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pedals</td>
<td>9 18 14 5 7 10</td>
<td>-10</td>
<td></td>
</tr>
<tr>
<td>RPM</td>
<td>22 36 12 15 8 30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Air Speed</td>
<td>36 31 21 18 23 27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Altitude</td>
<td>22 29 19 12 15 15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyclic Control</td>
<td>35 22 21 10 15 15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Items</td>
<td>31 36 24 24 24 30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean PPDR Item-Weighted Score</td>
<td>80 63 88 85 85 82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traditional Grade</td>
<td>47 50 80 50 50 44</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>22*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The t-test was used to determine the significance of the amount of difference between the two
groups of check pilots. The one-tailed test was used for the null hypothesis that the check pilots who
were experts in the PPDR system were as similar or less similar to each other as were the check pilots
who were only oriented to the PPDR system. The symbol * indicates a difference that is significant at the
.05 level of confidence.

**These scales constituted over half of the items on the PPDR.

*Based on the percentage of "average" and "above average" grades given. The PPDR was not
referred to in assigning the traditional grades in 1957.

TRYOUT OF THE 1958 VERSION OF THE PPDR

Procedure

The PPDR's were modified on the basis of practical experience and
data obtained during the 1957 experimental administration. Revisions
of the Intermediate PPDR were relatively minor, consisting largely of
changes in format making it easier for the check pilot to determine quickly
where to record his observations. A few items which had been shown
to serve no purpose were eliminated, and others were added where it
had been found that student performance was not described sufficiently.
Modifications of the Advanced PPDR were substantial. The type of specific scale used in the Intermediate PPDR items was substituted for the more categorical type of scale which had been used in the 1957 Advanced PPDR. Many ineffective scales were eliminated on the basis of experience, and a set of maneuvers requiring take-offs and approaches over a tree, both into the wind and crosswind, was added.

Examples of the format used for the 1958 version of the Intermediate and Advanced PPDR's are presented in Figures 5 and 6. Both PPDR's are described in detail in the Manual of Instruction for the use of the PPDR's (12).

In the 1958 experimental tryout, 12 check pilots were trained to administer the PPDR. Six of the check pilots were civilian flight commanders or other responsible training administrators with the Southern Airways civilian contract school at the USAPH, Camp Wolters, and six were military pilots who were part of the monitoring military check section at the USAPH.

The training program was administered to the 12 check pilots by the two LIFT pilots who had participated in the 1957 evaluation. The 1958 training was somewhat more comprehensive than that given in 1957. It lasted one week and consisted of (1) a three-hour detailed presentation and discussion of the individual scales in the Intermediate and Advanced checks; (2) two hours of in-flight orientation in the use of the PPDR's, conducted by the two LIFT staff pilots; (3) practice with at least one student; (4) a final "evaluation" ride with a LIFT pilot simulating student performance; (5) a procedure for identifying markedly different individual check pilot standards, and partially modifying these standards (requiring approximately five hours of classroom work for each check pilot).1

Following the check pilot training program, two successive check rides were administered to each of 50 Intermediate and 50 Advanced student pilots to obtain estimates of check pilot agreement (reliability). The first ride was always administered by one of the civilian pilots, and the second by one of the military pilots.

The four scores computed for each PPDR check ride were essentially as described for the 1957 tryout. However, in 1958 each check pilot (military and civilian) scored the PPDR immediately after completion of the check ride and provided the error pattern-weighted score. Also, the check pilots were required to base the "traditional" score on a careful review of the PPDR results.

Results of the 1958 PPDR Tryout

Ride/ride relationships for the 1958 PPDR tryout are presented in Table 8. The error pattern-weighted and the traditional scores are the most reliable. The traditional score for the PPDR tryout in 1958 is reliably higher than that for 1957.

1 This procedure is described in more detail in Chapter 3, pp. 28-29.
Sample Record Sheet from 1958 Version of the Intermediate PPDR

<table>
<thead>
<tr>
<th>PANEL</th>
<th>RPM</th>
<th>26</th>
<th>28</th>
<th>30</th>
<th>32</th>
<th>33</th>
</tr>
</thead>
<tbody>
<tr>
<td>End/Overshoot</td>
<td>Sh</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>alt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rate/Climb</td>
<td>Slow</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pedals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Line of Descent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flight Path</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rate/Climb</td>
<td>Slow</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pedals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPM</td>
<td>29</td>
<td>30</td>
<td>31</td>
<td>32</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>Entry</td>
<td>Shallow</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sight/Pic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R/A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALT</td>
<td>-100</td>
<td>-50</td>
<td>0</td>
<td>50</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Gnd/Taxi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pedals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final Turn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPM</td>
<td>29</td>
<td>30</td>
<td>32</td>
<td>33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pedals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NORMAL APPROACH

Figure 5
Sample Record Sheet From 1958 Version of the Advanced PPDR

RPM

End/Oess

ALT

Pedals

Descent

(Over Barrier)

Line of Descent

Pedals

RPM

Entry

Approach Angle

Shallow

StEEP

II. LOW RECONNAISSANCE and APPROACH

A/C CONTROL DURING PATTERN

ALT

RPM

Wind/Forced landing areas

Observation/Angle of sight

Pattern flown with respect to:

I. HIGH RECONNAISSANCE

CONFINED AREA OPERATION

Figure 6
These results provide evidence that subjective judgments can, under controlled conditions, serve to provide a reliable general estimate of student proficiency.

Table 8
Correlations Between Rides, 1958*
(Camp Wolters)

<table>
<thead>
<tr>
<th>PPDR Score</th>
<th>Intermediate PPDR</th>
<th>Advanced PPDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item-Weighted</td>
<td>.17</td>
<td>.42*</td>
</tr>
<tr>
<td>Error Pattern-Weighted</td>
<td>.42*</td>
<td>.52*</td>
</tr>
<tr>
<td>Total Error</td>
<td>.14</td>
<td>.37*</td>
</tr>
<tr>
<td>Traditional Grade (PPDR-based)</td>
<td>.48*</td>
<td>.47*</td>
</tr>
</tbody>
</table>

*The number of students in 50 in all cases except for the Intermediate error pattern-weighted score, where it is 49. The symbol * indicates significance at the .05 level of confidence.

It was stated earlier that a flight proficiency evaluation system should reflect to a substantial degree the consistency presumed to exist in student flight performance from early to later training. Data presented in Table 1 and in Appendix B indicate that interrelationships between the Army's traditional check scores and training grades are low. By comparison, the relationships between the 1958 PPDR scores given by the military check pilots (on the second of two rides by each student) and the training grades given in the Pre-Solo, Intermediate, and Advanced phases of training are substantially higher. To facilitate comparison, Table 9 presents the relationships of training grades with check scores given at Camp Wolters in 1957 and with those given during the PPDR administration in 1958.

For the Advanced training phase, the PPDR error pattern-weighted scores and traditional scores (based on the PPDR check ride) should show a relatively high relationship to training grades. This is particularly significant because the Advanced phase of training includes the maneuvers that are most similar to those a helicopter pilot would have to perform in tactical flying. The relationships between traditional grades (PPDR-based) and training grades are also high for the Intermediate training phase. However, the relatively low relationships of the Pre-Solo scores and the Intermediate PPDR scores to training grades may suggest that there is less consistency in student performance in the early phases of training as compared with the later phase. Miller (14) has suggested that the crucial source of unreliability of check rides is the lack of consistency in pilot performance from day to day. The relatively high reliability of results for the Advanced PPDR may suggest that this conclusion is appropriate only for the early stages of training. The somewhat higher relationships between Intermediate check scores and Advanced training grades further support this interpretation.
Table 9
Correlations Between Training Grades and Check Grades (1957) and Training Grades and PPDR Scores (1958)*
(Camp Walters)

<table>
<thead>
<tr>
<th>Check Grade or Score</th>
<th>Training Grade</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre-Solo</td>
<td>1957</td>
<td>1958</td>
</tr>
<tr>
<td>Traditional Check Grade</td>
<td>N = 55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Solo</td>
<td>.45*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intermediate</td>
<td>.14</td>
<td>.44*</td>
<td></td>
</tr>
<tr>
<td>Advanced</td>
<td>.24</td>
<td>.10</td>
<td>.45*</td>
</tr>
<tr>
<td></td>
<td>N = 50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Solo</td>
<td>.58*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intermediate PPDR Score</td>
<td>N = 55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Error</td>
<td>.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item-Weighted</td>
<td>.20</td>
<td>.14</td>
<td>.23*</td>
</tr>
<tr>
<td>Error Pattern-Weighted</td>
<td>.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traditional (PPDR-based)</td>
<td>.35*</td>
<td></td>
<td>.37*</td>
</tr>
<tr>
<td>Advanced PPDR Score</td>
<td>N = 50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Error</td>
<td>.50*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item-Weighted</td>
<td>.48*</td>
<td>.55*</td>
<td></td>
</tr>
<tr>
<td>Error Pattern-Weighted</td>
<td>.52*</td>
<td>.55*</td>
<td>.42*</td>
</tr>
<tr>
<td>Traditional (PPDR-based)</td>
<td>.55*</td>
<td>.60*</td>
<td>.51*</td>
</tr>
</tbody>
</table>

*The symbol * indicates significance at the .05 level of confidence.

Since relationships between check scores and training grades, particularly for the Advanced PPDR, are of a magnitude that requires relatively high reliability of measurement, it appears that the PPDR evaluation system is basically sound. The marked improvement in the reliability of the concurrent traditional score may be attributable to (1) the diagnostic data obtained with the PPDR, (2) the necessity to review the PPDR's to determine over-all scores, and (3) the PPDR check pilot training program. However, an additional factor must be considered: Both the civilian (Southern Airways) training and military check section personnel at Camp Walters were devoting every effort between the 1957 and 1958 tests to improvement of the training and monitoring system. Undoubtedly these efforts resulted in increased standardization as well as improved training. This is suggested by the somewhat higher relationships in 1958 than in 1957 between training grades and pre-solo check grades (which were not based on the PPDR system). Unfortunately, data reflecting relationships between training grades and the traditional check grades were not analyzed just before the 1958 tryout. This would have provided a more complete control for the comparison of
training grades with check grades and of the PPDR with the traditional check system. Thus, the data presented do not constitute proof, but are only substantiating evidence that the PPDR system was more reliable than the traditional check system in 1958.

Substantial improvement in reliability is indicated, however, by the higher interrelationships between training grades in 1958 (ranging from .61 to .74) as compared with those for 1957 (ranging from .10 to .45). The sizable increase in training grade interrelationships in the 1958 PPDR tryout suggests that the evaluation of training rides as well as check rides had become more standardized.

However, even more standardization is necessary since there was still considerable variation among check pilots in the PPDR’s administered in 1958. The percentages of all possible errors that were scored by the 12 check pilots on selected items and for all items of the PPDR’s, as well as PPDR-derived scores, were computed and the means and standard deviations are presented in Appendix Table D-1. Although the one-week training program given for the 1958 tryout appears to have made check pilot standards more uniform than in 1957, it did not eliminate check pilot differences. It is noteworthy, however, that a major contribution of the PPDR’s is the extent to which they allow specification of some of variation in check pilot standards.

The PPDR system itself is substantially more diagnostic and more reliable than the traditional system. However, either a more intensive check pilot training program or a check pilot selection program, or more likely both, must be initiated if the remaining substantial effects of check pilot biases on flight proficiency evaluation reliability are to be further reduced.
Chapter 3

APPLICATION OF THE PPDR SYSTEM

CHARACTERISTICS OF THE PPDR SYSTEM

The prototype flight check evaluation system developed in this study consists of (1) Intermediate and Advanced PPDR booklets on which personnel serving as check pilots can score specific maneuvers on standardized, relatively objective scales; (2) a training program to familiarize the check pilot with the concepts and techniques involved in the PPDR system and to give him practice in administering check rides using the PPDR; (3) classroom training for check pilots in scoring standard PPDR's—that is, an identical set of PPDR's for actual check rides—to allow identification of specific areas in which the check pilots' standards of evaluation are atypical; (4) methods of scoring the PPDR, the most promising of which is the error pattern-weighted score which reflects both the importance of each maneuver, in the judgment of expert opinion, and the check pilot's evaluation of over-all performance of each maneuver.

The PPDR system requires that the same flight test situation be presented to each student pilot. The type and number and, insofar as possible, the sequence of maneuvers included in a flight check is rigorously standardized. This fulfills the fundamental principle of sound evaluation that all students be exposed to conditions which are as nearly identical as possible. The existence of variables that cannot be controlled, such as weather and differences in flight characteristics of aircraft, makes it even more essential that controllable factors be standardized.

The PPDR provides a detailed and permanent record of the student's performance on a flight sample of critical maneuvers. The record can be analyzed in detail to diagnose student performance or to compare check pilot observations with those of other check pilots. The flight performance sample utilized in the PPDR system is realistic; it has been selected on the basis of a complete analysis of training maneuvers, tactical flying requirements, and expert pilot opinion. Most crucial maneuvers are included in the PPDR check ride.

SCORING STANDARD PPDR's AS PART OF THE 1958 TRAINING PROGRAM

The requirement that check pilots use similar standards in recording their observations and in scoring the data that they have recorded cannot be overemphasized. One method of determining whether check pilots are using similar standards is to have them evaluate the same
performance and then compare their evaluations. This was attempted as part of the LIFT study.

During the one week of training that check pilots received in 1958, each of the 12 check pilots was presented with an identical set of 10 completed PPDR descriptions of actual student flight performances. The PPDR's were selected from those administered as part of the 1957 evaluation program and represented a wide range of student performance. The cover sheets on which final scores and information about the student had been recorded were removed so that the pilots would not have any initial bias.

The check pilots assigned a rating from 0 to 10 (0, dangerous; 1-2, unsatisfactory; 3-5, below average; 6-8, average; 9-10, above average) to each of more than 100 maneuvers and maneuver components for each PPDR. These ratings were multiplied by maneuver weights which had been determined by a group of expert check pilots on the basis of difficulty and criticality of each maneuver. A total score was then determined by summing the weighted ratings.

It should be noted that in scoring these 10 PPDR's the 12 pilots were required to evaluate only recorded descriptions of the flight performance. No actual flight checking was involved.

For the 12 check pilots, a score for each of the 10 standard PPDR's was obtained. Correlations between pairs of check pilots' evaluations ranged from .82 to .99, indicating considerable agreement in scoring between the check pilots. However, the differences between check pilots, even within this limited range, appear meaningful in terms of agreement in scoring actual flight checks.

CLASSROOM SCORING AGREEMENT AND RIDE/RIDE RELATIONSHIPS

The relatively simple classroom technique described above shows promise as a method for quickly pinpointing differences in check pilot standards that would produce differing results in actual flight checks. Following the administration of the PPDR's in 1958, it was possible to select pairs of check pilots who had checked the same students and to compare the agreement of their standards in the classroom scoring with the agreement of the scores given by them to the same students during a flight check.

Table 10 shows the relationships of PPDR flight check scores for check pilot pairs whose classroom scoring agreements were from .82-.99 (all pairs), .91-.99, and .95-.99. It is clear from Table 10 that more agreement in the classroom scoring does mean considerably more agreement in actual flight check evaluation for the Intermediate

Attempts were also made to present the same flight performance to check pilots by flying two observers at the same time and by presenting student performance on film, but these efforts were not successful. The studies on interobserver relationships are presented in Appendix C.

*This procedure is used to obtain the error pattern-weighted score (see p. 18). Means and ranges of maneuver weights are given in Tables 3 and 4.
Table 10
PPDR Flight Check Scoring Agreement for Check Pilot Pairs Compared With Classroom Standard PPDR Scoring Agreement*

<table>
<thead>
<tr>
<th>PPDR Score</th>
<th>Correlations of Flight Check Scores* for Check Pilot Pairs Whose Scores on Standard PPDR’s Correlated:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.82-.99</td>
</tr>
<tr>
<td>Intermediate PPDR</td>
<td>Item-Weighted</td>
</tr>
<tr>
<td></td>
<td>Error Pattern-Weighted</td>
</tr>
<tr>
<td>Advanced PPDR</td>
<td>Item-Weighted</td>
</tr>
<tr>
<td></td>
<td>Error Pattern-Weighted</td>
</tr>
</tbody>
</table>

*Standard scoring was performed on the same 10 Intermediate PPDR booklets by all check pilots.

The symbol * indicates agreement that is significant at the .05 level of confidence. All 50 students are represented for the Intermediate and Advanced correlations for all check pilot pairs. For check pilot pairs with agreement of over .91, the number of students in 42 and 44 for the Intermediate and Advanced PPDR, respectively; for pairs with agreement of over .95, the number is 33 and 33, respectively.

PPDR. A trend in the same direction, but less pronounced, is shown for the Advanced PPDR. However, it must be remembered that the classroom method for comparing standards was based only on Intermediate PPDR records. It would be expected that the Advanced PPDR agreement would be much better predicted with a classroom technique for the Advanced PPDR, which can easily be developed and applied.

On the basis of these results, it seems probable that training of check pilots in scoring standards PPDR’s can increase uniformity of standards and consequently lead to greater reliability of the evaluation system. Since an increase in reliability is critical to future training methods research and to improvement of flight training by training supervisors, still more effort should be directed toward development of uniformity of standards among check pilots.

DEVELOPMENT OF THE PPDR SYSTEM FOR OPERATIONAL USE

The data obtained in this study provide the basis for further development of the PPDR system, by means of:

1. Refinement of the PPDR and scoring method.
2. Extension of the training of personnel serving as check pilots. An aviator assigned to duty as a check pilot has the necessary flight qualifications and requires only training to become qualified in the use of PPDR’s. Selection may be necessary where check standards are extremely lenient and cannot be modified.
3. Establishment of an information system which will provide feedback on training results (a) to students for determining specific areas where extra training is necessary, (b) to instructors to inform them of specific weaknesses in their instruction, (c) to command personnel regarding the effectiveness of the over-all program of instruction, and (d) to check pilots, showing where, over time, their standards are not sufficiently uniform.
REFERENCES


Appendix A

PRIOR RESEARCH ON THE USE OF OBJECTIVE MEASURES IN FLIGHT PERFORMANCE EVALUATION

At least one research effort has, with some success, been directed toward attempting to improve the reliability of the grade resulting from the traditional subjective system. Crawford and Da' ey (5) reported a technique for using Air Force flight instructors' comments written on the backs of grade slips. Greater reliability of evaluation resulted from their method than from use of the grade alone. While the technique may be cumbersome for regular use in a training program, the study did indicate that instructors and check pilots are capable of more reliable evaluation of student flying proficiency than they manifest in the regular grading system.

The efforts of research personnel to reduce the effects of differences in check pilot standards and to otherwise increase the reliability and diagnostic capacity of flight proficiency evaluation were directed primarily toward making the evaluation system more objective. In the systems that have been developed, research personnel have tried to increase the extent to which the check pilot observes and describes rather than evaluates during the actual check ride. The larger subjective judgments are reduced to smaller specific judgments (e.g., too much left pedal during the first take-off; over-controlled on the third landing) or, in scoring, a subjective score is assigned to each error rather than to the totality of errors. Description has been assumed to be an essential characteristic of a diagnostic flight performance evaluation, and to be fundamental to its reliability.

As early as 1939, a research attempt was made to devise a means of obtaining more objective and detailed, as well as more reliable, information from flight proficiency measures. The resulting Ohio State Flight Inventory (6) was directed toward increasing the objectivity of flight proficiency measures. Other research efforts along these lines prior to 1947 are summarized by Ben-Avi (1), and those before 1962, by Erickson (2). In one of the most successful studies, reported by Gordon (9) and Nagay (15), a system of evaluation for airline pilot proficiency was devised that depended largely on objective and detailed in-flight records. This system provided a ride/ride reliability of .70, one of the highest yet reported. The reliability was based on the relationship between two successive administrations of the same check ride to the same student by different check pilots. Of course, the airline pilot's activities are more procedural and require less frequent
and less gross control adjustments than do the lighter aircraft on which most flight proficiency research has been done.

A well-conceived research effort conducted for the Navy in 1952 (29) did not result in an increase in the reliability over that obtained in the traditional system. The objective evaluation method which was devised proved no more reliable at the pre-solo (ride/ride relationship of .32) and instrument (ride/ride relationship of .33) stages than the traditional subjective method (.42 and .41, respectively). It is noteworthy that the reliability of the traditional method reported in the Navy study was higher than in most studies. The authors attribute the low reliability of the experimental system to day-to-day fluctuations in student performance rather than to errors of measurement, citing Miller (14, p. 361) for support. "Different check pilots" is also listed as a reason for this low reliability, along with weather and aircraft differences. Considering other flight proficiency research successes, these explanations hardly seem to be adequate.

It should be noted that in the Navy study there was considerable resistance to the objective check on the part of the instructors. Sixty-nine per cent of the instructors who participated in the tryout considered the in-flight use of the objective booklets dangerous. This reaction may be accounted for by the facts that (1) one of the checks was used at the pre-solo stage; (2) the format of the booklets in which the check pilots recorded their observations required considerable "head-in-cockpit" time to find out where to record; and (3) inadequate training in the use of the booklets was given the check pilots. However, check pilot aversion to objective checks has been encountered to some extent in most studies.

Probably the most definitive flight evaluation work has been accomplished by the Basic Pilot Training Research Laboratory of the Human Resources Research Center, Air Training Command, Goodfellow Air Force Base, San Angelo, Tex. The work described in this report was largely based on the Air Force precedent. The developmental aspects of the Air Force work are described by Smith, Flexman, and Houston (19) and Smith and Flexman (18). The objective method developed was relatively reliable in comparison with the traditional system (most estimates of ride/ride relationships averaged above .50), but the reliability varied considerably from one application to the next, ranging from .17 to .69 (12). However, the diagnostic capability of this flight proficiency description system was of great value. Excellent examples of its use for this purpose are presented by Flexman et al. (8), Ornstein et al. (16), and Houston (11). In these reports, detailed, objective information about specific errors made by students at various stages of training was presented which demonstrated the kind of valuable analysis which is made possible by an objective flight evaluation system, as compared to the traditional subjective system.

*This view is also held by certain Air Force researchers; see Bray (3).

*Efforts by HamRRO to develop a pre-solo helicopter check had to be discarded because of the safety factor.
In the various research efforts, increasing objectivity and requiring subjective judgments to be more specific have usually resulted in higher reliability and almost always have produced greater analytic capacity in comparison with the traditional method. But the increases in reliability of check grades have not been as great as is desired, and the fluctuating reliability of the objective check has plagued researchers. Apparently, the requirement for check pilots to attend to and describe, or judge (where description is not possible), specific aspects of student performance is, of itself, no guarantee of high reliability. Check pilot biases seem to be manifested in “relatively objective” measures as well as in subjective measures, and this probably accounts for low or fluctuating reliability. Thus, primary attention should be accorded the problem of reducing differences in check pilot standards so that the more objective measures can be used reliably and for detailed diagnosis of training programs.
Appendix B

RELATIONSHIPS BETWEEN CHECK GRADES
AND TRAINING GRADES
IN THE ARMY'S FIXED WING TRAINING PROGRAM

Table B-1
Correlations of Fixed Wing Check and Training Grades,
Camp Gary, 1957-1958

\(N=100\)

<table>
<thead>
<tr>
<th>Check Grade Inter-correlations</th>
<th>25-Hour</th>
<th>.20</th>
<th>&lt;.10</th>
<th>.33*</th>
<th>&lt;.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>50-Hour</td>
<td>.20</td>
<td>.20</td>
<td>.14</td>
<td>.26*</td>
<td>.10</td>
</tr>
<tr>
<td>Basic Instrument</td>
<td>-</td>
<td>-</td>
<td>.26*</td>
<td>.10</td>
<td>.22*</td>
</tr>
<tr>
<td>90-Hour</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Training Grade Inter-correlations</th>
<th>25-50 Hour Check</th>
<th>After 50 Hour Check</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up to 25 Hour Check</td>
<td>.47*</td>
<td>.41*</td>
</tr>
<tr>
<td>25-50 Hour Check</td>
<td>-</td>
<td>.50*</td>
</tr>
</tbody>
</table>

Check and Training Grade Correlations

<table>
<thead>
<tr>
<th>Training Grade</th>
<th>25-Hour</th>
<th>50-Hour</th>
<th>Basic Instrument</th>
<th>90-Hour</th>
<th>Advanced Instrument</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up to 25 Hour Check</td>
<td>.32*</td>
<td>.32*</td>
<td>.10</td>
<td>.10</td>
<td>.20</td>
</tr>
<tr>
<td>25-50 Hour Check</td>
<td>.37*</td>
<td>.32*</td>
<td>.17</td>
<td>.22*</td>
<td>&lt;.10</td>
</tr>
<tr>
<td>After 50 Hour Check</td>
<td>.33*</td>
<td>.33*</td>
<td>.10</td>
<td>.20</td>
<td>.22*</td>
</tr>
</tbody>
</table>

*a*The symbol * indicates a significance at the .05 level of confidence.

*bAll training grades after the 50-hour check were combined because these training phases were very short.
Table B.2
Correlations of Fixed Wing Check and Training Grades,
Fort Rucker, 1957-1958
(N = 100)

<table>
<thead>
<tr>
<th>Check Grade Intercorrelations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Check Grade</td>
</tr>
<tr>
<td>Advanced Progress Check-Contact (APC)</td>
</tr>
<tr>
<td>Basic Progress Check-Instrument (BPC)</td>
</tr>
<tr>
<td>Final Progress Check-Instrument (FPC)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Training Grade Intercorrelations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Grade</td>
</tr>
<tr>
<td>Up to Advanced Progress Check-Contact</td>
</tr>
<tr>
<td>Advanced Progress Check-Contact to Basic Progress Check-Instrument</td>
</tr>
<tr>
<td>Basic Progress Check-Instrument to Final Progress Check-Instrument</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Check and Training Grade Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Grade</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Up to Advanced Progress Check-Contact</td>
</tr>
<tr>
<td>Advanced Progress Check-Contact to Basic Progress Check-Instrument</td>
</tr>
<tr>
<td>Basic Progress Check-Instrument to Final Progress Check-Instrument</td>
</tr>
</tbody>
</table>

*The symbol * indicates significance at the .05 level of confidence.
Appendix C

ATTEMPTS TO STUDY INTEROBSERVER RELATIONSHIPS

Interobserver studies are traditionally an integral part of flight proficiency measurement research (19). A method was sought in this study for obtaining interobserver agreement data by placing two check pilots in the same aircraft, both evaluating the student's performance simultaneously. Unfortunately, neither H-23 nor H-13 helicopters, for which the PPDR's were being developed, were capable of safely carrying three people through several of the primary training maneuvers, particularly with a student pilot. The H-13H, with its more powerful engine, was used in an attempt to have two check pilots observe an instructor pilot who simulated student performance. However, the added weight substantially altered the performance of the aircraft during critical maneuvers such as autorotations, maximum performance take-offs, and steep approaches. Under high-density altitude conditions the performance of these primary maneuvers with two passengers, even by an expert pilot, approached being dangerous.

In order to study interobserver agreement, a helicopter (Cessna YH-41) somewhat similar in size and general configuration to the H-23 and H-13 and capable of carrying a pilot and three passengers, was obtained and attempts were made to adapt primary maneuvers to this aircraft. The flight characteristics of the YH-41 were sufficiently dissimilar to the H-13 and H-23 that quite different procedures were required to execute primary maneuvers. Had the project been continued, the results would probably have been applicable only to the YH-41. The YH-41 was experimental at that time and three successive mechanical failures terminated the investigation. Thus, initial attempts to obtain in-flight interobserver data failed.

If the efforts to obtain interobserver data had been successful, there would still have been the problem of obtaining a permanent, accurate, independent record of the actual performance. As interobserver efforts did fail, attempts to record actual student flight performance became even more important, particularly because of the need to allow for comparison of actual performance records with check pilot records. Prior research had successfully used a series of photographs of the instrument panel to obtain partial records of student performance (19). HumRRO research personnel attempted to adapt to the H-13 and H-23 helicopters a camera arrangement which would photograph the instrument panel and the horizon during flight at the same time that a check pilot applies the experimental PPDR's. This approach was unsuccessful at first because of inadequate knowledge of photographic techniques and
a shortage of time, personnel, and money. A method which did appear to work was developed too late to be included in the final data collection phase in the summer of 1958.

Had the photographic methods been successful, only about 25 percent of the check items could have been recorded, and approximately four hours per check ride would have been required of a trained clerk to translate filmed information into useful data. Because of budget limitations, this technique was not considered for further study.
Appendix D

VARIATION AMONG CHECK PILOTS IN SCORING THE 1958 PPDR's

Table D-1

Means and Standard Deviations of Percentages of Errors Scored by Check Pilots on Selected PPDR Items and of PPDR-Derived Scores, 1958*

<table>
<thead>
<tr>
<th>Item</th>
<th>Intermediate PPDR</th>
<th>Advanced PPDR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>PPDR Item&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pedals</td>
<td>10.7</td>
<td>7.2</td>
</tr>
<tr>
<td>RPM</td>
<td>20.6</td>
<td>13.4</td>
</tr>
<tr>
<td>Air speed</td>
<td>29.1</td>
<td>6.8</td>
</tr>
<tr>
<td>Altitude</td>
<td>16.8</td>
<td>5.4</td>
</tr>
<tr>
<td>Ground track</td>
<td>16.1</td>
<td>5.4</td>
</tr>
<tr>
<td>All Items</td>
<td>20.4</td>
<td>6.0</td>
</tr>
<tr>
<td>PPDR-Derived Score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item-weighted</td>
<td>84.6</td>
<td>5.1</td>
</tr>
<tr>
<td>Error pattern-weighted</td>
<td>70.2</td>
<td>6.5</td>
</tr>
<tr>
<td>Traditional&lt;sup&gt;b&lt;/sup&gt; (PPDR-based)</td>
<td>55.4</td>
<td>14.7</td>
</tr>
</tbody>
</table>

*Of the 12 check pilots, one was not available for the Intermediate PPDR analysis, and another was not available for the Advanced PPDR analysis. Thus, 11 check pilots are represented in each statistic in this table.

<sup>a</sup>These items constituted over half of the items on the PPDR's.

<sup>b</sup>Based on the percentage of "average" and "above average" grades given.
ACKNOWLEDGMENTS

The authors express their sincere appreciation to the military and Southern Airways civilian check pilots at the U.S. Army Primary Helicopter School at Camp Wolters, Tex., who worked with the HumRRO research staff during the summers of 1957 and 1958 in experimentally administering the Pilot Performance Description Records. Their criticisms and suggestions, as well as their willingness to submit their profession to careful examination, provided a substantial share of the basis for the work described herein.
# DISTRIBUTION LIST

<table>
<thead>
<tr>
<th>Location</th>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBI</td>
<td>DISTRIBUTION LIST</td>
<td>BEST AVAILABLE COPY</td>
</tr>
</tbody>
</table>

**Notes:**
- BEST AVAILABLE COPY
- DISTRIBUTION LIST

---

**Best Available Copy**

**DISTRIBUTION LIST**

- 100 REGIONS BIV HUMMO
- B3 REGIONS BIV HUMMO
- 2 BIV HUMMO, FT WICKER
- CBI, FT PISKIE
- 1 CBI, FT WICKER
- 2 REGIONS FT HUMMO
- 3 REGIONS FT HUMMO
- 4 REGIONS FT HUMMO
- 5 REGIONS FT HUMMO
- 6 REGIONS FT HUMMO
- 7 REGIONS FT HUMMO
- 8 REGIONS FT HUMMO
- 9 REGIONS FT HUMMO
- 10 REGIONS FT HUMMO
- 11 REGIONS FT HUMMO
- 12 REGIONS FT HUMMO
- 13 REGIONS FT HUMMO
- 14 REGIONS FT HUMMO
- 15 REGIONS FT HUMMO
- 16 REGIONS FT HUMMO
- 17 REGIONS FT HUMMO
- 18 REGIONS FT HUMMO
- 19 REGIONS FT HUMMO
- 20 REGIONS FT HUMMO
- 21 REGIONS FT HUMMO
- 22 REGIONS FT HUMMO
- 23 REGIONS FT HUMMO
- 24 REGIONS FT HUMMO
- 25 REGIONS FT HUMMO
- 26 REGIONS FT HUMMO
- 27 REGIONS FT HUMMO
- 28 REGIONS FT HUMMO
- 29 REGIONS FT HUMMO
- 30 REGIONS FT HUMMO
- 31 REGIONS FT HUMMO
- 32 REGIONS FT HUMMO
- 33 REGIONS FT HUMMO
- 34 REGIONS FT HUMMO
- 35 REGIONS FT HUMMO
- 36 REGIONS FT HUMMO
- 37 REGIONS FT HUMMO
- 38 REGIONS FT HUMMO
- 39 REGIONS FT HUMMO
- 40 REGIONS FT HUMMO
- 41 REGIONS FT HUMMO
- 42 REGIONS FT HUMMO
- 43 REGIONS FT HUMMO
- 44 REGIONS FT HUMMO
- 45 REGIONS FT HUMMO
- 46 REGIONS FT HUMMO
- 47 REGIONS FT HUMMO
- 48 REGIONS FT HUMMO
- 49 REGIONS FT HUMMO
- 50 REGIONS FT HUMMO
- 51 REGIONS FT HUMMO
- 52 REGIONS FT HUMMO
- 53 REGIONS FT HUMMO
- 54 REGIONS FT HUMMO
- 55 REGIONS FT HUMMO
- 56 REGIONS FT HUMMO
- 57 REGIONS FT HUMMO
- 58 REGIONS FT HUMMO
- 59 REGIONS FT HUMMO
- 60 REGIONS FT HUMMO
- 61 REGIONS FT HUMMO
- 62 REGIONS FT HUMMO
- 63 REGIONS FT HUMMO
- 64 REGIONS FT HUMMO
- 65 REGIONS FT HUMMO
- 66 REGIONS FT HUMMO
- 67 REGIONS FT HUMMO
- 68 REGIONS FT HUMMO
- 69 REGIONS FT HUMMO
- 70 REGIONS FT HUMMO
- 71 REGIONS FT HUMMO
- 72 REGIONS FT HUMMO
- 73 REGIONS FT HUMMO
- 74 REGIONS FT HUMMO
- 75 REGIONS FT HUMMO
- 76 REGIONS FT HUMMO
- 77 REGIONS FT HUMMO
- 78 REGIONS FT HUMMO
- 79 REGIONS FT HUMMO
- 80 REGIONS FT HUMMO
- 81 REGIONS FT HUMMO
- 82 REGIONS FT HUMMO
- 83 REGIONS FT HUMMO
- 84 REGIONS FT HUMMO
- 85 REGIONS FT HUMMO
- 86 REGIONS FT HUMMO
- 87 REGIONS FT HUMMO
- 88 REGIONS FT HUMMO
- 89 REGIONS FT HUMMO
- 90 REGIONS FT HUMMO
- 91 REGIONS FT HUMMO
- 92 REGIONS FT HUMMO
- 93 REGIONS FT HUMMO
- 94 REGIONS FT HUMMO
- 95 REGIONS FT HUMMO
- 96 REGIONS FT HUMMO
- 97 REGIONS FT HUMMO
- 98 REGIONS FT HUMMO
- 99 REGIONS FT HUMMO
- 100 REGIONS FT HUMMO

---

**Best Available Copy**
Improving Flight Proficiency Evaluation in Army Helicopter Pilot Training—George D. Greer, Jr., Wayne S. Smith, and Capt. James L. Hatfield (U.S. Army Aviation Human Research Unit, Fort Rucker, Ala.)


A method was devised for evaluating helicopter pilots' end-of-flight performance in primary helicopter training on the basis of a standard check ride evaluated with more objective measures. The measures—tuned the Intermediate PFDOR (Pilot Performance Description Record) and the Advanced PFDOR—consist of scales for the critical measures given in primary helicopter training, on which the check pilot can record his observations of each component of performance during the actual flight. The PFDOR system of evaluation was found to be more reliable and diagnostic than the method used in the past. In addition to the PFDOR booklet, the new system includes a training program for check pilots in the use of the PFDOR and classroom practice in scoring the PFDOR's for the correction of typographical errors of evaluation.