Handling Qualities and Pilot Evaluation

1984 Wright Brothers Lectureship in Aeronautics

Robert P. Harper Jr.
Arvin/Calspan, Buffalo, New York

George E. Cooper
G.E. Cooper Associates, Saratoga, California

Introduction

The handling qualities of airplanes have been a subject of considerable interest and concern throughout the history of piloted flight. The Wright Brothers were successful in achieving the first powered flight in large measure because they provided adequate flying qualities for the task at hand. As they became capable of flights of longer duration, they altered the handling qualities of their flying machine to improve piloting performance and to accomplish the expanded tasks. They maintained, throughout, a balance between the amount of stability (or instability) of their airplane in flight and the pilot's ability to control its movements; they achieved a balance among the airplane's stability, the airplane's controllability, and the pilot's capability.

"Handling qualities" represent the integrated value of those and other factors and are defined as "those qualities or characteristics of an aircraft that govern the ease and precision with which a pilot is able to perform the tasks required in support of an aircraft role." [1] From this definition, it is clear that handling quality is characteristic of the combined performance of the pilot and vehicle acting together as a system in support of an aircraft role. As our industry has matured in the 82 years since the first powered flight, the performance, size, and range of application of the airplane have grown by leaps and bounds. Only the human pilot has remained relatively constant in the system.

In the beginning, the challenge was to find the vehicle which, when combined with the inexperienced pilot-candidate, could become airborne, fly in a straight line, and land safely. Longer flights required turns, and ultimately there were additional tasks to be performed. As greater performance capability was achieved, the airplane was flown over increasingly greater ranges of altitude and speed, and the diversity of tasks increased. The continuing challenge was—and still is—to determine what characteristics an airplane should have so that its role can be achieved in the hands of a relatively constant pilot.

This challenge has been difficult to answer; the problem is that the quality of handling is intimately linked to the dynamic response characteristics of the airplane and human pilot acting together to perform a task. Since the pilot is difficult to describe analytically with accuracy, we have had to resort to experiments in order to experience the real system dynamics.

A further problem is in the evaluation process, the judging of differences in handling qualities. One would expect to instrument the aircraft, measure the accuracy of task performance, and separate good from bad in that way. The human pilot, however, is adaptive to different airplanes; he compensates as best he can by altering his control usage for differences among airplanes as necessary to accomplish the task. This compensatory capability makes task performance relatively insensitive to differences among airplanes, but at the same time, heightens the pilot's awareness of the differences by altering the total workload required to achieve the desired task performance.

The airplane designer, then, is presented with a formidable task. He must design an airplane to be of good dynamic quality when operated by an adaptive controller who resists analytic description. Experience—if carefully documented and tracked—is helpful, but the rapidly changing technology of flight (and the nomadic nature of the industry engineers) causes at least some part of each new aircraft development to break new ground.

It is the purpose of this paper to examine this subject of handling qualities, defining first the dynamic system and discussing its constituent elements. Next, some historical perspective is introduced to illustrate that the quest for good handling has continued to be a challenge of substantial proportions from the Wright Brothers' beginning to the present day. The most modern methods of evaluating handling qualities place heavy emphasis on simulation and evaluation through experiment. The techniques, practice, and consideradons in the use of pilot evaluation are reviewed; recommendations are made which the authors believe would improve the quality of the evaluation data and the understanding of the pilot-vehicle system.

The Pilot-Vehicle Dynamic System

Fundamental to the subject of handling qualities is the definition of the system whose handling is to be assessed. Aircraft and flight control designers often focus on the dynamics of the vehicle, since that is the system element whose characteristics can be selected—the pilot is not readily alterable. The piloted-vehicle dynamics, however, are very much affected (and set) by the pilot's actions as a controller; he is a key element in the system. In the functional diagram of Fig. 1, the pilot's role is delineated as the decisionmaker of what is to be done, the comparator of what's happening vs what he wants to happen, and the supplier of corrective inputs to the aircraft controls to achieve what he desires. This, then, is the system: the pilot and aircraft acting together as a closedloop system, the dynamics of which may be significantly different from those of the aircraft acting without him—or open loop, in the context of acting without corrective pilot actions. For example, the aircraft plus flight control system could exhibit a dead-beat (no overshoot) pitch rate response to a pitch

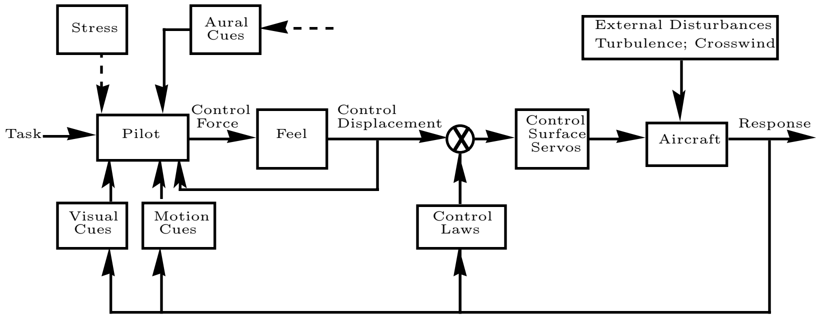


Figure 1: Pilot-vehicle dynamic system.

command input, and yet be quite oscillatory in the hands of a pilot trying to land. For this case, the pilot would say the handling qualities are poor in the landing task. Many engineers would look at the open-loop response, however, judge it to be of good dynamic quality, and expect that it would land well.

Note that the engineer in this case is basing his judgment on the observation of the dynamics of only part of the system. Here is the source of a fundamental problem in the study of handling qualities: the complete dynamic system is seen only by the pilot. Normally, the pilot sees only the complete system (unless, as a test pilot, he puts in special control inputs and observes the aircraft response). Normally also, the engineer sees only the aircraft response. So when they talk about the quality of the dynamic response, the pilot and engineer are often referring (at least in concept) to the characteristics of two different systems. Little wonder that the two groups sometimes have difficulty understanding each other!

Looking further at the system diagramed in Fig. 1, one sees other system elements that affect the pilot's actions as a controller, and therefore affect the closed-loop dynamics. The cockpit controls, for example: the forces, the amount of movement, the friction, hysterisis, and breakout—all are different in different aircraft, and each has an effect on the system dynamics. An experiment of interest to both pilots and engineers is to allow the pilot (or engineer) to attempt a specific task such as acquiring and tracking a ground target in a shallow dive for the same aircraft with different pitch control force gradients. As compared to the handling with a reference force gradient, the evaluator will describe the aircraft response with lighter forces as quicker and with some tendency to overshoot (or oscillate) when acquiring the target. With a heavy force gradient, he describes the response as slower, even sluggish, but with no overshoot tendency. The oomments describe dynamic differences for a change made to the statics of the open-loop aircraft, and are explainable in terms of the closed-loop pilot-aircraft system—much as the dynamics of aircraft plus flight control system are altered by changing the system gains.

In a similar way, one can appreciate that changing the amount of cockpit controller displacement can affect the ability to track a target—small displacement gradients evoking comments about quick response but a tendency to overshoot, larger gradients bringing complaints of sluggishness but attendant steadiness. This is understandable in terms of the closedloop pilot-airplane system but most puzzling in the context of the airplane-alone dynamic response.

The effects of the task can be illustrated in Fig. 1 by altering the parameters (feedbacks) that the pilot employs in making the system produce the desired response performance. Using pitch rate and normal acceleration to maneuver can be expected to generate different dynamics as compared to using pitch angle and pitch rate to acquire and track ground targets. Thus the handling qualities of a given aircraft are taskdependent: what is preferred for one task may be less desirable for another. An example here is the effect of 150 ms of pure time delay in pitch: not really noticeable in aggressive maneuvering with an otherwise good airplane response, but it causes a substantial pilot-induced oscillation (PIO) when the pilot tries to stop and hold the nose on a selected point.

The cockpit displays can have major effects on the handling qualities for task by affecting what is displayed (and therefore controlled). Display sensitivity (in terms of symbol movement per parameter variation) affects the pilot-aircraft dynamics, and so also do display dynamics (especially time delays). The effects of computational and other delays in the presentation of the visual scene may generate significant differences between ground simulation and the real world.

Consider now the pilot himself. His role in the system is all important; what do we know about him as a system element? He is certainly a product of his background, training, and past experience, and each affects his controller capabilities. He is affected by health and stress level. In general, although pilots tend to view themselves as individualists, their behavior as elements in the pilot-aircraft system are sufficiently homogeneous that we can create aircraft to carry out missions and count on all trained pilots to be able to fly them without individual tailoring. In fact, the process of selecting pilots (physical, mental, and educational standards) and training them (training syllabi, check rides, license standards) tend to strengthen their commonality. This is not to say they are all equally good controllers: it does postulate that what is designed to be good handling with one pilot is not likely to be significantly worse with another.

Since the pilot's characteristics as a controller are adaptive to the aircraft, his capabilities are significantly affected by training and experience. One finds that it may take a while for a pilot trained to handle a tricycle-landing-gear airplane to find, learn, or re-learn his rudder technique and avoid groundlooping a tail-dragger, but learn it he does. How long it takes may vary greatly among pilots, but the final achievement exhibits much less variability.

The depiction of the pilot-aircraft system is useful when comparing ground simulation to actual flight. One readily sees the potential effect of the presence (or absence) of motion cues as affecting the presence (or absence) of a feedback path in the pilot-aircraft system, with the resultant effect on system dynamics. The moving-base ground simulator introduces attenuation and washout dynamics into the motion feedback paths; the fixed-base ground simulator eliminates all motion feedback; both alter the dynamic result as compared to the real flight case, although less so with simulators having extensive movement capability. It is essential to know when compromises in or lack of motion fidelity may affect evaluation results. In some cases, a test pilot will recognize the disparity and provide guidance in interpretation of results, but the final answer in many cases will only be known when ground and flight evaluations are obtained for the same aircraft and task.

This system of pilot, aircraft, flight control system, and displays is indeed complex and intricate. It is inherently difficult to analyze in the precise equations of the engineering community and, in fact, the subject of handling qualities and the pilot-vehicle dynamic system is largely ignored in most of our universities. Consequently, we are continuing to produce engineers for the industry who have little education and training in this subject. A commendable exception is the curriculum at the United States Air Force and Naval Test Pilot Schools, where pilots and engineers receive both ground and in-flight instruction and experimental experience in this subject.

To better appreciate the development of the subject of handling qualities and pilot evaluation, let us next review their historical development.

Historical Perspective

The record of aeronautical progress shows a consistent pattern in which changes in aircraft design to increase airplane performance have led to handling qualities problems; the solutions have, in turn, frequently created new problems. The application of new technology, new mission demands, and associated tasks, conducted under a variety of environmental conditions, have also added their share of problems. The problems, their solutions, and the ongoing need for handling qualities criteria to avoid the problems have posed a continuing challenge to handling qualities specialists and the test pilots with whom they work.

The record also shows that each generation of an aircraft has become increasingly useful and productive. It can be expected that demands for increased productivity will persist in the future and these teams will face new problems that are similar to the past in nature if not in detail. It would therefore be instructive to trace the historical record, placing less emphasis on the obvious airplane design changes that produced the problems, and focusing more on the general character of the handling problems and on the manner in which the test pilots conducted the evaluations that contributed to their solution.

From such a review, we may gain some insight into any changes that may have occurred in the test pilot's role and whether the tools and techniques employed by him in the assessment of aircraft handling qualities have kept pace with the demands created by advanced and often complex systems. The reader is cautioned, however, that the review is conducted to set the stage for the material which follows, and is by no means exhaustive or complete; it is limited by the knowledge of the authors and their particular contact with the world of handling qualities. Certainly, other work of significance that is not mentioned was performed here and in other countries; other nations were facing similar problems and generating similar solutions and contributions.

Balancing Controllability with Instability

The Wright Brothers and Their Airplane

Two fundamental problems had to be solved for man to fly: the performance of his vehicle had to be adequate to become and remain airborne, and the control had to be adequate to maintain attitude, adjust flight path, and land in one piece. Lillienthal solved part of the problem for gliders. The Wright brothers were the first to do it for powered flight.

The success of the Wright brothers was due in no small measure to the attention they paid to controllability. They discovered that longitudinal balance could be obtained through a horizontal rudder and lateral balance achieved by changing the spanwise lift distribution. [2] The state of the art with respect to handling qualities was understandably rudimentary. For the first flight, to barely maintain control was a difficult enough goal in itself. They overcontrolled most of the time, and conducted many experiments and trials to improve controllability and handling after that initial flight. Working side by side during design and development, and sharing the roles of flight test engineer and test pilot, Wilbur and Orville avoided problems of communication that would in later years mark the interdependence of these roles.

Later, in attempting to market their airplane to the government, they had to meet a specification relating to handling qualities that was rather vague but appealing in its objectivity; to wit, “It should be sufficiently simple in its construction and operation to permit intelligent men to become proficient in its use within a reasonable length of time.” [3] This vagueness contributed to difference in interpretation that set the stage for conflict and negotiation between vendor and user in a pattern that would be repeated again and again for many years to come.

The Wright brothers’ success was especially remarkable when one reviews some of the problems they encountered. Early tests in 1902 revealed that “this combination...cathedral angle in connection with fixed rear vertical vanes and adjustable wing tips was the most dangerous used...” [2] This ultimately led them to a movable vertical vane and then to an interconnect between the rudder and the wing warping cables. By 1905, they had adopted wing dihedral and made rudder control independent of wing warping.

Perkins [4] noted that “in today’s language, the Wrights were flying longitudinally unstable machines with at first, an overbalanced elevator. They were flying machines that had about neutral directional stability, negative dihedral and with a rolling control that introduced yawing moments... From their own descriptions of their flights, they were overcontrolling most of the time, and many flights ended when at the bottom of an oscillation, their skids touched the sand.”

During early flights, atmospheric turbulence or windshear were encountered on several occasions, causing their airplane to suddenly roll or drop—in one case a distance of 10 ft to land flat on the ground. In considering a solution, it was stated, "The problem of overcoming these disturbances by automatic means has engaged the attention of many ingenious minds, but to my brother and myself, it has seemed preferable to depend entirely on intelligent control." [2]

World-War-I-Era Airplane Characteristics

Based on evaluations of the handling qualities of several World War I airplanes which were recently rebuilt and flown in the United States and United Kingdom, we gain the impression that any increases in performance that were achieved in this period were not matched by improved handling qualities.

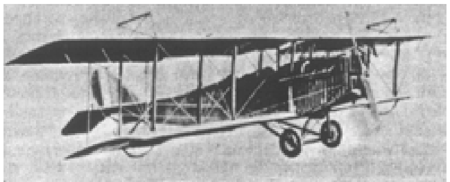
Two well known airplanes in this category in the United States are the Curtis JN-4 "Jenny" and the Thomas-Morse S.4C (Fig. 2). Nissen [5] has noted that both were unstable longitudinally and directionally. Pilots had to learn to fly them by reference to pitch attitude, as force feel was unrealiable. In flying the Jenny, there was a distinct tendency for pilots to overcontrol in pitch. The saying was, "Don't fly it; drive it." The Jenny was flown by the "wind-in-face" technique for directional control. The Morse was even more unstable, requiring a forward push on the stick in turns and exhibiting a tendency for overcontrolling with the rudder. Landings were not difficult because of high drag and a short landing roll, but might require a push instead of a pull on the stick during the flare and touchdown.

In the United Kingdom, the Sopwith Camel (Fig. 2) was reported to be an accurate replica of the original airplane but without the "awesome gyroscopic effects of the rotary engine and its enormous propeller." [6] The report states in part:

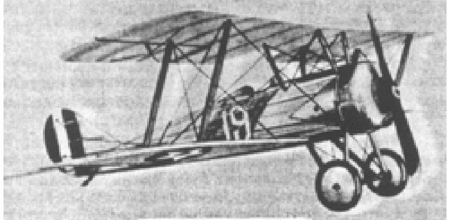
"Once in the air...the pilot is faced with almost total control disharmony. The Camel is mildly unstable in pitch and considerably unstable in yaw, and both elevator and rudder are extremely light and sensitive.. .the ailerons are in direct and quite awe-inspiring contrast. The Camel.. .has four enormous...barn doors [for ailerons] which require an equally enormous force to be moved quickly. And when you have moved them, the wing section is so degraded... that the roll response is very slow indeed... . At the same time, aileron drag is quite staggering. If you take your feet off the rudder bar and bank to the left, the Camel will instantly yaw sharply to the right and keep going... ."

The Bristol Boxkite (Fig. 2) is a pre-World War I airplane which is similar in appearance to the Wright Brothers' airplane.

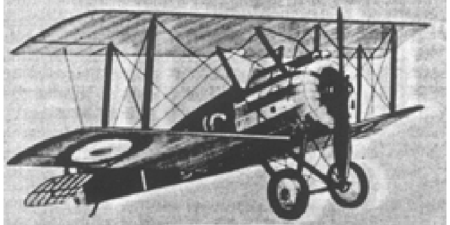
"...At 45 mph, tailplane lift overcomes the combined power of foreplane and elevator and the machine is now intent on a downward outside loop. This actually happened in the old days, and the aviator who was not strapped in fell into the underside of the top wing. The machine completed a half outside loop, stalled inverted in the climb and entered an uncontrolled inverted falling leaf with stopped engine.. .The only good thing about the whole story is that the



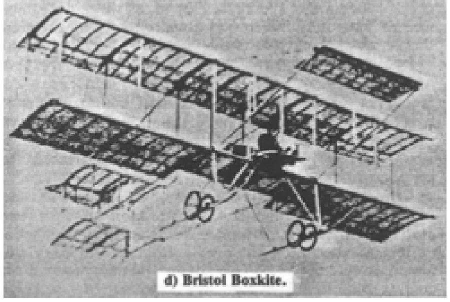
a) Curtiss JN-4.



b) Thomas Morse S.4C.



c) Sopwith Camel.



d) Bristol Boxkite.

Figure 2: World War I era aircraft.

Boxkite fluttered down and disintegrated, but so slowly that the aviator was completely unharmed.” [7]

These descriptions of the World War I airplanes are extracted from the commentary of pilots who built and flew replicas later, not the original test pilots. What the original test pilots said is not known, but is must be assumed that the role of the test pilot in this era was limited to completing the test flight safely so that he could report on the characteristics he observed. How thoroughly and lucidly he reported is not known, but one can infer that in many cases the designer didn’t have much solid information on which to base his attempts at improvements. Realize, also, that there were no data recording systems to show the designer what the airplane characteristics actually were in comparison to what he thought they might be; he had only the pilot’s comments and his own ground observations. Nor was there any organized framework of dynamic theory from which he could have evaluated the design significance of quantitative recorded data.

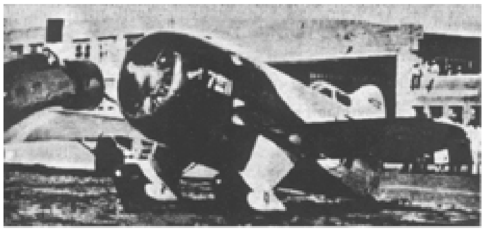
Quest for Stability

One encouraging development in the overall picture of events in this era was the recognition of the need for scientific research relating to airplane design. In the United States, the National Advisory Committee for Aeronautics, NACA, was established in 1915, with its first laboratory located at Langley Field, Virginia.

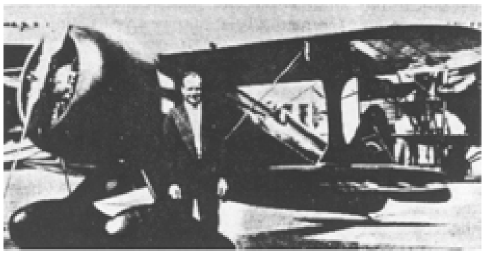
This period saw the early development of the wind tunnel and the application of flight research, largely for wind tunnel correlation. One of the first flight tests for stability and control was performed in the summer of 1919 by Edward Warner and F.H. Norton [8] at Langley Field with Edmund T. “Eddie” Allen, the sixth Wright Brothers Lecturer, doing most of the flying. While the need for special research pilots who would be qualified to make appropriate observations had been noted in 1918, it was not until 1922 that flight tests for handling qualities appear to have started. Early flight research concentrated on the test pilot’s observations of deficiencies in stability and control, stalling and spinning characteristics, and takeoff and landing performance.

Early Flight Research

During the 1920’s, enormous public interest in aviation was generatee by the exploits of early racing and barnstorming pilots and by Lindbergh’s flight across the Atlantic. This was a period during which NACA was building its flight research capabilities and evaluating a variety of aircraft to accumulate results that could be generalized. The techniques used relied largely on the subjective opinions of the test pilots, which were written down and formed a data base. Originally, instrumentation used to document the aircraft characteristics was primitive, with the test pilot relying largely on his kneepad, stopwatch, and a spring balance to measure control forces. However, NACA soon developed



a) GeeBee R-1.



b) Laird Super Solution (Doolittle).

Figure 3: Racing airplanes.

photographic recording instrumentation which enabled accurate documentation of aircraft stability and control characteristics for correlation with pilot opinion.

Early Racing Plane Characteristics

Indicative of the kind of handling qualities that characterized airplanes of this era is General Jimmy Doolittle's report [9, 10] on his experience with the GeeBee and the Laird racers (Fig. 3). He describes the GeeBee Racer as both longitudinally and directionally unstable. With its small vertical tail surface, which was blocked out at small angles of sideslip, response to gusts would excite a directional instability that was particularly bad during landing; during high-speed flight, the directional instability could be handled fairly easily. One landing experience was related in which, with the aircraft almost ready to touch down, the rudder was kicked, which started a directional oscillation that was impossible to control. A quick burst of power, however, straightened the aircraft out, and a safe landing was completed. A larger rudder and vertical fin were added, which improved the directional stability characterstics considerably; however, Gen. Doolittle relates one incident where he was making a simulated pylon turn at 4,000 ft and, during the entry, sufficient sideslip developed that the aircraft executed a double snaproll.

The Laird Racer had insufficient wing cross-bracing which permitted the wing to twist in flight. At any time, if the angleof-attack were changed, the aircraft would tend to diverge. It was found that a hard bang on the stick laterally would adjust the wing alignment and place it back in balance. One must marvel at the approach a test pilot must take toward his airplane that would enable him to find that kind of solution to a problem.

In the aeronautical community at large, the test pilot's role was developing a more professional nature. However, many of the handling qualities improvements were obtained by "cut and try" methods, and not all pilot evaluations were conducted with expertise and objectivity. For example, one contract test pilot was known to guard against revealing his kneepad observations to the engineers but insisted on interpreting his observations himself and conveying only what he thought the solution or "fix" might be. This is a rather extreme example of poor communications between test pilot and engineer. Engineers and designers who were misled by such tactics developed an appreciation for the type of test pilot who objectively and accurately reported his observations before attempting to recommend solutions.

Stability, Control, and Open-Loop Dynamics

Documentation and Criteria Development

The airplane designs of the late 1930's began the reflect benefits from an increased understanding of static stability and, to a much lesser extent, dynamic stability, emanating largely from NACA flight research. Contributing to this understanding were the data obtained by correlating pilot opinion ratings of observed stability characteristics with measured characteristics for a variety of different airplanes. An example of the class of data being obtained in this period is shown in Table 1 from NACA Report 578, [11] published in 1936. Test pilot observations of the dynamic (phugoid) and static stability of eight different aircraft are summarized here. The research pilots were asked to evaluate static stability in terms of "stiffness" and factors affecting stiffness, using a rating scale of A to D, in which A corresponded to the greatest stiffness. Dynamic stability was merely noted as stable (s) or unstable (u) for various aircraft conditions.

These data are of interest in showing the qualitative (rather than quantitative) nature of initial attempts to formalize the use of subjective pilot assessments as a means of developing handling qualities criteria. More refined developments of this technique took place in later years.

Meanwhile, the new airplanes were found to fly reasonably well. In a pattern that persists to this day, problems and deficiencies in handling qualities often tended to be identified with the fringes of the operating envelope, and with unusual mission requirements and environmental conditions.

Spurred by strong support from NACA management, flight research relating to handling qualities was accelerated. Allen of Boeing and NACA test pilots, collaborating closely with engineers, formulated specific test maneuvers to acquire data that could be used in conjunction with subjective pilot opinion to form design critera. [12, 13] The process was aided by the newly developed photographic recording instrumentation that provided more precise quantitative flight data for analysis. The resulting flight test procedures and evaluation techniques used by NACA were then used by military test pilots as part of the evaluation and acceptance of their own aircraft. Using Warner's tentative checklist [14] as a starting point, Soule [15] and Gilruth [16] and their staffs quickly translated the rapidly accumulating data from flight tests of a variety of airplanes into more refined criteria. Probably the first effort to set down a specification for flying qualities was performed by Edward Warner when asked by Douglas to do so for the DC-4. [4]

Documentation of the handling characteristics of available aircraft continued throughout this period, aided by the continuing development of improved techniques for assessing airplane handling qualities. Deliberate efforts were made to provide opportunities for test pilots during this and following periods, to fly and evaluate a variety of different aircraft; this farsighted policy was invaluable in enabling them to develop objectivity and overcome the biases which occur when background and experience are limited.

As a result of the combined contributions of flight data documentation and assessment, wind tunnel studies, and theoretical analysis, the first comprehensive sets of military flying qualities specifications were issued by the Navy Bureau of Aeronautics [17] in 1942 as NAVAER SR-119 and by the U.S. Army Air Force in 1943 as AAF-C-1815. [18] These specifications reflected the flight test procedures used by test pilots in documenting the "open loop" response of airplanes, supplemented by subjective assessment of "feel" characteristics obtained through "pilot induced" maneuvers. They confirmed the fact, noted by Allen [13] that "flight testing was becoming a more exact science, combining accurate quantitative data with the pilot's qualitative report." This "aerodynamic" approach to flight testing "dealt largely with the aerodynamic forces and moments acting on the airplane."

World War II HandlIng Problems

During World War II, combat planes were being pushed to the limits of their operational envelopes and beyond. Adverse effects of compressibility were encountered, as sonic speeds developed over the airfoil sections during high-speed dives. A major effort was made to understand and correct the handling deficiencies that resulted from heavy "tuck-under" control forces, aeroelastic control reversal effects, and buffeting, to mention only a few. This effort, mounted largely by NACA and the military services with industry support, established a pattern of research effort that utilmately resulted in penetration of the transonic barrier.

After the war ended, the stability and control problems and other deficiencies in handling qualities encountered during the war years further stimulated aeronautical research and development. As a result, immediately afterwards technological breakthroughs were made that contributed to penetration of the sonic barrier but left unanswered the question of which of several aircraft designs capable of enabling supersonic flight would be best—swept wings, delta wings, or low-aspect straight wings. It was obvious to the military that pilots would need this supersonic capability for the next generation of combat aircraft, but it was not obvious which of the several wing designs and configurations that could provide supersonic performance would be best overall for the mission.

An exploratory period followed in which all three wing planforms; including some without tails, were incorporated into airplane designs. A new series of handling qualities problems was exposed as all types were investigated and research was conducted to find solutions. These solutions frequently involved the use of more complex control systems (e.g., powered controls and stability augmentation systems). And these fixes were not without problems of their own that taxed the engineer/test-pilot team.

Flight Simulation and Closed-Loop Dynamics

Simulator Development and Use

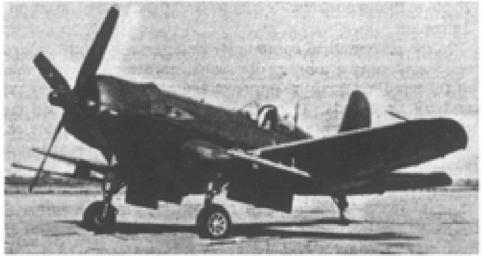
While not as glamorous or stimulating to the test pilot as flying and evaluating the "compressibility problem" or the new generation of aircraft capable of transonic flight, new technology was developing in this period which was to have a dynamic and revolutionary impact on the role of the test pilot. This was the development of the variable-stability airplane and advances in computer technology. Modified Grumann and Vought World War II fighters were independently developed into variable-stability research aircraft (Fig. 4) at NACA's Ames Aeronautical Laboratory (AAL) and at Cornell Aeronautical Laboratory (CAL), respectively, the latter with U.S. Navy support. Subsequently, taking advantage of advances in computer and servo-mechanism technology, follow-on variable-stability aircraft become known as in-flight simulators.

These developments had a dramatic impact on handling qualities research and the development of handling qualities criteria. The variable stability aircraft were quickly put to use: the NACA F6F [19, 20] and the CAL/Navy F4U [21] each examined the effects of different lateral-directional response characteristics. Their longitudinal handling qualities were not variable. Variations in longitudinal handling qualities were made possible by the conversion to variable-stability airplanes of a Douglas B-26 [22] in 1950 and a Lockheed F-94A [23] in 1953 by CAL for the USAF, and a YF-86-D by NACA Ames in 1957. [24] NACA Ames and CAL [25] also developed separate lateral-directional variable-stability F-86E airplanes. The development of the first all-axis variable-stability airplanes included a Twin-Beech in 1950 and a Lockheed NT-33 in 1957 by CAL for the USAF, and a North American F1OO by NACA.

These airplanes, aside from being the first of a new breed called "fly-by-wire," employed power control systems and analog computer electronics to permit the systematic variation of the airplane's static and dynamic response characteristics. Different sets of response characteristics were set up in actual flight, and the pilot performed maneuvers to assess the quality of handling. The results of these experiments were reported, and so began what was to become a growing body of systematic in-flight simulation experiments, the results of which would form the basis of a new set of flying quality specifications for military airplanes.

Later, important research developments in in-flight simulation were the Tn-

a) NACA F6F.



b) Navy/CAL F4U.

Figure 4: The first variable-stability airplanes.

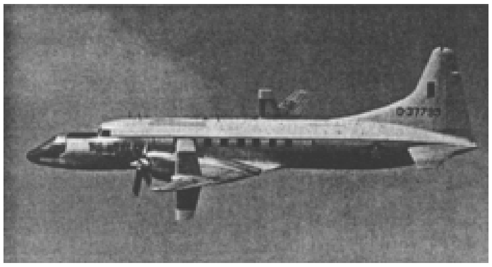


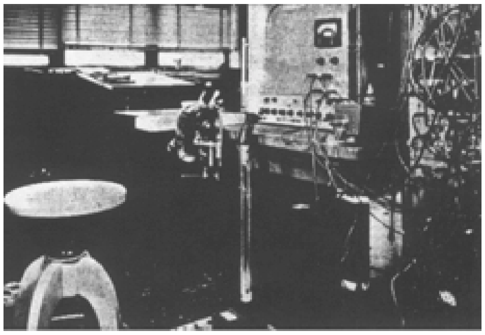
Figure 5: USAF C-131H Total In-Flight Simulator (TIFS).

Service (later Navy) X-22A and the NASA Ames X-14 V/STOLs; the USAF C-131H Total In-Flight Simulator (TIFS); the Calspan Learjet; the Lockheed Jetstar, developed by CAL for NASA Dryden; the Boeing 707-80; the Canadian National Research Council Bell 47G and Jet Ranger, and NACA Langley CH-47 variable stability helicopters; the two Princeton University Navions; the French Mirage 3; the German HFB 320; and the Japanese Lockheed P2V-7.

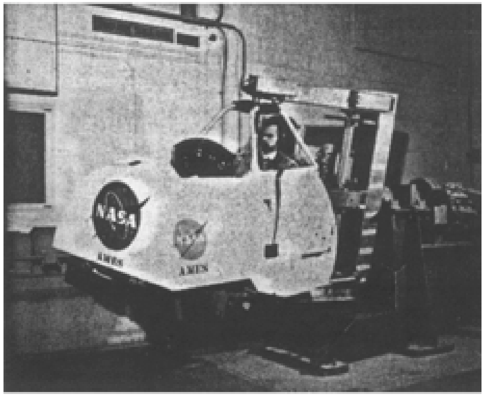
In addition to their uses for research in handling qualities, the in-flight simulators took on a role of simulating real aircraft designs prior to first flight. The list of aircraft that have been simulated is too long to document here, but it is impressive both in diversity and number, and continues to grow. The aircraft include several from the X-series, lifting bodies and the Space Shuttle, fighters, bombers, V/STOL, transports, and trainers; their flight-control systems range from simple mechanical to high-order, complex digital. Generally, the physical appearance of the in-flight simulators was vastly different from the aircraft simulated, and onlookers often questioned the wisdom of believing the evaluation results. The proof of the technology came, however, with the first flight of the real airplanes: time after time, the test pilots attested to the close handling qualities correspondence of in-flight simulation to the real airplane. This is a tribute not only to in-flight simulation technology, but also to the accuracy with which our industry can predict the parameters of the analytical models upon which all simulation depends.

A current USAF in-flight simulator is shown in Fig. 5. Development of a new fighter in-flight simulator called VISTA (Variable-stability In-flight Simulator Test Aircraft) is being planned by the USAF to enter research use in 1990.

Shortly after the introduction of the airborne simulators, it became evident that advances in analog computer technology made ground-based simulators technically feasible, and NACA Ames embarked on research programs using this research tool. In contrast to the early Link trainer, the flexibility and capability of these devices for enabling pilots to conduct stability and control evaluations systematically was recognized early. The first ground simulators (Fig. 6), however, provided little more than the opportunity for pilots to examine variations in specific stability and control parameters in simple maneuvers, and this often under conditions of less-than-perfect responses. These simulators were far removed from having the capability to enable pilots to fly more complex



a) Rudimentary fixed-base.



b) Early moving base.

Figure 6: Early research ground simulators.

mission-oriented tasks. This situation ushered in a new era in demands on the pilot. Faced with these ground simulator limitations, pilots had to call upon all of their background and experience to extrapolate their observations to the real world flight situation. They had to identify and separate deficiencies and limitations in the simulator from those of the simulated airplane. To do so required sufficient familiarization to enable the pilot to adapt to the simulation. Even when this was provided, the demand for pilot extrapolation was often so great that at times the pilot's confidence in his evaluation results was compromised. It became even more important that the test program objectives be consistent with what reasonably would be expected from the pilot.

The accelerating trend in reliance on ground-based simulators emphasized the importance of developing improvements in fidelity and capability to overcome their other limitations, an objective to which NACA, the military services and industry exerted strong efforts. As simulator designs improved, the need for pilot extrapolation diminished and a welcome increase in confidence in simulator results occured. A modern NASA research simulator is shown in Fig. 7.

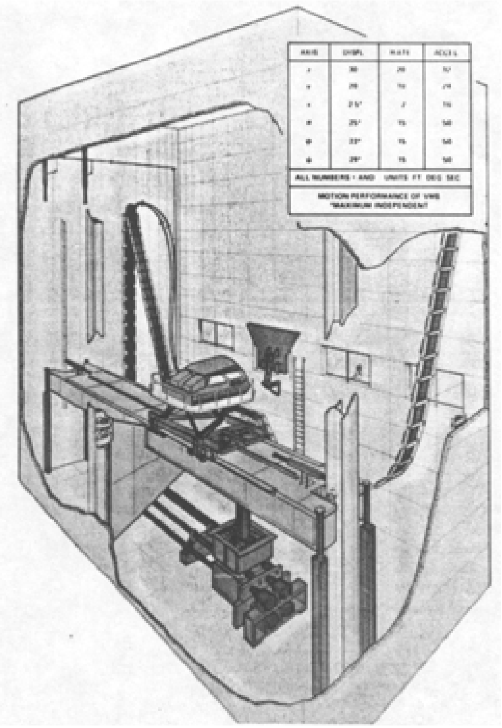


Figure 7: Modern 6-DOF motion-based research ground simulator (NASA Ames VMS).

The industry test pilot found himself more than ever before involved in the subjective assessment of new aircraft designs through flight simulation, because those aircraft manufacturers who were involved in some aspect of the pilot-vehicle interface found it increasingly desirable to employ ground simulation in their design and development process.

During the late 1940's and the 1950's, a substantial amount of experimental research was conducted and reported, which led to a greatly improved understanding of airplane handling qualities. Much of this work focused on the use of the ground and in-flight simulators to set up various sets of airplane response characteristics, with pilots carrying out several tasks to evaluate the handling quality. Engineers and pilots, often working together, then correlated the comments and ratings with the airplane parameters that were varied. From these efforts, it was learned, for example, [22, 23] that the longitudinal short period frequency could be made so high as to be objectionable—where previously engineers had expected that faster responses were always better. The evaluation pilot's description of the high-frequency, well-damped cases as "oscillatory, especially when tracking the target" led engineers to conceptually model the piloted airplane as a closed-loop system, with the pilot as an active dynamic controller, in order to explain the pilot comments. The closed-loop concept led engineers to attempt the analytic modeling of the pilot, so that analysis could be used to predict piloted-airplane behavior in new, untried situations. The early modeling efforts were beneficial more for the understanding which they imparted than for the hard, accurate handling qualities data that they produced; but the impact which they had was profound and valuable, and it shaped the direction of much of the experimental work in handling qualities.

Quantification of Pilot Opinion

The combined capability of these ground and in-flight simulators provided engineers with very powerful tools with which to condtsct systematic evaluations of airplane handling qualities. They emphasized even more forcefully, however, the need for improved communication between pilot and engineer; particularly as any simulation, no matter how sophisticated, will differ in some way from the actual vehicle and a real-world environment.

A common language with defined terminology was needed. For each set of handling qualities evaluated, some means for quantifying the pilot's overall assessment was required, and pilot rating scales were developed. Interpretation of pilot ratings, the terminology used, and statements regarding quality of observed response varied widely throughout the aeronautical community. Definitions of the terms used and some early pilot rating scales were developed and used independently at CAL and AAL.

The first widely used rating scale was introduced in 1957, along with a discussion of the subject "Understanding and Interpreting Pilot Opinion." [15] Subsequent efforts to correct deficiencies in this "Cooper Scale," as it was referred to, were spurred by interest abroad and in the United States and resulted in an interim revised scale being introduced in 1966. [26] A final version of this Cooper-Harper rating scale was published in 1969 as NASA TN D-5153. [1] Frank O'Hara, then Chief Superintendent, RAF, Bedford, was particularly helpful in stimulating this collaborative effort to standardize terminology and definitions, and to develop a standard rating scale suitable for international use. [27]

As the quantification of pilot opinion developed, the industry experienced a period of decreased funding in aeronautics in deference to missiles and space. A direct result was an effort to accomplish more R&D through flight simulation, which meant increased participation by evaluation pilots. Test-pilot training was augmented by the introduction of CAL's in-flight simulators into the curriculum of the Navy and Air Force Test Pilot Schools in 1960 and 1963, respectively. Ground simulation followed.

Criteria from Piloted Evaluations

The military flying specification, AAF-C-1815, gave way in 1954 to a new version, MIL-F-8785, [28] which incorporated some of the data from early piloted simulation experiments. These data bases were growing very rapidly, however; and in 1969, Chalk and associates at CAL, under contract to and working closely with Charles Westbrook of the USAF Flight Dynamics Laboratory, formulated a new, almost revolutionary verison of the specification. This version, adopted as MIL-F-8785B, [29] incorporated the pilot-in-the-loop research results, formulated concepts of flying quality levels of desirability based on the Cooper-Harper rating scale, aircraft states (including failure states), performance envelopes, and other concepts to deal with the new world of flight control augmented airplanes. Another innovation of MIL-F-8785B was an extensive report giving "Background Information," which documented the basis and supporting data upon which each requirement was based. [30] A revised version, MILF-8785C, was issued by the USAF in l980 [31] and is currently in use while another revision is in preparation.

Electronic Flight Control Systems

As we look to the future, it appears that electronic flight control and cockpit display system technology should be capable of providing whatever the pilot desires. However, we hear much about the development problems of the sophisticated electronic flight control systems in the YF-16, YF-17, F-18, and Tornado fighter aircraft, as well as in the Space Shuttle, in that they exhibited serious and, in some cases, dangerous handling quality problems (pilot-induced oscillations, PlO) which were not predicted by analysis or ground simulation. This subject has been examined by a number of handling qualities specialists [32, 36] and a common problem seems to have been excessive time delays introduced through the design of complex higher-order fly-by-wire (FBW) control systems. A contributing factor was the unrecognized deficiencies in some simulators or in the simulation experiments and interpretation of results. A number of important lessons have already been drawn from these experiences; those directly related to handling qualities and pilot evaluation are summarized here.

1. Flying qualities criteria have not kept pace with control system development, in which high-order transfer functions have introduced undesirable delays.

2. Deficiencies in simulators are often not recognized, thereby adversely affecting the results.

3. The development process should include both ground and airborne flight simulation.

4. Inadequate communication among the various engineering specialties, management, and test pilots has seriously compounded the problems of conducting sophisticated simulation experIments.

It becomes increasingly apparent that successful application of new technology is critically dependent upon the test pilot's evaluation of handling qualities and the tools (simulator facility and experiment design) with which he and the project engineers have to work.

There is certainly nothing in these lessons that cannot be applied to the next generation of aircraft, if the industry heeds them. It is of interest to note that the next generation European Airbus, the A-320, plans to adopt new technology that has not yet been incorporated in a civil transport. [37] In addition to an advanced fly-by-wire control system, without mechanical back-up, the A-320 will replace the conventional control column with side-located hand controllers and will employ a full panel of electronic displays. The success of this development would appear to rest, in part, on their approach to the above lessons.

We have noted several significant lessons to be applied to the science of flight simulation. We will next consider guidance and recommendations for engineers and pilots in the conduct of handling quality evaluations.

Methods for Assessing Pilot-Vehicle System Quality

AnalytIcal

All of the elements shown in Fig. 1 should be represented or considered when evaluating the handling qualities of an aircraft. For engineers, the most satisfying and potentially instructive means of dealing with the system would be through computational analysis. Each element would be represented by an analytical model with which the output could be computed for specified inputs. The elements would be arranged and interconnected in the form of Fig. 1 and the total system responses could be computed for specificed inputs. More importantly, the dynamics of the system could be assessed and related to the characteristics of the elements which form the system. This form of analytical representation was pioneered by Tustin [38] but has been further developed and evangelized by McRuer [39] and others. McRuer has had a profound influence on our understanding of the interactive dynamics of the pilotvehicle system, especially by his dedication to the need for parallel development of analysis and experimentation. His work stimulated other researchers such as Krendel, [40] Hall, [41] Anderson, [42] and Neal and Smith [43] to contribute analysis methods based upon experimental data.

The major difficulty with analysis has been the analytical representation of the adaptive human pilot. The other elements of the system can be accurately represented; the pilot and his actions are only partially understood. One troublesome aspect of the pilot is in knowing what variables he is sensing and acting upon in supplying his corrective control inputs. For example, as he performs the air-to-air fighter task, he will close different loops during different portions of the tasks: perhaps normal acceleration and pitch rate for corrective pitch inputs during the acquisition turn, and pitch angle error and pitch rate during tracking. What does he sense in landing: altitude, altitude-rate, pitch angle and rate, ground speed? How do the amounts of each of these vary? What strategy does he employ to select among the variables as the maneuver progresses, and what "gains" does he use? Work of very high quality has been done in these areas, but we are still of limited capability in predicting the pilot's dynamic behavior, especially in new situations. As quoted by McRuer, [39] the words of Cowley and Skan in their 1930 paper [44] describe the difficulties of pilot-vehicle analysis even today: "A mathematical investigation of the controlled motion is rendered almost impossible on account of the adaptability of the pilot. Thus, if it is found that the pilot operates the controls of a certain machine according to certain laws, and so obtains the best performance, it cannot be assumed that the same pilot would apply the same laws to another machine. He would subconsciously, if not intentionally, change his methods to suit the new conditions, and the various laws possible to a pilot are too numerous for a general analysis."

Experimental

Experimental methods are the other means of assessing the quality of the pilot-airplane combination. Experimentation involves the combination of the pilot and either the real vehicle or a simulation (ground or in-flight) of the vehicle in the accomplishment of the real task or a simulation of the real task. For design purposes, simulation is employed, and the preponderant use is of ground simulation, although in-flight simulation is assuming a growing role during aircraft development.

There are two general data outputs of the experimental methods: performance measurement, and pilot evaluation. Because the pilot is adaptive, performance measurement should include not only how well he is doing (task performance), but also how much effort the pilot is supplying (workload) to achieve that performance. Workload is used to convey the amount of effort and attention, both physical and mental, that the pilot must provide to obtain a given level of performance. [1] The meaningfulness of task performance measurement data is always bounded by realism of that which is measured. For example, much effort has been devoted to measuring tracking performance expressed in statistical measures of aim wander taken from 30 to one mim tracking runs on a nonmaneuvering target; the real fighter pilot deals generally with an aggressively maneuvering target on whom he needs only to achieve the correct solution long enough to fire. These are really two different tasks. Workload measures are difficult, too: physical workload of various kinds have been measured; mental workload measurement is much more difficult. One without the other is an incomplete description of the total pilot activity being supplied. Roscoe [45] notes that "ideally, assessment or measurement of pilot workload should be objective and result in absolute values; at present, this is not possible, nor is there any evidence that this ideal will be realized in the foreseeable future. It is also unfortunate that the human pilot cannot be measured with the same degree of precision as can mechanical and electronic functions." Ellis [46] suggests the use of subjective pilot opinion based on a modified Cooper-Harper scale as a standard measure of pilot workload. McDonnell [47] introduced the concept that the pilot's ability to compensate implies that he has spare capacity. This idea was developed further by Clement, McRuer, and Klein [48] who suggested that workload margin be defined as the capacity to accomplish additional expected or unexpected tasks. Even if the workload measures are incomplete, they are important data for understanding and interpreting pilot evaluation data. One recent development which shows promise is the Workload Assessment Device (WAD) that Schiflett [49] has used to measure certain aspects of mental workload; it offers handling quality researchers a potential for comparative measures of this elusive parameter.

Pilot evaluation is still the most reliable means of making handling quality evaluations. It permits the assessment of the interactions between pilot-vehicle performance and total workload in determining the suitability of an airplane for the intended use. It enables the engineer not only to evaluate the quality but to obtain the reasons behind the assessment. It allows the engineer to devise and refine analytical models of pilot controller behavior for use in predictive analysis of new situations. But pilot evaluation is like most forms of experimental data: it is only as good as the experimental design and execution. Further, since it is a subjective output of the human, it can be affected by factors not normally monitored by engineers. The remainder of this paper attempts to provide guidance for the conduct of pilot evaluation experiments. The reader is cautioned that such guidance is constrained by the knowledge and experience of the authors; much work needs to be done in this field, and it is the hope of the authors that this effort will stimulate additional contributions to this important area.

Pilot Evaluation

Pilot evaluation refers to the subjective assessment of aircraft handling qualities by pilots. The evaluation data consists generally of two parts: the pilot's commentary on the observations he made, and the rating he assigns. Commentary and ratings are both important sources of information; they are the most important data on the closed-loop pilot-airplane combination which the engineer has.

Comment Data

Comment data are the backbone of the evaluation experiment. Commentary can provide the engineer with a basic understanding from which he can model the closed-loop system. It can tell the analyst not only that something is wrong, but also where he can introduce system changes to improve the handling qualities.

Comment data can be stimulated by providing a questionnaire or a list of items for which comments are desired. Such comment stimuli are best when prepared jointly by engineers and pilots and refined with use in practice evaluations. The evaluation pilot should address each item for every evaluation; otherwise, the analyst will not know whether the item was overlooked or unimportant.

Pilot Rating

Pilot rating is the other necessary ingredient in pilot evaluation. It is the end product of the evaluation process, giving weight to each of the good and bad features as they relate to intended use of the flight vehicle and quantifying the overall quality. A thorough discussion of the nature and history of pilot rating was set forth by the authors in NASA TN D-5153 [1] in 1969, along with a new pilot rating scale and methodology. That scale, usually called the Cooper-Harper scale, has been accepted over the ensuing years as the standard measure of quality during evaluation, where in previous years, several scales had been used. The use of one scale since 1969 has been of considerable benefit to engineers, and it has generally found international acceptance.

One problem has been that the background guidance contained in TN D-5153 has not received the attention that has been given to the scale. For this reason, some of the definitions and guidance of the earlier reference are included in the following discussion. The scale is reproduced here as Fig. 8.

Attention is first called to the “decision tree” structure of the scale. The evaluation pilot answers to a series of dichotomous (two-way) choices which will lead him to a choice of, at most, one among three ratings. These decisions are sometimes obvious and easy; at other times difficult and searching. They are, however, decisions fundamental to the attainment of meaningful, reliable, and repeatable ratings. These decisions—and, in fact, the use of the whole scale—depend upon the precise definition of the words used.

The first and most fundamental definition is that of the reguired operation or intended use of the aircraft for which the evaluation rating is to be given. This must be explicitly considered and specified; every aspect which can be considered relevant should be addressed. The rating scale is, in effect, a yardstick. The length measured has little meaning unless the dimensions of interest are specified. The definition should include what the pilot is required to accomplish with the aircraft; and the conditions or circumstances under which the required operation is to be conducted.

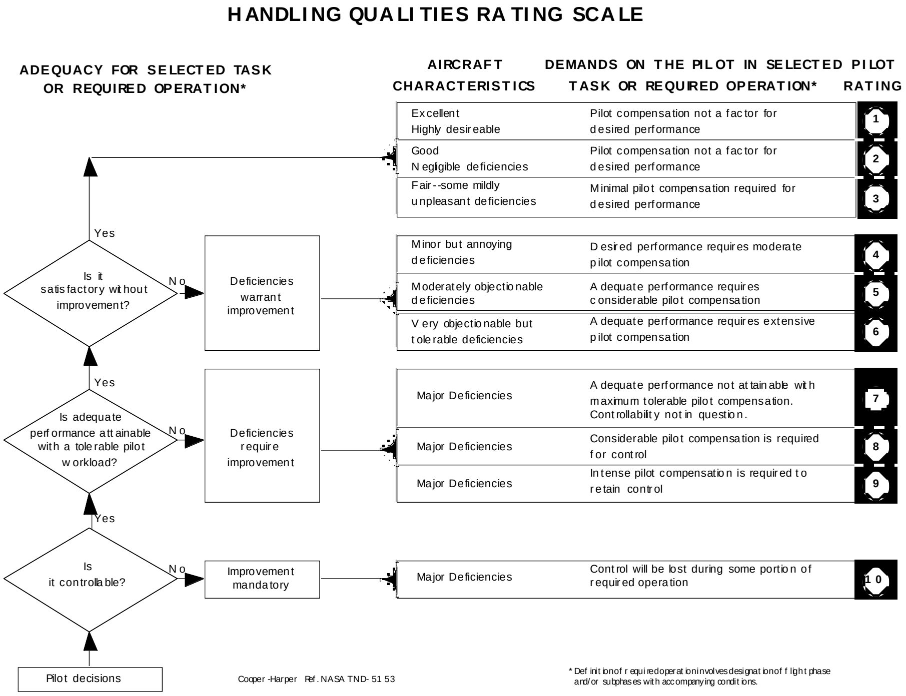


Figure 8: Cooper-Harper handling qualities rating scale.

The required operation should address and specify the end objective, the various primary and secondary tasks, the environment and disturbances that are expected to be encountered, the piloting population who will perform the operation—their range of experience, background, and training—and anything else believed to be relevant.

The first rating-scale decision questions controllability. Will the pilot lose control during some portion of the required operation? Note that he cannot adopt an alternative operation in order to retain control and call it controllable. If the required operation is, for example, air-to-air combat and a divergent pilot-induced oscillation (PlO) is experienced, it must be a rating 10 even if the PlO disappears when the combat maneuvering is abandoned. In other words, controllability is assessed in and for the required operation: to control is to exercise direction of, to command, and to regulate in the required operation.

The next question in the decision tree—if controllability is given an affirmative assessment—is concerned with the adequacy of performance in the required operation. And because of the adaptive character of the human pilot, performance cannot be divorced from the piloting activity which is being supplied. That activity includes both mental and physical efforts, so the term workload is used to address both aspects. "Tolerable pilot workload" requires a judgement from the pilot as to where the tolerable/intolerable workload boundary lies for the required operation. Performance is a quality of the required operation: how well is it being done? "Adequate performance" must be defined: how good is good enough? How precise is precise enough? If the answer to the adequacy question is negative, the required operation may still be performed, but it will either require excessive workload or the performance will be inadequately precise, or both. Logically, then, the deficiencies require improvement.

If the pilot judges the performance adequate for a tolerable workload, then the handling qualities are adequate—good enough for the required operation. They may, however, be deficient enough to warrant improvement so as to achieve significantly improved performance of the required operation. That judgement is faced by answering the next question: "Is it satisfactory without improvement?" A "no" answer implies that the deficiencies warrant improvement; that with improvement will come significant benefit in terms of workload reduction or performance improvement. A "yes" answer says that further improvement in the deficiencies will not produce significant improvements in the required operation, even though a better handling aircraft can result.

It can be seen by again referring to the rating scale that after the decision tree has been addressed, the rating process has narrowed to a choice among three rating at the most. Further, one can see that selection of a rating of 3.5, 6.5, or 9.5 is an indication that the pilot is avoiding making a fundamental decision, one that he is the most qualified to make.

Having completed the decision tree process, the pilot encounters additional considerations to be assessed in reading the descriptors in the individual rating boxes. Note that these are additional considerations. They do not stand alone; they depend for sufficiency on the decision tree judgments. Therefore, anyone using this scale by using only the individual rating descriptors—without proceeding through the decision tree process—is inherently missing important steps in logic and handling quality considerations.

To choose among the three ratings, there are different considerations depending upon where the decision tree answers led you. In the 7, 8, 9 area, the consideration is controllability: how dependent is it on pilot compensation for aircraft inadequacies? Inadequacy for the required operations has already been stated; how dependent are we now on the correct pilot actions to stay in control? A rating of 7 says it is not really a problem; a 9 says that intensive effort, concentration, or compensation is required to hang in there, still doing the task.

In the 4, 5, 6 region, a different consideration is required. Performance, although adequate, requires significant pilot compensation for aircraft deficiencies. A distinguishing quality is the level of performance achieved. As noted earlier, the level of performance achieved depends upon how taxed the pilot is in producing that performance, so a level of performance is always qualified by a specified level of pilot compensation. Two levels of performance are specified: desired and adequate. Both must be defined, as thoroughly as possible, as to what are the handling quality performance parameters and what levels (quantitative, if possible; qualitative otherwise) are desired—or adequate—for the required operation. With a rating of 4, desired performance requires moderate pilot compensation. A rating of 6 relates adequate performance to extensive compensation, while a rating of 5 attributes adquate performance to considerable pilot compensation.

The experimenter and evaluation pilot should examine all aspects of the required operation in regard to the consideration of performance. For example, roil "racheting" (oscillatory roil-rate response) during target acquisition in air-to-air combat maneuvering could hardly be considered as desirable performance, even if final tracking did not exhibit this characteristic. Neither would a roll PlO during lineup from an offset landing approach be considered desired performance, even though the desired touchdown zone parameters could be accomplished.

It may be possible to achieve desired performance with extensive compensation (more than "moderate" of PR = 4). and this combination does not appear explicitly in the scale. The authors believed that this situation—worse than PR = 4—would result in a PR = 5 (adequate performance/considerable compensation). The evaluation pilot can assess this by backing off from achieving desired performance to an adequate performance level and noting the amount of compensation required. Should the circumstances arise that adequate performance required substantially less than considerable compensation (moderate or minimal, for example), the evaluation pilot would have to resort to a PR = 4.5 to delineate the situation.

Also included in the 4, 5, 6 region are adjectives to classify the severity of the deficiencies which warrant improvement. It is possible that the choice of one of these descriptors could conflict with the performance/compensation selections. This has not appeared to be a problem, but if it should be, the pilot should note the conflict and the reason for his final selection.

In the 1, 2, 3 region of the rating scale, it can be seen that the performance/compensation considerations are less discriminating. This is hardly surprising since the handling qualities in this region are already "satisfactory without improvement." Desired performance is attainable with no more than minimal compensation. Here the adjectives "excellent," "good," or "fair" are dominant in separating satisfactory airplanes, one from another. This is the region for pointing toward design goals—the best direction to go in seeking handling improvements.

Evaluation Experiments

Guidance is offered in this section for the conduct of handling quality evaluation experiments. The material is organized into three sections which address the issues of planning, execution, and analysis. The most general application is the conduct of simulation experiments where the pilot flies a representation of the actual flight vehicle. In its simplest form, the simulator device may be a chair, broomstick controller, and oscilloscope display. Or, more likely, it may be a much more sophisticated form of ground simulator (fixed- or movingbase) or in-flight simulator. Finally, the evaluation may be in the real airplane, but even here the required operation may be simulated to some extent. The issues discussed here largely arise from the fact that the experiments involve human subjects, the primary data gathered are their evaluation comments and ratings, and the evaluation situation seldom fully represents the real world situation for which evaluation data are needed.

Planning

This part is often left to the engineers to do, but it should be a joint effort of pilots and engineers, working together to achieve a sound experiment from which meaningful, understandable data can be gathered, analyzed, and reported. The talents of both engineer and pilot are required for the planning stage.

Definition of the Required Operation

This should be as complete as possible. Initial focus should be upon the real world operation for which the data is desired; then, with the existing evaluation tools, one should examine what can be done to assess that situation. The tasks which the piloted airplane must perform, the weather (instrument, visual) and environmental conditions (day, night) which are expected to be encountered, the situation stressors (emergencies, upsets, combat), the disturbances (turbulence), distractions (secondary tasks), the sources of information available (displays, director guidances)—all these and more need to be considered. Secondary piloting tasks (voice communication, airplane and weapon system management) as well as primary tasks should be considered as they affect the attention available and total pilot workload.

Evaluation Situation/Extrapolation

In comparing the real world versus the simulation situation, the issue arises as to how to deal with the differences. Some would have the pilot assess only the simulated operation; others would have him use the simulation results to predict/extrapolate to the real world operation. The issue must be discussed and addressed a priori; otherwise, different pilots may produce different results. An important aspect to this question of extrapolation is: if the pilot doesn't do it, who will? And what are his credentials for doing so? Some differences (especially simulator deficiencies) would seem to be primarily left to the engineer to unravel (perhaps with the aid of a test pilot), for it is difficult for the evaluation pilot to fully assess the effects, for example, for missing motion cues or time delays in the visual scene. But when the simulation tasks do not include all of the real situation, one would perhaps rather depend upon the pilot to assess what he sees in the simulator in the light of his experience in the real-world tasks.

User Pilots

This is the specification of the group of pilots who are to perform the required operation in real life. The range of background, experience, training, health, stamina, and motivation should be assessed and specified.

Selection of Evaluation Pilots

The evaluation pilots represent the user population and should be experienced in the required operation. For some research applications, the required operation has not yet been performed (e.g., re-entry before space flight was achieved) and the evaluation pilots must gain their experience by a thorough study of the required mission maneuvers, tasks (primary and secondary), and circumstances.

Number of Evaluation Pilots

A classic handling qualities experiment [50] showed that a few pilots evaluating for a longer period of time produced the same central tendency of the rating excursions as a larger group conducting shorter evaluations. What was lost with the larger group, however, was the quality, consistency, and meaningfulness of the pilot comment data. Based upon this and other experiences, it is generally recommended to use only a few pilots (sometimes only one) until the experiment has matured through the engineer's understanding of the comment and rating data. His task of sorting out, organizing, and digesting the comments and ratings to understand the pilotairplane system is complex and often frustrating. By working closely with one or a few pilots initially, the engineer can often acquire this understanding sooner. He can then expand the evaluation pilot sample to test his new-found hypothesis of the pilot-vehicle system, revising his original theories as necessary.

Blind Evaluations

There may be differing opinions, especially among pilots, as to the advisability of allowing the evaluator to know the parameter values used to describe the aircraft which he is evaluating. Many experienced evaluation pilots, however, strongly voice objections to being told the parameter values; they find it more difficult to comment on a pitch oscillation that they experienced when the parameter values given to them indicate that the pitch damping is high. Obviously, closedloop pitch oscillations can occur even when pitch damping is high due to high pitch frequency, light forces, small controller motions, response time delays, or other causes. But pilots who know the parameter values may feel inhibited; they may attribute the oscillation to their inappropriate control technique and downplay the problem. It is therefore recommended that evaluations be conducted with the evaluator blind to the parameter values.

For purposes other than evaluation, the pilot can be told the parameter values to enhance his knowledge and training.

Repeated Evaluations

It is highly desirable to include repeat evaluations in the experimental matrix. The pilot should not know that he is evaluating something he has seen before; assign different identifying numbers to the repeat configurations. The consistency of his comments and ratings for repeat configurations provide an assessment of the experiment itself; variability points toward uncontrolled/unknown factors in the experiment that should be identified and dealt with. Rating variability should be less than one pilot rating, in terms of the expected variation of the rating from the central tendency.

As for how many repeats to plan for in one experiment, one would like to confirm the hypothesis that the pilot's variability is less than one rating, and proceed accordingly. If the rating variability is independent of the parameter values, not many repeats would be required; but if one wants to examine rating variability at several parameter values, the expetiment size and cost grow correspondingly. There is also a tradeoff between number of repeats and number of evaluation pilots in an experiment of a given size. A precise answer does not exist. A rule-of-thumb might be to increase the number of evaluations by one-third to one-half to account for repeat evaluations.

Length of Evaluation

The presumption in most evaluation experiments is that the results apply for a trained pilot. Therefore, the evaluator should fly each configuration of handling characteristics long enough to become "trained" before giving his evaluation and rating. How long this is isn't generally known (and would probably vary with the parameter values, anyhow) so the recommended practice is to let the evaluator decide. Instruct him to repeat the tasks which represent the required operation enough times so that he wouldn't expect his evaluation to change if he performed them again. Then use this evaluation time/number-of-runs structure for the experiment, always giving the evaluator the option of additional time on any configuration which needs it.

Specification of Performance Standards

In planning an evaluation experiment, the performance to be achieved in the required operation must be defined. The full range of performance should be addressed as far as possible. For example, if landing approach and touchdown is the required operation, one might identify a touchdown spot on the runway, surrounded by a rectangle which defines dispersion limits fore and aft and left and right of the touchdown spot. The size of the rectangle would be smaller for desired performance as compared to adequate performance. The size of the rectangle for a given level of performance (desired or adequate) could be larger for cases where turbulence, crosswind, or decreased-headwind/tailwind increased the task difficulty. The specific size of these rectangles are difficult to specify a priori—they would be different as compared to real flight if the simulator fidelity was low—but the concept should be discussed and the pilots should work out standard definitions of performance before the formal evaluations begin.

The experimentors should attempt to address (even if only qualitatively) all aspects of performance important in the required operation. In the example above, one should include airspeed and angle-of-attack control; sink rate at touchdown; and pitch, roll, and yaw attitude at touchdown. Even with all these considerations, one has to include what happened prior to touchdown: a roll oscillation prior to touchdown could hardly be considered “desired performance” of the closedloop pilot-airplane combination, even if all of the touchdown parameters were met.

Some experiments have used a “desired performance” touchdown zone and parameters that were invariant with turbulence levels and crosswind components. This leads to what appears to be a paradox in that the pilot rating gets worse for a given aircraft as the task difficulty increases, implying that airplane deficiencies have appeared which now need correcting. In reality, only the physical environment in which the task must be performed has worsened, and one would expect the touchdownd precision to worsen somewhat. The real question is: how much of the worsening is task-demand related, and how much is it deficient-handling-qualities related? Experienced pilots are able to judge this reasonably well from having flown different airplanes through a large spectrum of environmental conditions, and their guidance should be sought during the planning and evaluation phases.

Execution

The best advice in carrying out an experiment is always to execute the experiment as planned. In handling quality experiments involving human subjects, however, not all of the experiment can be preplanned and certain issues must be addressed during the conduct of the experiment. In the following material, the authors discuss certain of these issues and offer some guidance.

Pre-Evaluation Familiarization

Several factors which may affect the evaluation data can be dealt with by conducting a pre-evaluation phase. The pilot can familiarize himself with the tasks, the experimental procedures, and the use of the comment card and rating scale. He can experiment with the performance standards that were defined during rhe planning stage, revising the numerical values or modifying the selected parameters. During this period, he will improve his state of training and proficiency in the tasks and gain familiarity with the simulator.

During this phase, it is helpful to allow the pilot to evaluate handling qualities that span the range of the rating scale; that is, let him see good, bad, and in-between characteristics. This is perhaps less important with experiencei evaluation pilots, but it can be an important factor with operational pilots whose experience is confined to one or two airplanes. They may not realize the handling improvement (or degradation) that is possible, and may tend to rate relative to their recent experience and differently from more widely experienced pilots. This pre-evaluation experience also gives the pilots time and opportunity to deal with the trauma of making the rLting scale decisions and worrying later if they were doing it consistently.

Simulation Experiment Deficiencies

The pilot should be encouraged to critique the experiment and simulator early in the experiment, especially during the pre-evaluation phase. Any deficiencies should be noted, discussed, and corrected. If correction is not feasible, it may be possible to alter the experiment to account for the deficiency, or else call upon the pilot to attempt to account for the inadequacy as best he can in his evaluations. Whatever course is pursued, it should be identified and discussed in the report of the experimental results.

Evaluation Time/Number of Runs

The evaluation phase is when the time assumptions made during the planning phase can be tested. As noted earlier, the pilot should be allowed to evaluate long enough to let him feel comfortable in summarizing his comments and ratings; long enough so that additional time would not significantly change his evaluation.

Blind Evaluations and Pilot Morale

The characteristics of the aircraft that are being evaluated and the results of other pilot evaluations should remain unknown to the pilot until his participation in the experiment is concluded. This especially means that evaluators should not attend the debriefing of another evaluator unless they have completed their program. Evaluators can discuss the tasks but not the evaluation results and ratings.

The engineer should anticipate that the pilot will be troubled by the uncertainty about how well he is doing. A pilot is accustomed to immediate feedback on the quality of his piloting performance, and he may experience uncertainty and doubt about his evaluation performance and repeatability. The pilot tends to view the engineer as one who has all the data and is testing the pilot to see if he measures up; when in fact, the engitleer only knows the aircraft parameters and needs the pilot to tell him the character of the pilot-aircraft combination. It is helpful for the engineer to remind the pilot that only he can produce the needed answers, and give him a periodic pat-on-the-back regarding the quality of the results in order to allay his concerns over his performance as an evaluator.

Terminology

Encourage the pilot to use plain, comfortable language to describe his evaluation results. Whenever standard terminology (as defined in NASA TN D-5153, for example) is insufficient, ask him to define words and phrases which are unfamiliar to the engineer. Engineering terms should generally be avoided, especially if they have definite meanings associated with the aircraft-alone response (such as frequency, damping, stick force per g, etc.).

Data Taking

The evaluation comments and ratings should be given during and at the end of the evaluation, before the next evaluation commences. These data should be tape recorded for later transcription and analysis. Engineers often monitor the evaluation comments, taking notes of pertinent points. This procedure is helpful in both monitoring the conduct of the experiment as well as pinpointing any problems in understanding the pilot's terminology. If the pilot's comments are transcribed, the pilot can assist in the editing of the commentary for errors in transcription.

Pilot Rating

The rating scale should be posted in the simulator and accessible for easy reference. The full decision tree should be traversed each time a rating is given, preferably aloud, so the engineer can witness the decision process. When the choice has narrowed to one of the three ratings, it is helpful t. reject the other two, announcing why each is not proper for the evaluation at hand. In this way, one arrives at the proper ratings by first saying why the rejected ratings don't apply.

Analysis

There are only a few general guidelines that will be offered here for the analysis of pilot evaluation data. Probably the most fundamental one is to quote Calspan's Chief Test Pilot of many years, Nello Infanti: "The pilot is always right." For this statement to be taken at face value, however, one must assume objectivity and freedom from bias on the part of the evaluating pilot. Objectivity requires a sincere desire for truth, implemented by skillful observation and accurate reporting. Although this implicit faith in the pilot's comments may be abrasive to some engineers who have had to deal with test pilots over a number of years, to others who have learned new things by trying to understand pilots' comments, it may make good sense. The presumption here is that the pilot's words are correct; if they seem unsound, it is probably because the engineer's understanding is insufficient. The key issue is that the pilot is an intelligence who is the only one present in the dynamic system under evaluation; he has the capacity to report to engineers information which would be difficult to obtain by other means. If we assume that the pilot's comments have a specific basis and meaning, we can deduce a logical system framework which can support those comments. From that framework, we can model an analytical equivalent of the pilot as a dynamic system element and use that model to predict the performance of future systems.

Another issue with which every analyst must deal is variability in pilot ratings. Sometimes the variability is so great and apparently unstructured as to make even the most tenacious engineer wonder why he ever chose handling qualities as his special niche. But most experience with pilot ratings is much better, so that one should be stimulated to dig even deeper into the analysis when faced with rating variability. There must be a reason: the challenge is to find it.

The first place to look is in the design or execution of the experiment; rating variability often has its source in the experiment itself. The pilots may have different concepts of the required operation, leading them to emphasize different aspects of task performance—or even perform different evaluation tasks—and hence give different ratings. They may be using different performance standards, or including different factors in their performance definitions. They may be using the rating scale differently (ignoring the decision tree, for example), or they may not be using the same definitions of the words in the scale.

Another place to look for the source of rating variability is in the pilot's background and experience. Has he had the opportunity to experience a signficant spectrum of handling qualities from good to bad, either from flying different airplanes or from the pre-evaluation phase of the experiment? Until he has seen a wide range of quality, he may expect even the best aircraft to be difficult or demanding, and attribute piloting difficulties not to the aircraft but to himself. After all, that is the way pilots are trained; the aircraft is provided, the pilot trains and learns to fly it. Only test pilots are given the luxury of blaming piloting difficulties on the aircraft.

Although it is fortunately rare, another source of variability must be mentioned. There are cases where the pilot rating has been influenced by factors outside of the experiment itself.

Pilots could be advocates of a design or concept—they may not want to expose deficiencies; they may be competing with other pilots to fly a mission and don't want to call attention to their piloting difficulties; they might want to appear more capable as a pilot by giving better ratings. These items illuminate a characteristic of a desirable evaluation pilot: he should be highly motivated to carry out the objectives of the experiment, to “call them the way he sees them,” and let the engineer figure out what it all means.

We are fortunate that the piloting profession has, throughout the years since the first airplane flight, followed the highly professional example set by Orville and Wilbur Wright and contributed so greatly and uniquely to our knowledge and understanding of the pilot-aircraft system. This paper is dedicated to the memory of all those pilots and engineers who have preceded us and made possible our current knowledge and understanding.

Concluding Observations

1. The never-ending search for increased aircraft performance and mission productivity continues to bring forth new technology to challenge our capability to provide good handling qualities.

2. The increasing use of flight simulation, both groundbased and airborne, has enabled great strides to be made in understanding the nature of the pilot-vehicle dynamic system, and substantial quantities of handling qualities data has been generated.

3. Flight simulation is bringing increased opportunities as well as challenges for the test pilot to affect the design of modern aircraft.

4. For simulation results to be accurate and informative, the experiments must be meticulously planned and executed through a close, cooperative relationship between engineer and evaluation pilot.

5. Subjective pilot evaluations which produce comments and ratings continue to be the primary means for judging the quality of handling.

6. The nearly universal use of the Cooper-Harper rating scale for handling qualities assessment is not commensurate with the general lack of access to and familiarity with NASA TN D-5153 (which gives background guidance, definition of terms, and recommended use). This report should be reprinted and given distribution as widely as the printed rating scales themselves.

7. There will be further improvements in flight simulators to enhance fidelity and reduce extrapolation, but an evaluation pilot must always be alert to simulator limitations and their effect on his evaluations.

8. Limitations inherent in simulators and simulation experiments are not always recognized. Pilot-engineer collaboration and communication are vital to the product of valid, useful data.

9. Engineers should always validate and document the performance of their flight simulators.

Acknowledgment

We wish to express our appreciation to Maurice D. White and Charles R. Chalk for their helpful review and editing support during preparation of this manuscript, and to those other individuals who were kind enough to read our drafts and offer suggestions.

Table 1a Example of early flight research correlating stability characteristics with pilot opinion ratings
(Table II from Ref. [11])

[S, stable; U, unstable]

Airplane	Elevator fixed Throttle closed	Elevator fixed Throttle full	Elevator free Throttle closed	Elevator free Throttle full
Fairchild 22	S	S	S	S
Martin T4M-1	S	S	S	S
Consolidated NY-2	S	U below 49 mph	S	S
Boeing F4B-2	S	U below 58 mph	S	U below 56 mph
Verville AT	S	U below 58 mph	S	U below 57 mph
Douglas O-2H	S	U below 83 mph	U below 91 mph	S
Fairchild FC2-W2	S	S	S	Ubelow 61 mph
Martin XBM-1	S	S	S	Ubelow 70 mph

Table 1b Example of early flight research correlating stability characteristics with pilot opinion ratings
(Table III from Ref. [11])

	Observed characteristics				Measured characteristics	
		Factors affecting stiffness				
Airplane	Stiffness	Elevator force	Elevator movement	Pitching in rough air	Period	Damping
Fairchild 22	D	C	B	A	A	A
Martin T4M-1	C	A	A	B	B	B
Consolidated NY-2	C	B	B	A	B	B
Boeing F4B-2	D	C	C	C	A	C
Verville AT	C	C	B	B	D	B
Douglas O-2H	A	A	A	A	D	D
Fairchild FC2-W2	B	B	B	B	C	A
Martin XBM-12	A	A	C	D	C	D

A is used to designate airplanes that are stiffest, require the greatest elevator forces and movement, do most pitching in rough air, and have the shortest periods and the greatest damping.

References

[1] Cooper, G.E. and Harper. R.P., "The Use of Pilot Rating in the Evaluation of Aircraft Handling Qualities," NASA TN D-5153. April 1969.

[2] The Papers of Wilbur and Orville Wright, Vol. 1, 1898-1905; Vol. 2, 1906-1948.

[3] KelIy, F.C., *The Wright Brothers,* Bantam Books, New York, Nov. 1983 edition.

[4] Perkins. C.D., "The Development of Airplane Stability and Control," AIAA Paper No. 69-1137, 1969 Von Karman Lecture.

[5] Personal conversation with James M. Nissen, restorer and pilot of antique airplanes, and former NACA test pilot, 1984.

[6] Lecombcr, B., "Flying she Sopwith Camel and Fokker Triplane." *Flight International,* April 8, 1978.

[7] Williams, N., "On Gossamer Wings," *Flight International,* Jan. 9, 1975.

[8] Warner, E.P. and Norton, F.H., "Preliminary Report on Free Flight Tests," NACA TR 70, 1919.

[9] Personal conversation with Gen. James Doolittle, 1983.

[10] Thomas, L. and Jablonski, E., *Doolittle: A Biography,* Doubleday & Co., New York, 1976.

[11] Soule, H.A., "Flight Measurements of the Dynamic Longitudinal Stability of Several Airplanes and a Correlation of the Measurements with Pilots' Observations of Handling Characteristics," NACA TR-578, 1936.

[12] Gough, MN., "Notes on Stability from the Pilot's Standpoint," *Journal of the Aeronautical Sciences,* Vol. 6, Aug. 1939.

[13] Allen, E.T., "Flight Testing for Performance and Stability," *Journal of Aeronautical Sciences,* Vol. 10, Jan. 1943.

[14] Cooper, G.E., "Understanding and Interpreting Pilot Opinion," *Aeronautical Engineering Review,* March 1981.

[15] Soule, HA., "Preliminary Investigation of the Flying Qualities of Airplanes," NACA Report 700, 1940.

[16] Gilruth, R.R., "Requirements for Satisfactory Flying Qualities of Air?lanes," NACA TR 755, 1943.

[17] Anon, "Specification for flying Qualities of Piloted Airplanes," NAVAER SR-I 19, 1942.

[18] Anon, "Stability and Control Requirements for Airplanes," AAF-C-1815, 1943.

[19] Kaufman, William M., Smith, Allen, Liddell, Charles J. Jr., and Cooper, George E., "Flight Tests of an Apparatus for Varying Dihedral Effect in Flight," NACA TN 1788, Dec. 1948.

[20] Liddell, C.J. Jr., Van Dyke, R.D., Jr. and Heinle, D.R., "A flight Determination of the Tolerable Range of Effective Dihedral on a Conventiional Fighter Airplane," NACA TN 1936, Aug. 1949.

[21] Graham, D. and James. C.W., "A Flight Investigation of Minimum Acceptable Lateral Dynamic Stability," Cornell Aeronautical Laboratory Rept. No. TB-574-F-3, April 30. 1950.

[22] Campbell, G. and Newell, F., "Flight Evaluations of Variable Short Period and Phugoid Characteristics in a B-26," Cornell Aeronautical Laboratory Rept. No. TB-757-F-1 I, Wright Air Development Center, TR-54-594, Dec. 1954.

[23] Harper. R.P., "Flight Evaluations of Various Longitudinal Handling Qualities in a Variable Stability Jet Fighter," Wright Air Development Center TR-55-299, July 1955.

[24] McFadden, N.M., Vomaske, R.F., and Heinle, D.R., "Flight Investigation Using Variable-Stability Airplanes of Minimum Stability Requirements for High-Speed, High-Altitude Vehicles," NASA TN-D-779, April 1961.

[25] Bull, G. and Kidd, E.A., "Air-to-Air Tracking with Linear and Non-Linear Yaw Damping," Wright Air Development Center TR-55-223, June 1955.

[26] Harper, R.P. Jr. and Cooper, G.E., "A Revised Pilot Rating Scale for the Evaluation of Handling Qualities," Cornell Aeronautical Laboratory Rept. No. 153, 1966.

[27] Cooper, G.E. and Harper, R.P. Jr., "The Use of Pilot Rating in the Evaluation of Aircraft Handling Qualities," AGARD Rept. 567, April 1969.

[28] Anon, Military Specification, Flying Qualities of Piloted Airplanes, MIL-F-8785, Sept. 1954.

[29] Anon, Military Specification, Flying Qualities of Piloted Airplanes, MIL-F-8685B (ASG), Aug. 1969.

[30] Chalk, CR., Neal, T.P., Harris, TM. et al., Background Information and User Guide for MIL-F-8785B (ASG), "Military Specifscation—Flying Qualities of Piloted Airplanes," Air Force Flight Dynamics Laboratory, TR-69-72, Aug. 1969.

[31] Anon, Military Specification, flying Qualities of Piloted Airplanes, MIL-F-8785C, Nov. 1980.

[32] Smith, R.E. and Bailey, RE., "Evaluating the flying Qualities of Today's Fighter Aircraft," AGARD FMP Symposium on Combat Aircraft Maneuverability, Oct. 1981.

[33] Berry, D.T., "Flying Qualities: A Costly Lapse in Flight Control Design?" AIAA Guidance and Control Conference, April 1982.

[34] Hoh, RH., "Bring Cohesion to Handling Qualities Engineering," *Aeronautics and Astronautics,* June 1983.

[35] Harper, R.P. Jr., "Handling Qualities of flight Vehicles," *Israel Journal of Technology,* Vol. 21, 1983.

[36] Powers, B., "Space Shuttle Pilot-Induced Oscillation Research Testing," NASA Technical Memorandum 86034.

[37] Anon, Airbus Industries, "A-320 Flight Deck, The Logical Path," Issue 2, June 1984.

[38] Tustin, A., "The Nature of the Operator's Response in Manual Control and Its Implications for Controller Design," *Journal of the IEEE,* Vol. 94, Part hA, 1947.

[39] McRuer, D., "Development of Pilot-In-The-Loop Analysis," Journal of Aircraft, Vol. 10, No.9, Sept. 1973.

[40] McRuer, D.T. and Krendel, ES., "Dynamic Response of Human Operators," Wright Air Development Center TR-56-523, Oct. 1957.

[41] Hall, lAM., "Effects of Controlled Element on the Human Pilot," Wright Air Development Center, TR-57-509, Aug. 1958.

[42] Anderson, RO., "A New Approach to the Specification and Evaluation of Flying Qualities," Air Force flight Dynamics Laboratory, TR-69-120, June 1970.

[43] Neal, T.P. and Smith, R.E., "An In-Flight Investigation to Develop Control System Design Critria for Fighter Airplanes," Air Force Flight Dynamics Laboratory TR-70-74, Dec. 1970.

[44] Cowley, W.L. and Skan, S.W., "A Simplified Analysis of the Stability of Airplanes," R&M No. l333, March 1930, Aeronautical Research Committee, London, England.

[45] Roscoe, A.H., "Assessing Pilot Workload," AGARDograph No. 233, Chapter 1, NATO, Feb. 1978.

[46] Ellis, GA., "Assessing Pilot Workload," AGARDograph No. 233, Chapter 2, NATO, Feb. 1978.

[47] McDonnell, J.D., "Pilot Rating Techniques for the Estimation of Evaluation of Handling Qualities," Air Force Flight Dynamics Laboratory TR-68-76, Dec. 1968.

[48] Clement, W.F., McRuer, D.T., and Klein, RH., "Systematic Manual Control Display Design," Conference Proceedings No. 96, Guidance and Control Displays, AGARD, Paris, 1971.

[49] Schiflets, S.G., "Evaluation of a Pilot Workload Assessment Device to Test Alternate Display Formats and Control Handling Qualities," Naval Air Test Center TR SY-33R-80, March 1980, Patuxent River, MD.

[50] Kidd, E.A. and Bull, G., "Handling Qualities Requirements as Influenced by Pilot Evaluation Time and Sample Size," Cornell Aeronautical Laboratory Rept. TB-1444-F-l, Feb. 1962.