A Study of Student Perception of the Validity and Reliability of Learner Centered and Collaborative Grading in University Flight Training Assessment

by
Francis H. Ayers, Jr., Ed.D
Chairman, Flight Training Department
Embry Riddle Aeronautical University
Daytona Beach, Florida

2007-2008 FITS Grant Extension
TASK (6) FAA/Industry Training Standards

May 2008

Abstract

A Study of Student Perception of the Validity and Reliability of Learner Centered and Collaborative Grading in University Flight Training Assessment, Ayers, Francis H., Jr., 2008: Keywords: Alternative Assessment/Adult Learning/Learner Autonomy/Learner Controlled Instruction/Self-assessment

This report was designed to determine the student perception of the validity and reliability of learner-centered grading in a university flight training program. The target university planned to implement a newly developed learner-centered flight training syllabus and was uncertain of its effect on the student population. The university's existing flight training program utilized a traditional teacher-centered grading system and grade symbols with unknown results. The new system utilized a collaborative approach to lesson grading as well as objective, performance-based grade symbols. Using 7 research questions, this study sought to determine the student perception of the validity and reliability of each portion of the new grading symbols and the collaboration as well as the new grade symbols.

Using a 2-part qualitative and quantitative experimental research design, the researcher compared the student and flight instructor pre- and postexperiment perceptions of the validity and reliability of traditional and 2 separate and distinct variations of the learner-centered grading systems. From the literature, criteria to measure validity and reliability were developed and, with the assistance of a formative committee, incorporated into pre- and postexperiment survey instruments. Two separate grade instruments were developed and administered to the 2 experimental groups. A 3rd group utilized the current grade system in place at the university. Each group was administered the pretest and performed 5 iterations of the assigned grading system. Once these iterations were completed, each participant was immediately administered the posttest. The results were analyzed and reviewed by a summative committee. The results from the survey instruments were used to create the recommendations for this study.

The study revealed that student-instructor collaboration in the grading process as well as the addition of objective, performance-based grade symbols demonstrated statistically significant increases in perceived grade validity and reliability. The study produced 4 major recommendations. The primary recommendation was that the university adopt the learner-centered grading system described in the study.

Table of Contents

Chapter 1: Introduction

Assessment and grading procedures exert a significant influence upon student self-esteem and performance (Crocker, Quinn, Karpinski, & Chase, 2003; Holmes & Smith, 2003). In order for student assessment to exert a positive influence on student training, procedures should be valid and reliable (Salvia & Ysseldyke, 2007). Anecdotal evidence and some early statistical data suggested that serious shortcomings existed in these areas within the student assessment systems in use in the flight training curriculum of a major aeronautical university. As the university transitioned to a new form of flight training, it seemed prudent to examine the perceived validity and reliability of the current and future approaches to flight training assessment.

The flight training industry, at the behest of the Federal Aviation Administration (FAA) and in concert with several major universities, had begun a transition from a more traditional and pedagogical approach to flight training to an andrologogical approach. This learner-centered approach embraced constructivist theories that had entered the educational discourse in the last half of the 20th century (Knowles, Holton, & Swanson, 1998; Wright, 2002). The adult learner-centered approach placed a renewed emphasis upon student involvement across the entire spectrum of the learning process to include performance assessment and evaluation (Anderson, 1998; Stefani, 1998). University leaders made the decision to embrace this new learner-centered, FAA/Industry Training Standards (FITS) approach to flight training that included a learner-centered grading (LCG) philosophy (Connolly, Summers, & Ayers, 2005).

The setting for this study was a private, aviation-oriented university in the southeastern United States. The study focused on the validity and reliability of the assessment system used in the flight training program at the university. The flight training

program was the laboratory portion of the Aeronautical Science, 4-year degree program. Flight training students flew approximately 200 hours in small, single, multiengine aircraft as well as flight simulation devices and earned FAA approved pilot proficiency ratings. Approximately 960 flight students were taught by approximately 120 flight instructors at any given time. This yielded a student:instructor ratio of 8:1. The flight instructors who taught these students were generally recent graduates as well as 3rd- and 4th-year university students who had earned their FAA flight instructor certificates. The majority of this instruction was conducted in a one-on-one, student-teacher environment. Graduates of this program moved on to careers as commercial pilots in airline, corporate, and military flight programs.

Flight training at this institution as well as elsewhere was a highly scripted and regulated form of cognitive, affective, and psychomotor skill development that was subject to frequent performance assessment (Department of Transportation, 1999). Two types of assessment were accomplished. Internal (formative) assessment was accomplished on a daily basis by the flight instructor. In the course of a typical course of study, the instructor might grade the student 30 or more times. Each grading session consisted of several individual maneuver or event grades and an overall progress grade. Failure to achieve a passing grade in a lesson resulted in a reaccomplishment of that lesson.

The second form of assessment, external evaluation (summative), was accomplish one to three times in each course and carried greater significance. The external evaluation, usually referred to as a check ride or standardization flight, was similar to an academic final examination. During a single session, the student was required to demonstrate mastery of all the skills required in the syllabus (Department of

Transportation, 1999). An independent examiner normally conducted this evaluation.

Nature of the Problem

The problem that this study addressed was the failure of the assessment system currently in use in the flight training curriculum to provide valid and reliable feedback to students and instructors. Although flight instructors were given basic guidance on student performance assessment, the execution of the actual lesson grading appeared to be less consistent and predictable across different instructors and different periods within the training curriculum. Students who scored acceptably well in early training appeared to score poorly just prior to significant external evaluations. The university chief pilot and the assistant chief pilot were all too familiar with this pattern. Often, a series of satisfactory performances gave way to a series of unsatisfactory performances and repeated lessons as the student prepared for the summative flight check ride (I. J. Grau, personal communication, March 1, 2007; K. L. Byrnes, personal communication, February 28, 2007). Anecdotal evidence also suggested that individual differences in the understanding and application of assessment procedures may have resulted in grade variations between essentially similar student performances. This evidence suggested the presence of inconsistent and subjective grading behavior.

Holmes and Smith (2003) noted that students voice confusion at grades that appear increasingly subjective as they progress through the curriculum. Poor student perception of the validity and reliability of assessments may lead to reduced student self-esteem and motivation. Failures in these key areas may lead to reduced participation in the learning experience and reduced student performance levels. However, according to Kohn (1994), “supportive assessment” (p. 4) policies and procedures may exert a very useful and positive influence over the entire learning process.

Purpose of the Project

The purpose of this study was to conduct an evaluation study of student perception of the validity and reliability of the assessment tools and systems in use at a major aeronautical university flight program. This research provided an increased understanding of the assessment system in use and its effect upon the flight training program and student success. The study compared the current assessment system to a new form of flight training assessment that was soon to be adopted by the university. Students and their instructors were asked to evaluate three distinct assessment approaches to determine which system was perceived to be more valid and reliable.

Significance of the Problem

Wright (2002) noted that aviation education lags behind secondary and higher education in the development of new approaches to training. More specific to this study, little useful guidance was found to guide flight instructors and their students in current methods of performance assessment (Department of Transportation, 1999). Thus, students participating in the university flight training program used assessment instruments that evolved through common practice, rather than any purposeful scientific inquiry. Because of this, students and instructors within the flight program appeared to be susceptible to counterproductive behaviors such as grade negotiation, grade inflation, and grade irrelevancy that were found in other academic arenas that relied on traditional grading systems (Baines & Stanley, 2004).

The target university had entered into a grant with the FAA to implement the FITS training methodology within the university flight department. The FITS training methodology consists of three main components. These are scenario-based training, single pilot resource management, and LCG. Scenario-based training and single pilot

resource management had been the subject of significant prior research. LCG, specifically as it applies to aviation, was defined as a generic concept with little support in the literature. However, further research revealed significant LCG research in other educational disciplines.

The researcher was the chairman of the flight department at a major aviation university and was tasked with integrating the FITS training methodology and, more specifically, the new approach to assessment within the existing flight curriculum. This study attempted to understand and report on the students' perception of the validity and reliability of this assessment method before it was implemented department-wide.

The researcher reported to the dean of the college of aviation and, ultimately, to the university provost who provided support and encouragement for the study. Additionally, the results of this study were to be integrated into formal FAA training documents under revision through a grant to the university.

Research Questions

This study was guided by the following research questions:

1. What does the literature suggest about the validity and reliability of traditional grading procedures in aviation or other more conventional classroom education programs?

2. What does the literature suggest about the validity and reliability of LCG procedures in aviation or other more conventional classroom education programs?

3. What does the literature suggest about the appropriate method to determine perceived assessment validity and reliability through quantitative or qualitative means?

4. How should the perceived validity and reliability of flight student assessment programs be evaluated?

5. How do the participants (instructors and students) in the study rate the validity and reliability of traditional grading techniques? In this form of grading, the instructor assigns the student performance task grades using the traditional grading scale currently in use in the flight training department.

6. How do the participants (instructors and students) rate the validity and reliability of LCG techniques if a traditional grading scale is utilized? In this form of grading, the students self-assign performance task grades using the traditional grading scale currently in use in the flight training department. These data help determine if learner involvement in the grading methodology produces a separate effect from the actual grading scale used.

7. How do the participants (instructors and students) rate the validity and reliability of LCG techniques when objective performance grading standards are utilized? In this form of grading, the students self-assign performance task grades using the objective performance grading developed by the FITS research team. Because the grading scale and the grading methodology were modified simultaneously, this question, determined the combined effect.

Chapter 2: Review of Related Literature

Introduction

The literature surrounding educational assessment was reviewed in an attempt to proceed from the general to the specific and to attempt to respond to the challenges posed by the research questions. The literature was examined to define the competing educational theories that guided the development and deployment of educational programs. Once these underlying precepts were more clearly understood, the literature revealed the assessment strategies that responded to these educational imperatives.

In addition to this broader search, the specific assessment issues that surrounded aviation education and flight training were examined. New approaches and current research were examined to help define possible ways forward. Specific issues facing flight educators at the university flight training program were examined in order to help define specific problems addressed by the study and identify possible courses of action.

Bloom (as cited in Bloom, Hastings, & Madaus, 1971) identified two competing views of education which significantly influence assessment objectives, methodologies, and uses. According to Bloom et al., the first views education as a selection process in which those “fitted by nature” (p. 1) for increased educational opportunities are culled from those not capable of continuing. This traditional view leads to a relatively static curriculum, in which knowledge is a finite and constant standard to be attained successfully by the student. This view fosters assessment methodologies that tend to stress the lowest levels of the taxonomy, recall, and understanding (Gall, Borg, & Gall, 2003). These assessment strategies tend to be clear, easy to execute, and simple to defend. However, they may not be suitable in a more complex educational environment that values education as a process, rather than a static goal.

A second view of education focuses on developing the student and is committed to improvement of the process (Bloom et al., 1971). The purest expression of this form holds that the student is a full partner in the learning process and has a voice in the content, style, and direction of the process (Knowles et al., 1998). As stated by Brookfield (1986), this "self-directed learning" (p. 47) requires an assessment system that provides active feedback to the student and the educator, which is utilized to improve performance in real time. Although the more traditional educational view emphasizes initial and summative evaluation, the developmental model places increased emphasis on the formative evaluation (Bloom et al.). Rather than focus solely on what has already been learned, the assessment system must reveal what is still to be learned and how it might be accomplished more efficiently. The research identifies education as an active and dynamic process and assessment as a tool for continuous development of the student and the curriculum. In order to be effective, the assessment system must present a valid and reliable indication of student progress as well as curriculum effectiveness across the span of the curriculum. For the purpose of this study, the concepts of validity and reliability were adapted from those used in the assessment of quantitative and qualitative research.

Gall et al. (2003) defined validity as the "meaningfulness and usefulness of specific inferences made from test scores" (p. 640). Although this definition addresses quantitative and qualitative research, it is no less applicable to student performance assessment. If a lesson grade is to be a valid representation of the student's performance, it should be meaningful and useful. The grade should convey the level of performance in a manner that accurately reflects the student's achievement in terms the student understands and accepts. The literature gave voice to a general displeasure with the lack

of accuracy and precision in the traditional grading process as well as recent inflationary grading trends that appeared in higher education (Baines & Stanley, 2004). Thus, grade validity appeared to be a valid starting point for the study. However, grade validity may be of little value without reliability.

Reliability of the lesson grade describes the repeatability of the measure of the performance by multiple raters over time. It is often referred to as test-retest reliability (Gall et al., 2003; Salvia & Ysseldyke, 2007). In terms of the specific demands of flight education, the instructor should be able to conduct frequent formative evaluations in such a way that they meet the following criteria. First, a specific grade should represent the same level of performance, despite the presence of multiple iterations. Second, the grade should represent the same level of performance, despite the presence of multiple raters. Finally, an external evaluator should be able to observe the grades of several students and make meaningful comparisons between individual student performances and published performance standards. The style and content of the grading system may exert a significant impact upon the validity and reliability of the assessment system.

Anderson (1998) wrote of the differences between traditional and alternative grading approaches and identified a clear linkage between the grading schema in use and the style of learning exhibited. The line is drawn between a traditional approach to learning in which the student is a fairly passive participant and an active learning process in which the student and teacher work together to achieve the student's goals. The former creates clear lines of authority and, often, results in the teacher as lecturer, examiner, and assessor. The latter, often described as the constructivist approach, creates blurred lines of power and authority between student and teacher (Duffy & Jonassen, 1992). Stefani (1998) postulated that student assessment should become a part of the learning process,

and the responsibility for student grading should be shared by student and teacher. However, this new level of power sharing might prove difficult to sustain within the traditional grading framework established by tradition and practice.

Current Approaches to Assessment and Grading

Before delving deeper into the related subjects of grading and assessment, it might be wise to understand the underlying concepts and terminology. Speck (1998) defined the related concepts of marking, grading, and assessing. The term *marking* is often used to describe comments and notations designed to justify a specific grade and provide feedback to the student. Used in this way, the markings of a teacher or professor will most often lead to the awarding of a grade and may contribute to the evaluation or assessment of the student.

Marking. Marking may be viewed as a subset of or the supporting documentation for the actual grade that is earned by the student (Speck, 1998). Although a grade is not always accompanied by markings from the teacher, in those instances when it is, it may prove helpful to develop understanding between student and teacher. However, markings on papers can also carry significant emotional impact if delivered in a thoughtless, uncaring, or unenlightened manner (Gopinath, 2004).

Holmes and Smith (2003), in a survey of student opinions on faculty grading, noted that, although the lack of feedback was the most significant perceived problem with faculty marking, the lack of positive or constructive feedback was equally negative. In the same study, teacher objectivity was questioned by a significant segment of students as well. Thus, marking may viewed as more subjective than objective. However, the feedback provided by the marking process may well have the potential to provide a positive influence. The question of subjectivity may also hold true for the awarding of

grades.

Grading. Speck (1998) wrote of different languages of grading as defined by the positivist and constructivist theories of learning. In the realm of the positivist, grading is a purely objective, right or wrong construct designed to identify and rank students by their mastery of specific factual bits of data. The true-false test may be the ultimate expression of positivist grading in which the responses provided are simple, clear, and either correct or incorrect. The simplicity of this type of grading is obvious and comforting, especially for a teacher who might worry about the dangers of grade negotiation and external pressures to alter marks for at-risk students (Baines & Stanley, 2004).

The constructivist might see the process of grading as a more holistic part of the learning process and the grade a central part of the students learning experience (Speck, 1998). Much more about constructivist grading is included in the section on nontraditional grading. However, this mention is included to highlight the fact that the language of grading is often influenced by the lens through which the educator views their role and the educational model to which they subscribe. Thus, the traditional idea of the grade may be simply an observation of the familiar, rather than an objective survey of the entire spectrum of grading behavior.

One frequent concern is that a single grade may not be descriptive enough to reveal the truth about the complexities of student learning. Milton and Edgerly (1976) noted that their research consistently revealed that “unidimensional symbols (often letter grades) are used to report on multidimensional phenomena (learning)” (p. 48). A single unit of measure is asked to describe student aptitude, motivation, attitude, and interest as well as similar qualities on the part of the teacher. This places a significant amount of trust in the construction of the symbol.

The familiar symbols that identify a specific grade are not as simple or traditional as it might seem on first observation. A review of the descriptive terminology associated with specific student grade symbols from 120 nations around the world reveals wide variation and little unanimity (World Educational Services, 2007). For example, the A through F grading system, based on a mathematical scale of 100 points, is widely accepted and used within the United States. However, it appears to be used by only a handful of nations. Only Canada, New Zealand, India, and a few other nations ascribe to this model. In the Russian Federation, arguably one of the larger systems in the world, a 5-point scale topped by the grade of *otlichno* (or excellent) is the standard. Iran employs a 20-point scale, Denmark employs a 13-point scale, and Albania employs a 10-point scale (World Educational Services). This variation in grading systems demonstrates a distinct lack of unanimity and may leave significant room for improvement and innovation. Understanding the wide variation present in grading is important because it directly contributes to the assessment of student learning.

Assessment. According to Salvia and Ysseldyke (2007), assessment "is a process of collecting data for the purpose of making decisions about individuals and groups" (p. 4). In a macrosense, these decisions may involve financial aid eligibility, course entry, or decisions concerning continuation or termination within a specific academic program. At the microlevel, as stated by Stefani (1998), assessment policies and practices "have a profound impact on the attitudes students take toward their work, their learning strategies, and their commitment to learning" (p. 339).

The accuracy, transparency, and transportability of assessment data appear to have a positive or negative impact on the learning process. Some educators are concerned that ranking and grading behavior present in the classroom reduces creativity and limits

learning, especially if applied too early or unwisely by the teacher. Kohn (1994) represented this point of view and called for a minimalist approach to student assessment based almost entirely on student need, rather than on institutional priorities. Kohn referred to the creation of a positive "learning climate" (p. 3) that must be carefully studied and adjusted to produce the optimum conditions for learning and student support.

Stefani (1998) noted that, as workers become more mobile and self-sufficient in a steadily globalizing economy, fundamental changes to education and assessment are taking place. As individuals are becoming more responsible for designing and implementing their own lifelong educational process, they should become much more of a partner in the very assessment process that identifies their levels of success or failure. For example, students might assist in the design of the scoring rubrics that define their academic progress and work in partnership with the instructor to provide relevant assessment data (Choinski, Mark, & Murphey, 2003). Although these progressive views of assessment and grading are present in the literature, further examination of the current practice within the specific field of aviation education may prove useful.

Assessment and grading. Assessment and grading have been an integral part of aviation education since the Wright brothers established the first civilian flight school in Montgomery, Alabama, in the spring of 1910. Orville Wright, coinventor of the airplane and the first civilian flight instructor, soon discovered that a careful assessment of individual capabilities and personality traits yielded a much higher probability of success (Ennels, 2002). However, nearly 100 years later, the key FAA document that informs the practice of flight instruction says little about student assessment and grading (Department of Transportation, 1999). This document takes a pedagogical view of flight training. It focuses on behavioral and cognitive learning strategies and establishes the preeminence

of the flight instructor as the primary source of performance feedback. The handbook explains the role of the postflight critique in the learning process and encourages positive as well as negative feedback. Additionally, it acknowledges a role for limited student participation in the evaluation process. However, little useful guidance on student assessment or grading is contained within this document. To find additional guidance, one needs to examine the contents of the practical test standard (PTS) documents produced by the FAA.

The PTS lists the detailed requirements for the attainment of specific aeronautical ratings and certificates authorized by the government (Flight Standards Service, 2002). Each document consists of a series of tasks with a verbal description of the actual tolerances and characteristics required for successful completion. For example, a steep turn maneuver is required for the attainment of the private pilot certificate. The PTS notes that this steep turn must be accomplished in level flight and goes on to define level flight as plus or minus 100 feet from the altitude at which the student began the maneuver (Flight Standards Service). It also defines specific bank angles and airspeeds that must be maintained throughout the maneuver. Only one standard is provided for successful completion of a given task. Thus, students might maintain their altitudes within 1 foot of the desired altitude or within 100 feet of the desired altitude, and both would meet the standard provided for the task. According to Flight Standards Service, the PTS also requires, for any specific task, the student to "demonstrate mastery of the aircraft with the successful outcome of each task performed never seriously in doubt" (p. 8).

Although the PTS provides the tasks, standards, and general performance guidance required for specific flight course, it provides little useful guidance for how each task might be graded during the learning process (Flight Standards Service, 2002).

During learning, the student will most certainly fail to meet the standard and fail under the pass-fail guidance established by the PTS. Provided with this guidance, an instructor might be justified in awarding only a fully successful or unsuccessful grade for each task. Because few students master the complex cognitive, affective, and psychomotor skills required for flight until after significant actual practice, students could reasonably be expected to be scored unsuccessful during a significant portion of the learning process. This constant reinforcement of failure may produce a negative effect upon student self-esteem and self-image and an associated negative impact upon performance (Crocker et al., 2003). At the other end of the spectrum, the award of a successful grade for a clearly unsatisfactory performance, for the purpose of student motivation may produce equally unpredictable results. More research at the individual institution and syllabus level is required to understand fully the use and impact of the grading system at the operational level.

No evidence of a systematic approach to grading procedure development, deployment, or analysis could be found at the university. No definitive grading development documentation was identified during the study. The system in use appeared to have been created based upon tradition and experience within the university and from across the field of aviation education (I. J. Grau, personal communication, March 1, 2007; K. L. Byrnes, personal communication, February 28, 2007).

The university employed three grading scales within the flight program. First, grades were administered for each flight course by the flight department professor of record. Second, each end-of-course flight examination was graded by an FAA approved flight examiner. Finally, each daily flight lesson was graded by the individual flight instructor. Each level of grading was conducted with the use of a different scale.

Flight courses such as the FAA-designated private or commercial rating course were graded using a conventional letter grade scale in the same manner as other university courses. This grade was arrived at by the use of a matrix (Byrnes, 2007). The matrix, based on a 100-point scale, awarded 10% of the course grade based on the opinion of the instructor and an additional 20% based on student attendance. The remaining 70% of the grade was awarded based upon the student's performance on intermediate and final FAA required flight examinations.

The course-grading matrix placed a high value on student performance during the first attempt at the check ride. Each check ride failure resulted in a 10% or one letter grade reduction (Byrnes, 2007). Additionally, a student who failed to pass a flight check ride on the third try or failed to show up for a flight more than three times automatically failed the course (Byrnes). The system based 70% of the student grade on performance in a specified examination, 20% on student behavioral issues, and 10% on the instructor's opinion of the student (Byrnes). Individual lesson grades were not considered in the determination of the overall course grade and were determined by using a separate process. This grade system is presented to provide context but was not examined within the study.

The second grade system present in the flight department was employed by FAA approved flight examiners during required student end-of-course check rides. The standards grading system was a pass-fail system utilizing a *satisfactory*, *unsatisfactory*, and *incomplete* grading scale. This grade system, although meeting the intent of the FAA guidance, is also presented for context but was not examined in this study (Flight Standards Service, 2002). The third grading scale, the individual lesson grading system, was the focus for this study.

Individual lesson grades were determined by the flight instructor immediately following each flight, simulator, or oral recitation lesson (Byrnes, 2007). The specific criteria for each grade were provided in written form to the instructor although not to the student. Until the fall semester of 2007, the actual grading procedures, as depicted in Table 1, were not taught or presented in written form to new flight instructors (Byrnes). Thus, the instructors' experience as a student (most flight instructors were graduates of the university flight program) would appear to have been their sole resource for determining how to grade effectively. Each grade was characterized by a single word that summarized the grade.

Table 1

University Lesson Task Grading Scale

Grade	Description
Outstanding	The student performs the task within approved standards, never deviating to the limits of the standard, and demonstrates complete mastery of the aircraft
Good	The student performs the task within approved standards, sometimes deviating to the limits of the standard, with the successful outcome of the task never seriously in doubt.
Minimum	The student occasionally exceeds the limits of the approved standard, with prompt corrective action taken when the tolerance is exceeded.
Unsatisfactory	The student does not demonstrate satisfactory proficiency and competency within the approved standard.
Incomplete	The line item is not completed.

Two specific grades were associated with measurable consequences for the student. A grade of unsatisfactory required that the entire lesson be graded unsatisfactory.

Further study revealed that an unsatisfactory grade was the only administrative tool available to the flight instructor to request a repeat of the current lesson (Byrnes, 2007). Thus, the award of an unsatisfactory grade exerts a significant immediate financial impact upon a student because lessons are paid for individually by the student, rather than by tuition or fees, in addition to any emotional-, motivational-, or performance-related effect. Additionally, a grade of incomplete required the student to complete the individual missed task during the first portion of the next lesson (Byrnes). Repeating the task might also slow student progress and increase the cost of the flight course, although to a lesser degree than an unsatisfactory grade. However onerous, neither of these grades has any impact upon the final grade received for the course.

The grades of outstanding, good, and minumim denote more detailed levels of performance as measured against the standards required by the PTS as well as a general standard for overall mastery of the aircraft (Byrnes, 2007). Figure 1 illustrates this point.

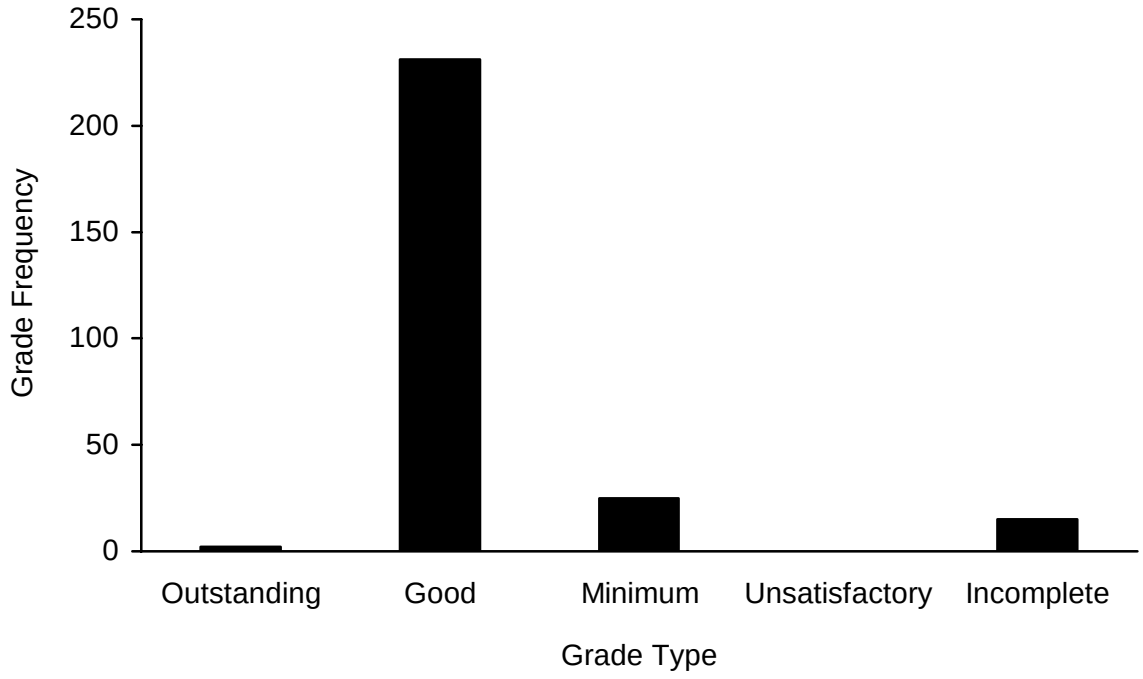

Figure 1. A graphic depiction of a small sample of student grades.

An examination of 20 randomly selected student records of flights that resulted in a satisfactory overall grade illustrated two predominate grade patterns that differed from what might be expected in a standard distribution of scores. The most common grade awarded to students appeared to be the grade of good that appeared to denote a wide variety of acceptable performances. This grade of good appeared in over 84.50% of lesson grades. At the other end of the distribution, the grade of outstanding appeared only twice in 271 separate grading opportunities or 0.73% of the time. This agreed with the observation of the flight department leadership. The university chief pilot noted that it is common knowledge that instructors used the grade of good as a default to signify any acceptable performance, regardless of quality (I. J. Grau, personal communication, March 1, 2007). The marginal grade denotes a less than acceptable performance and appears to serve as a warning to the student. Although no unsatisfactory grades appeared in this small sample, the role of the grade of unsatisfactory is, nonetheless, significant.

The university chief pilot noted that that the FAA requires a repetition of the lesson if a grade of unsatisfactory is awarded. He agreed that the grade of unsatisfactory appeared to be used to signal a requirement for additional training. The grade of unsatisfactory seemed to appear more frequently during those periods of the curriculum when an external evaluation was imminent. This second pattern of grading often emerged just prior to the instructor's recommendation for an FAA-required check ride. The award of a grade of unsatisfactory was immediately followed by additional student training until a grade of good was achieved at which time the check ride proceeded.

For example, Flight Unit 13 required the students to perform their first takeoffs and landings without the instructor on board the aircraft. The preceding lesson, Flight Unit 12, was the check ride by an external evaluator to determine the students' fitness for

this significant event. Thus, Flight Unit 12 was the last lesson in which an instructor could decide if the students were ready for the solo flights. The occurrence of the grade of unsatisfactory during Flight Unit 12 was more than double that for any other unit in the syllabus (see Figure 2), despite the fact that the students were graded on similar items during previous lessons. Thus, the grade of unsatisfactory appeared to constitute a request for additional training prior to a significant external evaluation as well as an objective or, possibly, subjective description of student performance.

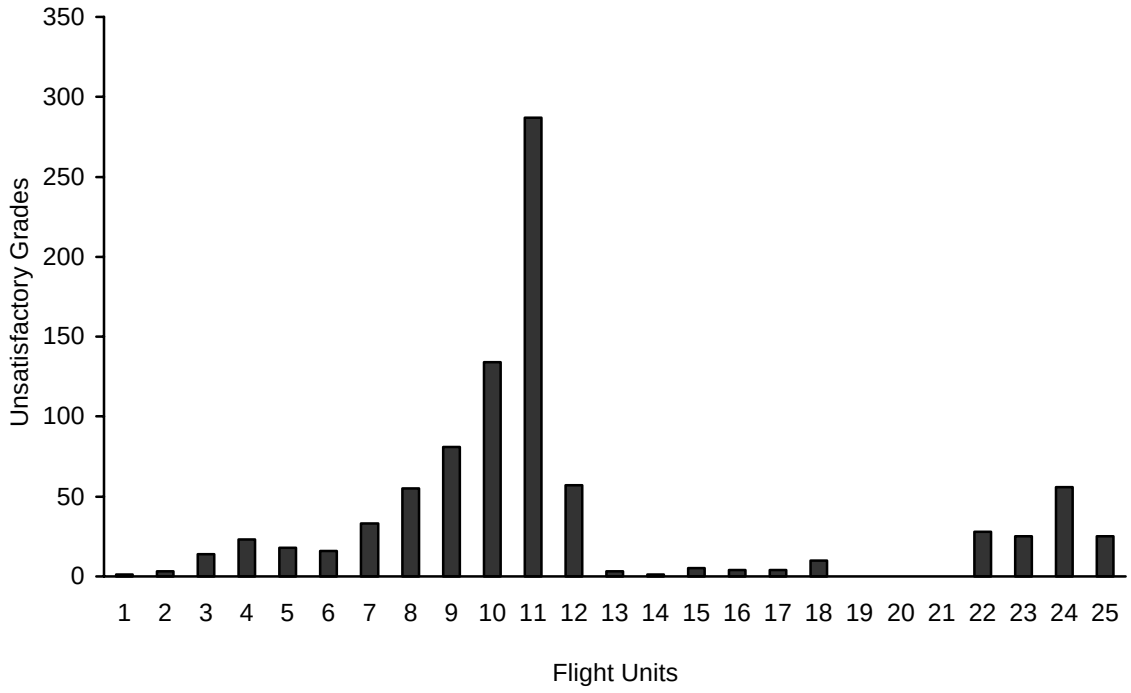

Figure 2. The number of individual lessons graded unsatisfactory by flight unit.

The grading patterns illustrated in Figures 1 and 2 raised significant questions about the purpose; validity; and, to a lesser extent, the reliability of grading in the flight department. Although the grade system may have had some input into the student learning process, it appeared to be more closely associated with the administration of the program (Hendrickson, Gable, & Manning, 1999). Grades appeared to be utilized by the individual flight instructor to motivate students as witnessed by the award of acceptable

grades early in the curriculum. Later in the curriculum, improved performances were often deemed unacceptable. Additionally, the grade of unsatisfactory appeared to be utilized as a de facto administrative tool to request additional training prior to significant events such as student solos or standardization flights. From these anecdotal data, one might reasonably draw the conclusion that the grade system present in the flight department was not solely dedicated to the purpose of documenting and supporting student learning.

Figures 1 and 2 suggest remarkable unanimity in grading procedures across the flight department. Although flight instructors appear to be reliable in their application of the grading system, this initial data suggested questionable validity of the actual task grades across the curriculum (Gall et al., 2003; Salvia and Ysseldyke, 2007). The university *Flight Instructor Orientation Handbook* set forth distinct standards for student grading (Byrnes, 2007). However, the anecdotal data presented, questioned the validity of these standards in practice.

Another practice observed in flight department grading behavior was the requirement that the instructor grade the student. However, relatively recent research identified the field of flight training more closely with a learner-centered and androgogical approach and made a case for increased learner participation in the assessment process. This program, begun in 2003, is known as FITS and has since become an industry standard for flight training (Connolly et al., 2005; Knowles et al., 1998).

Nontraditional Approaches to Assessment and Grading

A primary goal of the FAA (2003) FITS research effort is to enhance the general aviation pilots' aeronautical decision making, risk management, and single pilot resource

management skills. This involves the application of knowledge to a variety of ambiguous situations. Gagne, Briggs, and Wager (1992) theorized that this type of problem solving may be best taught by providing the student with a "larger and better organized knowledge base" (p. 72). The FITS approach seemed to indicate that the greater the experience and knowledge about the system, the greater the probability of success in problem solving. However, Gagne et al. expressed some doubt that these "executive or metacognition strategies [can be taught, instead, theorizing that learners develop them from a] variety of task oriented strategies" (pp. 74-75). These strategies pose relevant questions for those who desire a relatively simple approach to student knowledge attainment and performance assessment.

Anderson (1998) examined the beliefs and assumptions that underlie assessment and student grading and found two schools of thought. The traditional assessment school views learners as relatively passive participants in the learning process. Their primary source of learning is rote memorization of related and unrelated bits of knowledge. According to Anderson, assessment tools are developed to measure a representative sample of these "discrete and isolated" (p. 8) bits of knowledge. The entire assessment process is designed to rank students in a rigidly hierarchical system in which the instructor designs the assessment tool; identifies the standard; and, then, rates the students.

Milton and Edgerly (1976) found that these same properties limit the usefulness of student grading to little more than a record-keeping function. They called into question the value of this traditional approach for any sort of predictive use within the learning process. In one particularly vivid example, they cited an experiment in which students who had all received an A grade in a prerequisite course were prestratified by grade level

into a single class section. Yet, their final grades in the new section exhibited the same norm-referenced A through F grade distribution as other nonstratified sections containing a more normal distribution of students. When challenged on this outcome, the experienced professor found it difficult to believe that any change was required. This anecdote suggests that habit and experience appear to be significant determinants of grading behavior.

Another opinion and the one under consideration in this study is that a constructivist approach to learning may provide a better way to teach problem-solving skills and improve overall student learning (Duffy & Jonassen, 1992). Constructivism revolves around the development of a mental model or schema constructed by exposure to a realistic and complex environment. Learning occurs as the student explores the new environment with the guidance and council of the instructor or teacher. When adopted, the relationship between student and teacher changes significantly (Anderson, 1998). The two become collaborators in the learning experience that includes instructional and assessment strategies. Ideally, student and instructor become a team devoted to improving the learning process. The alternative assessment strategy that accompanies this approach to learning differs sharply with the more traditional methods described previously.

In the constructivist approach, assessment becomes an active component of the learning process. Grading is repurposed as a facilitator, rather than as a discriminator. The teacher and the student share in the task of learning assessment, building on the partnership aspects of collaborative learning, and taking advantage of the student's unique view of their own progress. Table 2 compares the two strategies. The increased emphasis on learning requires formative evaluation opportunities designed to predict

performance, rather than measure outcomes. Underlying all of this is the concept of power sharing between teacher and student (Anderson, 1998). Table 2 illustrates the difference between the two philosophies.

Table 2

Traditional Versus Alternative Assessment

Philosophy and assumptions	Traditional assessment	Alternative assessment
Learning strategy	Passive	Active
Purpose	Document learning	Facilitate learning
Abilities	Focus on the cognitive	Focus on all 3 domains
Assessment	Objective	Subjective
Power and control	Teacher centered	Shared
Process	Generally summative	Formative and summative
Learner-teacher collaboration	Fosters competition	Fosters collaboration

Note. From "Why Talk About Different Ways to Grade? The Shift From Traditional Assessment to Alternative Assessment" by R. A. Anderson, (1998). In R. S. Anderson & B. W. Speck (Eds.), *New Directions for Teaching and Learning: Changing the Way We Grade Student Performance. Classroom Assessment and the New Learning Paradigm* (pp. 5-16). San Francisco: Jossey-Bass.

One approach to a constructivist learning schema involves the application of well-designed flight scenarios that enable a student to construct an effective decision-making model (Connolly et al., 2005). This approach would appear to be most effective if flight students actually fit the psychological model of adult learners. Knowles et al. (1998) described several characteristics that separate adult learners from the more common field of pedagogy.

The primary characteristics of adult learning revolve around the more sophisticated self-concept, motivation, and orientation to the learning process of the learners (Caffarella, 2002). The adult learners may approach learning with a desired outcome in mind and come to the learning experience with some idea of how they might partner with the teacher or exert some control over the learning process (Knowles et al., 1998). Additionally, the learners bring life experiences and a readiness to learn, usually not observed in the pedagogical learning situation. Although there is some disagreement over the specific adult learning concepts, many scholars agree that the characterization of the individual learner has less to do with their chronological age and more to do with their self-concept and orientation to the task (Brookfield, 1986). One could make a reasonable, although oversimplified, assertion that the adult learners learn because they want, need, or desire to, whereas the pedagogical learners learn because they are required to. Flight training, by its very nature, appears a better fit with the former description. There are many reasons for this observation.

First, flight training is an expensive undertaking, with costs approaching $300 a lesson (university students average three lessons a week) with yearly student costs exceeding $15,000 per year, not counting tuition and room and board (Embry Riddle, 2007). Students voluntarily enter this expensive course of study, and most, if not all, attempt to exert significantly more influence over the pace and content of their training than traditional classroom students (K. L. Byrnes, personal communication, March 1, 2007). The learning is conducted in a one-to-one, teacher-student setting. Because of this close relationship, flight students have an increased opportunity to influence the course of study, especially when compared to that individual student in a large lecture hall setting (Department of Transportation, 1999; I. J. Grau, personal communication, March 1,

2007). Finally, flight instruction possesses an active psychomotor learning environment not generally found in traditional classrooms with an immediate feedback loop designed to improve performance in real time (Department of Transportation). Successful completion of the flight tasks may satisfy the upper level (self-fulfillment and intellectual stimulation) and lower level (safety and security) needs of the learner (Cafarella, 2002; Knowles et al., 1998). Those students who are drawn to this expensive, satisfying, and very personalized form of learning would appear to fit most definitions of the adult learner. If a nontraditional approach to learning might be required for this group of students, as the FITS program asserts, then, a nontraditional approach to the area of student performance assessment and grading may be appropriate as well.

Learner-Centered Assessment and Grading

Stefani (1998) noted that, for students to become "autonomous, independent, and reflective learners" (p. 339), they must develop self-assessment skills. She proposed a partnership between teachers and learners in which the students take an equally active role in assessment and grading. This approach immediately satisfies some of the major student criticisms of assessment relating to perceived arbitrary assignment of scores, disrespectful grading techniques, and incomplete information used to assign grades (Holmes & Smith, 2003). On the other hand, student self-assessment opens a discussion of learner objectivity and accuracy. This discussion may be addressed by a collaborative approach to the grading process that realizes that the actual purpose of the grade is to assist in the learning process (Boud & Falchikov, 1989; Kohn, 1994; Stefani). Although the question of methodology may have become a bit clearer, other voices have questioned the validity of the grade itself.

Butler (2004) argued that comments that truly reflect student performance may be

more meaningful without the assignment of a letter grade. According to Butler, this "comments only" (p. 37) approach to assessment removes the emotional stigma from the student and provides for a more mature reflection upon the competency of the student. Freed from the use of narrowly defined letter or numerical grades, the teacher is theoretically able to describe more accurately the student's actual performance. Although this approach might not be as useful in the highly regulated field of flight training as it is in grading an essay, it does beg the question, How does the actual grade support the purpose of the grading process?

Kohn (1994) agreed to some degree with Butler (2004) as he identified grading with three distinct goals: sorting, motivation, and feedback. He found none of these are as meaningful to the learner as the descriptive feedback that often accompanies the grade. In point of fact, the letter grade that may be of most value in the sorting role actually appears to have a negative impact upon the motivation of the learner. Lack of validity in traditional grading may further exacerbate the student's emotional response to grading. Noting this trend, Baines and Stanley (2004) called for a more frank assessment of student performance separate from the grade negotiation (also known as grade shopping) that has become a part of classroom life.

In the realm of aviation education, the role of self-esteem has been examined and found to be a significant factor, along with actual task proficiency, in determining student success (Davis, Fedor, Parsons, & Herold, 2000). The instructor-driven nature of traditional aviation grading appears to have a measurable positive or negative impact upon student self-esteem. Students often view grades as a reward for obedience, a method of obtaining self-esteem, or a gateway to scholarships and advanced training (Baines & Stanley, 2004). Additional research on this link between self-esteem and grading supports

a thesis that traditional grading techniques and procedures may have a profoundly negative impact upon students, especially those in more adult-oriented education situations (Crocker et al., 2003). More recently, Blickensderfer and Jennison (2006) examined the effectiveness of learner-centered debriefing techniques and concluded that they were more effective and well-received by students than more traditional teacher-directed methodology. Many thoughtful observers seem to agree that, as a minimum, the learner should be much more involved in the assessment process (Boud & Falchikov, 1989).

Some of the more interesting research on assessment methods exists within the arts. Merrill (2003) wrote of the value of recording musical ensembles to encourage meaningful self-reflection during multiple playbacks. Although music and flying may seem worlds apart, they are more similar than one might think. Both consist of the application of cognitive and psychomotor skills in concert with other participants (in the case of aviation, other pilots and air traffic controllers), and both contain significant elements of personal technique. A musician garners criticism for a poorly played passage in the same way an airline pilot may later hear about a rather hard landing from passengers. However, only when musicians begin to reflect upon their own work, do they possess the tools to improve performance (Merrill). Finding ways to reflect upon one's performance may be very valuable for student pilots as well as musicians. The question remains, How best to do it?

There appears to be a requirement for some structure within collaborative assessment as well as self-assessment. Gopinath (2004) noted that students often have trouble with assessment because their teachers neither understand the criteria nor know how to apply it. Before flight students can be expected to practice self-assessment, the

grading methods and scales must be clear and meaningful to instructor and student alike. This appears to lend support to the idea of objective performance standards that remain valid and reliable throughout the curriculum. A recent study that examined student perceptions of faculty grading techniques and procedures at a major midwestern university informs this discussion.

Holmes and Smith (2003) found that students and professors “differ in their perception of the meaning of grades” (p. 318). They noted that grades have a motivational role that goes well beyond mere performance assessment into the areas of learner involvement and participation. Holmes and Smith also observed that students may be either “grade oriented or learning oriented” (p. 319). The conflict between these two orientations may prove confusing to the student and teacher. However, the biggest irritant surrounding grades appeared to be the issue of fairness. Student survey results supported the assertion that unreliable or subjective grading and lack or real feedback by professors are the biggest irritants and roadblocks to learning. This issue of fairness speaks right to the heart of grade validity and reliability.

According to Connolly et al. (2005), based on the literature, FITS researchers attempted to develop a more reliable, and valid, LCG schema based on “desired student outcomes” (p. 8) within the domains found in aviation. The schema provides an objective framework for assessing student performance during all phases of the curriculum. It requires collaboration between the student and the instructor to arrive at an accurate assessment of performance. Others have approached this problem as well.

Hendrickson et al. (1999) discussed two approaches to the specific grade rubric that might be developed for a collaborative grading system. A norm-referenced approach may be used to arrive at an overall grade for the lesson. This schema might examine the

level of student preparation, participation, and overall effort as compared to other students within the curriculum. This approach recognizes the students' emotional and motivational needs and may allow the specific task grades to be more specific to student performance. The students are allowed to discover the reasons for their progress or lack thereof in a less threatening way. This overall lesson grade that might be more tolerant of individual failures on specific task grade should be considered a rating of level of effort and a tool for student motivation. However, within the specific lesson tasks, a criterion-referenced approach to task grading in which the grade is determined against a set standard allows the student to understand where they are and how far they have to go (Hendrickson et al.; Kuisma, 1999; Salvia & Ysseldyke, 2007). This allows a more realistic assessment of actual student performance by the instructor as the students progress toward their training requirements.

The FITS program approaches this problem through the use of a set of objective and descriptive grades as described in Table 3 (Connolly et al., 2005). The specific scale used assigns a descriptive grade that identifies the level of performance demonstrated by the student. A key indicator of success in flight training is the ability of the student to fly solo without assistance from the instructor (Department of Transportation, 1999). A performance-level descriptor that reflects required proficiency for unsupervised flight is utilized in the FITS methodology to describe the highest level of performance. This level, represented by the perform grade, sets a realistic expectation that the students performance will not be perfect. Rather, it describes a student who is constantly detecting errors and corrects them without assistance from the instructor (Connolly et al.). This is a significant requirement for solo flight.

The remainder of the grades, practice, explain, and describe, is meant to describe

objectively the students' cognition and performance of the required tasks and maneuvers. For example, at the practice grade level, the student will require active assistance from the instructor to complete the graded item.

Table 3

Sample Federal Aviation Administration Industry Training Standards Learner-Centered Grading Scale

Grade	Description
Perform	At the completion of the lesson, the student will be able to perform the activity without assistance from the instructor. Errors and deviations will be identified and corrected by the student in an expeditious manner. At no time will the successful completion of the activity be in doubt.
Practice	At the completion of the lesson the student will be able to practice the scenario activity with little input from the instructor. The student with coaching and assistance from the instructor will quickly correct minor deviations and errors.
Explain	At the completion of the lesson the student will be able to explain the scenario activity in a way that shows understanding of the underlying concepts, principles, and procedures that comprise the activity.
Describe	At the completion of the lesson the student will be able to describe the physical characteristics of the scenario activities.

The explain grade denotes a point at which the student understands and can verbalize the requirement but cannot perform it, even with assistance from the instructor. Finally, the describe grade denotes a condition in which the student can neither understand nor perform the task or maneuver but can describe its basic characteristics (Connolly et al., 2005). These grade descriptions have been in limited use since 2004 but have yet to be subject to any rigorous scientific examination. They represent an early attempt to develop an objective system that might accurately describe student

achievement in terms of the student's demonstrated cognitive and psychomotor abilities.

The FITS schema was due to be deployed in the university flight department during the 2007-2008 academic year. The researcher obtained permission to study and evaluate the deployment of this grading system or a similar approach from the FAA and the university. The collaborative grading methodology and the objective grade descriptors were put to the test beside variants of the existing traditional grading system. All involved understood that the research effort would evaluate the student and instructor perception of the grading process as well as independent grading scales to determine the causal factors for any observed behavioral changes. The study examined grading practice in terms of the related concepts of perceived validity and reliability.

Assessment Validity

Gopinath (2004), in a study of case study grading, found that, even with well-established criteria, two skilled graders may exhibit significant variation in the grades awarded for the same work. Although this is a problem with assessment reliability, the researcher found the primary issue was a lack of clarity and understanding between student and grader about the meaning of the individual grade. This resulted in a lack of individual grade validity. Four distinct factors emerged as causative factors. First, although the criteria for grading were established in writing, it was interpreted differently by each rater. Second, the actual letter grade was viewed as a reward or penalty by some graders and was often adjusted due the expected emotional impact (grade impact) on the student. Third, despite clear written standards, raters continued to compare the marks given to one student to the marks given another (relative grading). Finally, errors of omission caused differences in grades awarded (Gopinath). It is only a small leap to assume that a young flight instructor in a busy flight school environment might fall

victim to similar problems. These parallel weaknesses of grade impact and relative grading are addressed by additional research.

As previously noted, Butler (2004) advocated for “comments only” (p. 37) marking of written work. He concluded that the absence of a grade makes the assessment of the students work more accurate and useful, devoid of the emotional baggage associated with the grade on the part of student or instructor. Although students in the study initially bemoaned the lack of a letter grade, they appeared to embrace the emphasis on constructive criticism and adjusted their work accordingly. Thus, the validity of the grade might actually be improved by deleting the grade itself.

Meaningful parallels may also be drawn between the concept of validity as it relates to research and the same concept as it related to student assessment. Schaffner, Burry-Stock, Cho, Boney, and Hamilton (2000) discovered, in a study of elementary student perceptions of teacher grading processes, that examining classroom assessment procedures and outcomes helped redefine the word validity as it related to grades. Schaffner et al. noted that, “since grades are an accumulation of assessments, it seems to us (the authors) that that the validity of the various assessments and the grading process are important” (p. 10). Beyond a mathematical construct, they defined assessment validity as a process that adds value to the learning environment by understanding the social consequences, accuracy, meaningfulness, and fairness of the assessment system. Schaffner et al. went on to say, “Putting the child into the teaching, testing-grading cycle is critical to understanding the validity of the assessment process” (p. 11). This would appear to support increased student participation and grade symbol clarity as methods to improve validity.

One method to improve symbol clarity and student involvement is through the

concept of construct validity (Salvia & Ysseldyke, 2007). It is vital that the grading system is designed to measure those characteristics vital to success in the task. The teacher is traditionally assumed to be in the best position to observe and report on student progress. Yet, the individual most involved in the learning process, the student, may have a much clearer understanding of the subjective portions of the learning process and should be heard. Thus, an assessment instrument should be carefully designed to extract the most accurate assessment information from student and instructor and report on it in a meaningful and actionable manner (Salvia & Ysseldyke).

Individual flight lesson grading is one in a series of formative evaluations leading to a summative (criterion-based) examination, most often referred to as a check ride. In order for these formative evaluations to result in a successful end-of-course evaluation, the grading system should exhibit both concurrent and predictive validity (Messick, 1989). Restated, the daily lesson grade should give student and instructor a clear idea of the student's present position and predicted future performance against the established criteria. The grading pattern presented earlier in Figure 2 would appear to exhibit little predictive validity early in the training program (when passing grades predominated before students successfully completed tasks and maneuvers) followed by a dramatic increase in predictive validity (as grades begin to describe actual performance) just prior to the required summative evaluation. This grading pattern would appear to exhibit little concurrent or predictive validity and might prove to be somewhat confusing to the student. The challenge for the curriculum developer is to design an assessment system that clearly links student performance to required performance criteria. One method that might be useful in increasing construct validity as well as student instructor understanding and cooperation is the rubric.

The literature contained much information on the role of rubrics as a method to improve grade validity. Choinski et al. (2003) examined the use of rubrics as a means to inject objectivity into the grading of information resource research projects. Their research found the development of the rubric appeared to parallel the development of the syllabus and became an iterative process. Thus, feedback obtained through use of the rubric as a means to express course objectives and student performance was integrated directly back into the curriculum (Choinski et al.). This formative assessment process would appear to make the entire learning process from assignment development to student assessment more valid.

Similarly, Lunsford and Melear (2004) suggested that the use of rubrics in nontraditional education (scientific inquiry). In this application, rubrics improved grade validity and student performance by increasing the accuracy of the information, as well as the assignment expectations that are shared between student and teacher. According to Lunsford and Melear, the three step process they espoused defines the end product, identifies the correct “criteria and weight” (p. 37), and takes the additional step of identifying who should assign the grade. Self-evaluation, peer evaluation, expert judging, and similar related concepts are all presented as ways to improve the validity of the assessment process. There is ample room for integration of many of these ideas into aviation grading. The basis for a grading rubric already exists through the deployment of the PTS (Flight Standards Service, 2002). These standards for performance, although not developed jointly between student and instructor, lay out the end objectives required for success as the literature suggested (Choinski et al., 2003; Lunsford & Melear).

As previously noted, no specific assessment methodology is required by the FAA for formative evaluations. In essence, the rubric is incomplete. With the addition of a

clear and definable assessment system, the PTS can develop into a learner-centered rubric. However, the designer of an aviation assessment system has two related problems to solve. First, as previously discussed, the validity of the individual grade must be addressed. Second, any system of formative assessment must be designed to create a high degree of assessment reliability across many student and instructor pairings.

Assessment Reliability

Reliability, in terms of case study research, is described, according to Gall et al. (2003), as the "extent to which other researchers would arrive at similar results if they studied the same case, using exactly the same procedures as the first researcher" (p. 635). If the performance of a flight student can be viewed as the subject of a case study, then, given nearly identical student performance, different flight instructors should be expected to arrive at very similar assessments. If LCG is used, collaborative pairs of instructors and students should achieve the same result. Thus, if flight student assessment was to follow the scientific model, lesson grading should be considered successful if it is accurate and repeatable.

Meaningful parallels may be drawn between the concept of reliability as it relates to research and the same concept as it relates to student assessment. Feldt and Brennan (1989) discussed four relevant sources of measurement error that lead to a lack of reliability. The first of these consists of random variation in human performance. Individual humans perform at different levels of effectiveness. Thus, a grading system should be sufficiently clear, easy to use, and tolerant of human differences in its application and interpretation. Second, humans who might perform reliably in a similar situation, exhibit less reliability in different situations and scenarios. Thus, in the quest for greater reliability, the location and setting of the grading process might be examined

(Feldt & Brennan). Evaluator subjectivity is the third source of measurement error and is resident within the flight training program. Conversations with the chief pilot revealed that individual instructors exhibited very different views of student performance based on their previous experiences as students, their time as an instructor (he noted that instructors grade harder as they gain experience), and their position within the department (I. J. Grau, personal communication, March 1, 2007). A clearer system of grade descriptors may reduce the levels of subjectivity.

The reliability of grades may be subject to significant error across graders. Kohn (1994) noted that identical work submitted at different times to the same teacher or at the same time to two different teachers will often be graded differently. Kohn described grading as "a subjective rating masquerading as an objective assessment" (p. 2). This conflict between objective and subjective grading appeared often in the literature and appeared to exert a negative effect on grade reliability.

According to Feldt and Brennan (1989), a final cause of measurement error is "instrumental variables" (p. 107). Although this may refer to the error present in an electronic or mechanical instrument, it may also be present in the design and application of the instrument (or process) used to conduct the evaluation. Thus, if a poorly designed assessment instrument acts in a similar manner to a poorly designed measurement device, it may produce measurement error and unreliable data. A well-designed instrument should have an equal and opposite effect.

Research Methodology

A review of the literature led to the decision to utilize a pretest-posttest control group design that compared the experiences of three distinct groups of student-instructor pairs during an identical segment of the instrument flight simulation device training

conducted at the university (Gall et al., 2003). Three groups were required to accommodate a control group as well as two different but related experimental treatments. An experimental approach was selected due to the specific and measurable nature of the variables involved and the opportunity to hold others variables in check within the highly regulated flight training program. The survey instrument was designed to measure the student and instructor perception of validity, reliability, and overall effectiveness of three unique assessment methodologies. This research design facilitated a direct comparison of the effect of the type of assessment system employed on participant attitudes about grade validity and reliability.

Gall et al. (2003) noted that the pretest-posttest control group design effectively controls for threats to internal validity such as “history, maturation, testing, instrumentation, statistical regression, differential selection, experimental mortality, and selection-maturation interaction” (p. 405). Because the entire experiment was conducted within an approximate 3-month time period, the opportunity for other unplanned historical variables or participant maturation was greatly reduced. However, due to the high rate of turnover among flight instructors, experimental mortality might have been an issue, even in this short experiment. In the end, it turned out to be a relatively minor issue. Experimental mortality was addressed in more detail as the methodology was reviewed and the instruments and experiment were designed.

A review of the literature was conducted to find a valid survey instrument that might be used to measure the perception of validity and reliability of the grading process. Although no off-the-shelf instruments were found, several instruments were discovered that contributed to the development of the final survey. Schaffner et al. (2000) developed the Perception of Assessment of Teachers by Students Survey instruments that were used

to examine the attitudes and perceptions of 276 grade school children about the grading practices of their teachers. The instruments were validated by subject matter experts and pilot tested on 80 high school students. Schaffner et al. noted the "psychometric qualities of both versions, of both instruments, appear sound" (p. 10), and their study achieved statistically reliable results. The instrument utilizes a modified Likert scale (one version was modified with pictorial responses so early elementary school students might better understand the scale). The survey questions are simple, affirming statements that inquire about the fairness, validity, reliability, emotional impact, and student involvement in the grading process. The survey found strong correlation between fairness, internal motivation (locus of control), and student involvement in the assessment process.

Wooley and Wooley (1999) developed a survey to measure "teachers beliefs about teaching related to behaviorist and constructivist learning theories" (p. 3). The parallels between this work and the survey being created for this research are many. For example, the survey measures the teacher's preference for collaborative learning in conjunction with students and interested parents. It introduces the concepts of control into the survey as an indicator teaching style. The questions developed measured the differences between a teacher centered (behaviorist) and a learner-centered (constructivist) learning paradigm. This survey construct contributed much to the final survey design process.

The survey employs four categories, comparing the participants' beliefs about the two teaching styles, the teacher's management style, and the role of parents (Wooley & Wooley, 1999). Survey validation was conducted through a pilot group. Strong item correlation was achieved by the use of positive statements of belief. This survey instrument provided an excellent model for the one used in this study. However, it raised

questions about the accuracy of self-reporting by the participants.

The validity of self-reporting has been called into question due to errors of omission or commission on the part of the participants. Shim, Felner, Shim, and Noonan (2001) developed a classroom instructional practice scale to determine the various styles of instruction present in the classroom. It correlates positively with this study of instructor student grading behavior in the flight line classroom. According to Shim et al., this study of validity and reliability of self-reported data was based on the responses of "more than 25,000 teachers, in over 1,000 schools, across 5 years" (p. 8). The authors noted that grouped items such as the categories employed by Wooley and Wooley (1999) provided more reliable indicators. Shim et al.'s study found a "significant relationship" (p. 7) between teacher self-reporting and student reporting of the same event. A survey instrument based on this format could expect to achieve similar results, identifying the correlation between perceptions of validity and reliability in the grading process between instructor and student. This work provided relatively clear evidence to support the use of participant perception of validity and reliability in the survey instrument.

Summary

A review of the literature suggested that there is general agreement about the problems associated with student assessment and grading. Validity and reliability were called into question in various forms of student grading and assessment from the classroom to the music ensemble (Baines & Stanley, 2004; Merrill, 2003). Flight training, as witnessed by the development of the FITS program as well as the anecdotal data, appeared to be little different (Connolly et al., 2005). Although the idea that the role of grading has significantly changed from one of evaluation and sorting to one of maximizing learning has been around for several decades, the actual practice of grading

appears to have changed little over time (Michaels, 1976). The literature revealed considerable information about grade school, secondary school, and high school assessment. Little, if any, detailed empirical research was found that examined these concepts in aviation education. In the specific area of aviation training, no significant research could be found that challenged the currently held assumptions about the grading process and its impact upon the actual training of flight students.

Measuring the demonstrated validity and reliability of actual grading schema is a useful goal. Unfortunately, the time required for that level of effort was beyond the scope of this research effort. However, the literature showed that student and teacher perception of the validity and reliability of the grade schema would prove valid indicators of worth and effectiveness (Holmes & Smith, 2003; Shaw, 2004; Stefani, 1998). Thus, the challenge was to provide variations on the grading schema that incorporated a learner-centered approach and completed the partial rubric formed by the PTS documents. Once developed, they were deployed and tested to determine the perceived validity and reliability of each approach. From these data, reasonable conclusions were drawn for the way ahead.

Chapter 3: Methodology and Procedures

Introduction

The research methodology was a two-part qualitative and quantitative evaluation. The study consisted of eight procedures and utilized a pretest-posttest control group design that compared the experiences of three groups of student-instructor pairs during a segment of the instrument flight simulator training curriculum (Gall et al., 2003). Three groups were required in order to accommodate a control group as well as two different but related experimental treatments. An experimental approach was selected due to the specific and measurable nature of the variables involved and the opportunity to hold others variables in check. This research design facilitated the direct comparison of the effect of the type of assessment system on participant attitudes about grade validity and reliability.

Quantitative methods were used to evaluate the qualitative data obtained from the participants concerning the validity and reliability of the respective grading systems. The specific research questions were addressed through a review of the literature, the creation of the experimental treatments, the development of the survey instrument, and the collection and analysis of the data.

Participants

Two separate groups of participants in the study executed the experiment and provided independent feedback through the survey instrument. The first group consisted of approximately 73 instrument flight training students (64 actually completed the experiment) in the university training program. These students were expected to range in age from 18 to 22 years with an average age of 20 years. Participants were randomly selected from an instrument student pilot population of approximately 250 students.

Based on the Fall 2006 figures, participants were expected to be approximately 16% female. Eight percent of the students were expected to be of international origin. All of the participants spoke and read English, and most were 1st- through 3rd-year college students.

The researcher selected the student participants through the flight department scheduling and assignment system from all students enrolled in the instrument flight curriculum. These names were used to advertise an initial meeting and conduct a random drawing of candidates. The resulting candidates were invited to participate in the study.

The second group of participants was the flight instructors assigned to teach the first group of participants. This group was randomly selected based on their assignment to the student participants. Thirty-four flight instructors began the experiment, and 32 actually completed it.

Instruments

The researcher developed a single instrument to serve as a pre- and a postsurvey of student and instructor attitudes about the three different grading methods. The survey utilized a Likert scale to measure degrees of agreement with 38 positive statements divided into eight sections. Thirty questions were administered to all participants. Consistent with the literature, the survey instrument measured the participants' perceptions of validity and reliability as well as the related areas of collaboration, emotional impact, and overall impact and importance of the grading schema. Grading is a complex emotional and cognitive phenomenon, and these additional categories helped determine the participants' overall attitudes about grading, which may have had an impact on their views of validity and reliability. An additional section of the survey was administered to the second and third groups to measure the impact of the specific

collaborative and LCG techniques.

The researcher developed the instrument with the assistance of a formative committee, consisting of the university chancellor and two members of the FITS research team who were familiar with the research. The formative committee reviewed drafts of the survey and accompanying literature and provided written feedback. Two face-to-face meetings were held during the course of the survey preparation. The formative committee made several recommendations to reduce the total number of statements, add consistency to the remaining statements, and clarify the relationships between validity and reliability and the survey outcomes.

Once the survey was edited, the instrument was pilot tested on a representative sample of 10 senior instructor pilots. Respondents were asked to take the survey based on their experiences with the current grading scale and grading procedure in use. The surveys administered to the pilot test group were analyzed with a single sample test of means to identify any errors in the construction of the instrument. Fifteen of the original questions showed significant variation and were eliminated or edited for style and content. Additionally, all pilot test participants were asked to comment on the content, grammar, and clarity of each question. Based on these analyses, the survey instrument was reduced in size, and several questions that appeared to have multiple meanings were either eliminated or edited.

The survey was administered in two formats. The first format (see Appendix A) was for use by the student participants and contained belief statements that reflected upon their personal learning experiences. The second format (see Appendix B) was for use by the flight instructors and was edited for tense and subject-verb agreement to obtain their beliefs about the effect of the experiment on their students' behavior and attitudes.

In addition to the survey instrument, two separated grading forms for use in the study were developed. These forms were used to collect the grade data from the experimental group participants. The instruments consisted of simple representations of the grading scale and procedure used by Groups B and C. Group A did not require an additional grade sheet because it used the same grade procedure and grade descriptors as the current flight department schema.

Grading Instrument 1 (see Appendix C) was used by Group B and was designed to allow the student and instructor to grade each lesson separately and, then, come together to discuss their results. This instrument was designed to test the collaborative grading schema, independent of the LCG descriptors developed by the FITS team (Connolly et al., 2005; Stefani, 1998).

Grading Instrument 2 (see Appendix D) was used by Group C. It embraced the collaborative schema as well as a modified version of the LCG descriptors. The modifications to the grade descriptors were made after consultation with the pilot test group and the formative committee, including the original authors of the descriptors. The original FITS descriptors did not match the categories present in the university flight management system, and the experienced instructors believed this would be confusing. Changes were made, and all agreed that the change captured the essence of the LCG scale in a more practical way.

Procedures

Procedure 1. This procedure consisted of an extensive literature search to identify current and emerging concepts relating to student assessment. Procedure 1 related directly to Research Questions 1 and 2. Research Question 1 asked, What does the literature suggest about the validity and reliability of traditional grading procedures in

aviation or other education programs? Research Question 2 asked, What does the literature suggest about the validity and reliability of LCG procedures in aviation education programs?

The literature search identified a variety of teacher-centered and learner-centered performance assessment and grading techniques within the broad scope of androgogical and pedagogical practice. Strengths and weaknesses in traditional teacher-centered approaches were identified. From these, the FAA-sponsored, FITS-based LCG approach was selected for evaluation in this study. This approach consisted of collaboration between learner and teacher in the grading process as well as the use of objective standards for task evaluation. The literature identified a clear difference between the grading process and the meaningfulness of grade descriptors. This study examined process and descriptors as compared to the more traditional model.

The literature search included an anecdotal analysis of current grading procedures within the flight department. Interviews were conducted with the chief pilot and the assistant chief pilot to discern their views of the efficacy of the grading process in use. Based on their comments, a review of available electronic student progress reports revealed current procedures and highlighted significant shortfalls.

Procedure 2. This procedure consisted of a literature search to identify survey formats for use in the study. Procedure 2 directly related to Research Question 3: What does the literature suggest about the appropriate method to determine assessment validity and reliability through quantitative or qualitative means? This literature search attempted to define the meaning and purpose of the concepts of validity and reliability as they related to the task of student assessment. The concept of grade validity was found to be directly related to the accuracy and utility of the grade as it related to student learning.

Grade reliability refers to the repeatability of the grade, regardless of validity, by different raters and at different times when used to grade identical or nearly identical performances. Grade validity and reliability were found to be related concepts.

Procedure 3. This procedure consisted of the actual development of the survey instrument. Procedure 3 directly related to Research Question 4: How should the perceived validity and reliability of flight student assessment programs be evaluated? A single survey designed to measure participant opinions about assessment system reliability and validity was used in the completion of the study. The survey was constructed using a Likert scale for ease of data analysis and was used to assess instructor and student preferences about specific aspects of grading as well as the entire learning experience. The survey questions were constructed to measure specific participant attitudes about the issues of validity, reliability, fairness, self-esteem, and perceived effectiveness. The survey was validated by the following procedure.

The researcher convened a formative committee to assist in the development of the study instruments. The members of this group consisted of the university chancellor, the principal investigator of the FITS research team, and the human factors representative on the team. Two face-to-face meetings were held with the formative committee prior to and during the development of the survey instrument, the subsequent pilot testing of the instrument, and the development of the experimental grading instruments. The first draft of the survey instrument was submitted to the members of the formative committee for their review and comment. The formative committee provided valuable feedback to determine if the meaning of each question remained constant from the author to the respondent (Gall et al., 2003). These edits were used to craft a draft survey.

The researcher administered the draft instrument to a sample of the flight

instructor population (Gall et al., 2003). This pilot test of the survey instrument contained a series of questions about the survey that allowed individuals surveyed to comment on each question within the survey.

The pilot survey was administered in a controlled environment and contained suitable instructions to ensure a thorough and complete data collection. Upon further advice from the formative committee and after analysis of the pilot test results, the decision was made to develop two versions of the survey. The student survey (see Appendix A) was designed to be administered to the student participants. The instructor survey (see Appendix B) was modified to ask the flight instructors the same questions, only with reference to their students' learning process. The surveys were designed to be administered as a pretest and as a posttest.

Procedure 4. This procedure consisted of development of the actual grading instruments and assessment techniques to be used by participants within the study. Procedure 4 directly related to Research Question 4: How should the perceived validity and reliability of flight student assessment programs be evaluated?

Two grading formats were used. The first format (see Appendix C) consisted of the current grade descriptor format in use by the university flight training program (see Table 1). The second format (see Appendix D) used the objective grading criteria developed by the researcher (see Table 4). Two subject matter experts and members of the formative committee assisted in the development of these instruments. The members of this group consisted of the university chancellor and the principal investigator of the FITS research team. These individuals were selected due to their knowledge of the subject matter and interest in the results.

Table 4

Experimental Learner-Centered Grading Scale

Grade	Description
Performing	At the completion of the lesson, the student will be able to perform the activity without assistance from the instructor. Errors and deviations will be identified and corrected by the student in an expeditious manner. The student meets the practical test standard.
Practicing	At the completion of the lesson, the student will be able to practice the activity with input from the instructor. The student, with coaching and assistance from the instructor, will quickly correct minor deviations and errors. The student does not meet the practical test standard.
Learning	At the completion of the lesson, the student has been recently introduced to a task or maneuver and requires significant help from the instructor to complete it. The student is making good progress toward the practicing level.
Regressing	At the completion of the task, the student and instructor agree that the student does not fully understand or needs more practice to make progress. This grade requires the student and instructor to discuss the plan for the next lessons and may require additional training.

Procedure 5. This procedure consisted of recruiting and identifying participants for the study. Students enrolled in the flight program at the university were asked to participate in the study through e-mail and word of mouth advertisement. An initial meeting was held for participants. At this meeting, student participants were assigned to their experimental groups by the use of a random number generator, and instructor participants were identified by their preassignment (by the university, rather than by the researcher) to the individual students. In most cases, the students and their instructors had been together since the semester started, 2 months prior to the start of the experiment.

Procedure 6. This procedure consisted of the directed use of the three different grading methods by the student-instructor groups, to gather data for later analysis. Procedure 6 was directly related to Research Questions 5, 6, and 7: How do the participants (instructors and students) in the study rate the validity and reliability of traditional grading techniques?, How do the participants (instructors and students) rate the validity and reliability of collaborative LCG techniques if a traditional grading scale is utilized?, and How do the participants (instructors and students) rate the validity and reliability of collaborative LCG techniques when objective performance grading standards are utilized?

Participants were divided into three groups and asked to use one of three specific grading methodologies during a preidentified phase of flight training and a specific number of repetitions. The first methodology required no change to current procedure and measured the students' reactions to the normal grade system once their attention was called to the subject. The second 2 methodologies required the participants to complete modified grade sheets (see Appendixes C and D) prior to entering their grades into the students' records.

Outcome 1. Prior to the beginning of the experiment, all participants were asked to take the appropriate version of the survey instrument (see Appendixes A and B) based on their status as a student or an instructor. These data were used to generate baseline data on their attitudes and opinions about the flight department grading system.

Hypothesis 1. This hypothesis stated that that participants in Groups A, B, or C will not express a strong preference for or against the traditional grading system during their pretest survey. The null hypothesis stated that participants will express a statistically significant preference for or against the traditional grading system during the pretest

survey. Hypothesis 1 was directly related to Research Question 5.

Outcome 2. Group A conducted five grading repetitions utilizing the standard grading methodology in use at the university. All lesson grading was conducted by the flight instructor. This was the control group. The posttest survey was administered immediately upon completion of the Group A experiment. These data were used to determine the participants' attitudes toward the traditional grading schema they used for the specified period they used it. It also provided data to identify any gains in the experimental groups through the effect of expectancy (Salvia & Ysseldyke, 2007).

Hypothesis 2. This hypothesis stated that the participants in Group A will not express a statistically significant difference between the pretest and the posttest. The null hypothesis stated that there will be a statistically significant difference between the preferences expressed by participants in Group A between the pre- and the posttest. Hypothesis 2 was directly related to Research Question 5.

Outcome 3. Group B utilized the same grading scale as Group A. However, the students and instructors were required to grade specific lesson items independently of each other and compare their grades before deciding upon the final grade. Group B also conducted five grading repetitions. The posttest survey was administered immediately upon completion of the Group B experiment. These data were used to determine the participants' attitudes toward the LCG systems they used for the specified period they used it. These data helped determine if learner involvement in the grading methodology produced a separate effect from the actual grading scale used.

Hypothesis 3. This hypothesis stated that the participants in Group B posttest will express a statistically significant positive preference over the Group B pretest results. The null hypothesis stated that there will be no significant difference the preferences

expressed by the participants in Group B posttest results and those expressed in the Group B pretest results. Hypothesis 3 was directly related to Research Question 6.

Outcome 4. Group C followed the same collaborative LCG methodology as Group B. However, the grading symbols used were those described in Table 4, modified with the advice and consent of the formative committee. Group C was administered the posttest survey immediately upon completion of five grading iterations. These data were used to determine the participants' attitudes toward the LCG systems they used for the specified period they used it. This survey was administered to all three groups. The differences identified between the pre- and the posttest will determine the expected levels of agreement.

Hypothesis 4. This hypothesis stated that the participants in Group C posttest results will express a statistically significant positive preference over the Group C pretest results. The null hypothesis stated that there will be no significant difference the preferences expressed by the participants in Group C posttest results and those expressed in the Group C pretest results. Hypothesis 4 was directly related to Research Question 7.

Procedure 7. This procedure consisted of the statistical analysis of the data. Quantitative results were evaluated using Microsoft Excel and the SPSS software. The survey results from Groups A, B, and C were compared. A paired sample, two-tailed *t* test for differences between means was used (Gall et al., 2003). Significance was achieved at the $< .05$ level.

Outcome 5. Survey responses were compared between each of the three groups to identify significant differences between the respondents. A qualitative analysis of participants written comments from the daily grade sheets and the surveys was conducted to assist in interpreting the numerical results.

Hypothesis 5. This hypothesis stated that the perceived validity and reliability of the grading system utilized by the participants in Group B as measured by the survey instrument would be significantly increased over Group A. Statistical significance will be achieved at the < .05 level. The additional hypothesis stated that the perceived validity and reliability of the grading system utilized by Group C may be significantly higher than Groups A and B. Statistical significance will be achieved at the < .05 level.

The null hypothesis stated that there will be no significant difference between the perceived validity and reliability of the grading systems utilized by Groups B and C and the system utilized by Group A. Statistical significance will be measured at the < .05 level. The null hypothesis would not establish the perceived validity and reliability of the grading system utilized by the control group, rather, it would establish the failure of the experimental groups to attain any statistical advantage in perceived validity and reliability. Hypothesis 5 was directly related to Research Questions 5, 6, and 7.

Procedure 8. This procedure was the submission of the results to the members of the formative committee as well as a summative committee. The summative committee consisted of the dean of the College of Aviation, the chairman of the aeronautical science department, and the chief pilot of the flight department. Two face-to-face meetings were held. The summative committee provided valuable feedback on the data analysis and creation of the final report.

Procedure 9. This procedure was the creation of the final report. The recommendations contained in the final report were submitted to the university for review and action.

Outcome 7. The final report will be used to support maintenance of the current grading system or adoption of one or both of the experimental grading methods.

Limitations

The data collected in this study were predictive for only the flight program in the university under study. However, other collegiate flight programs as well as stand alone flight training programs may find the information useful as they examine their assessment processes.

Chapter 4: Results

Introduction

A private aviation university made a decision to embrace a new learner centered FITS approach to flight training that included a LCG philosophy. This learner-centered approach embraced constructivist theories that entered the educational discourse in the last half of the 20th century (Knowles et al., 1998; Wright, 2002). An adult learning-centered approach places a renewed emphasis upon student involvement across the entire spectrum of the learning process, including performance assessment and evaluation (Anderson, 1998; Stefani, 1998).

The researcher of this study sought to determine the most appropriate form of lesson grading consistent with the desired university approach to flight training. The following seven research questions were investigated:

1. What does the literature suggest about the validity and reliability of traditional grading procedures in aviation or other more conventional classroom, education programs?

2. What does the literature suggest about the validity and reliability of LCG procedures in aviation or other more conventional classroom education programs?

3. What does the literature suggest about the appropriate method to determine perceived assessment validity and reliability through quantitative or qualitative means?

4. How should the perceived validity and reliability of flight student assessment programs be evaluated?

5. How do the participants (instructors and students) in the study rate the validity and reliability of traditional grading techniques? In this form of grading, the instructor assigns the student performance task grades using the traditional grading scale currently

in use in the flight training department.

6. How do the participants (instructors and students) rate the validity and reliability of LCG techniques if a traditional grading scale is utilized? In this form of grading, the students self-assign performance task grades using the traditional grading scale currently in use in the flight training department. These data help determine if learner involvement in the grading methodology produces a separate effect from the actual grading scale used.

7. How do the participants (instructors and students) rate the validity and reliability of LCG techniques when objective performance grading standards are utilized? In this form of grading, the students self-assign performance task grades using the objective performance grading developed by the FITS research team. Because the grading scale and the grading methodology were modified simultaneously, this question determined the combined effect.

A formative committee was established to assist the researcher with the development of the research questions, the design of the procedures, and the development of the survey instrument. A summative committee was established to review the results of the experiment and assist the researcher in interpreting the data as they related to each of the research questions. The researcher received assistance from the university chancellor, dean of the College of Aviation, the principle investigator of the FITS program, the university's chief pilot, and the university's assistant chief pilot to conduct the experiment. All data were generated by the participants during the course of the experiment and were collected and compiled by the researcher.

A two-part qualitative and quantitative evaluation was used to measure flight student and flight instructor preexperiment and postexperiment attitudes about three

separate and distinct grading methodologies. The first group utilized the traditional grading methodology currently in use at the university that utilizes a traditional grading scale and in which the instructor determines all grades. The second group also utilized the traditional university grading scale. However, collaboration between the instructor and the student was introduced prior to establishment of the final grade. The third group utilized grading collaboration and a version of the FITS, objective-based grading scale developed by the researcher with the assistance of a formative committee.

Research Questions 1 through 4 were answered through the literature survey process. Research Questions 5, 6, and 7 were answered by conducting the experiment, gathering the data, and analyzing the statistical results. A paired sample *t* test of means was used to determine the significance between the pre- and the posttest data. Significance was determined at the $< .05$ level.

Results of Research Question 1

The answer to this research question appeared to lie in the intended and unintended purposes of the grading system in use. Although research specifically focused on aviation assessment and grading appeared to be minimal, the broader field of educational research indicated a bias, in practice, toward the first of two competing views of education. Bloom et al. (1971) defined the practice of educational assessment (and in a broader sense the practice of education itself) as one that either selects those that will ultimately succeed from a greater pool of entering students or one that attempts to educate all students according to their needs. Traditional grading scales were found to support the concept of grades as selection instruments, rather than grading as an active part of the learning process (Bloom et al.; Kohn, 1994). The literature described this positivist approach to grading as a simple, clear, and relatively unequivocal method of

student assessment based upon the authority possessed by the teacher (Baines & Stanley, 2004). However, other voices in the literature perceived learning as a more complex interaction between the student and the teacher and questioned the accuracy and repeatability of a relatively unsophisticated teacher-centered system (Anderson, 1998; Kohn; Speck, 1998).

Student grading policies and procedures were also found to have unintended purposes or consequences. The research suggested that students express significant displeasure with the accuracy, fairness, and emotional impact of grades (Crocker et al., 2003; Davis et al., 2000; Kohn, 1994). These problems surfaced in several studies and gave rise to a form of grade negotiation in which the student attempts to negotiate a grade for whatever purpose they deemed most useful (Baines & Stanley, 2004).

The symbols used to grade students were found to be quite diverse worldwide and not subject to any visible form of standardization or scientific analysis to determine validity or reliability (Gopinath, 2004; World Educational Services, 2007). In the absence of any scientific development, tradition and prior practice emerged as the source for grading systems and symbols.

A review of the university flight department grading and assessment systems yielded some specific problems with the traditional grading systems in use. An initial review, including a limited data collection and analysis of source documents that supported the university flight training system, revealed two distinct assessment patterns that suggested another purpose for the grade system. Interviews, observations, and some anecdotal data collection (see Figures 1 and 2) revealed that grades appeared to be used as a pass-fail device to signal a need for additional training. Additionally, early in the curriculum, students appeared to receive better grades than their performances warranted.

Later in the curriculum, as FAA-required check rides approached, the grades began to reflect the students' lack of readiness for the event and requirement for additional training. This suggested that flight instructors may have attempted to adapt the traditional grading scale for a different purpose other than learning enhancement.

No research or scholarship, save tradition and experience, could be cited as a source for the university flight grading system. However, conversations and interviews with flight department personnel revealed a common understanding of the weaknesses present in the current system (I. J. Grau, personal communication, March 1, 2007; K. L. Byrnes, personal communication, February 28, 2007). Although the grade system may have had some input into the student learning process, it appeared to be associated more closely with the administration of the program (Hendrickson et al., 1999). The grade systems de-facto function could be described, most accurately, as a student management tool.

The emergent problem is the real difficulty in measuring validity and reliability when the actual purpose of the grading system remains in question. The process of grading appeared to require an understanding of human motivation, emotional satisfaction, cultural diversity, and repeatability under diverse circumstances. Additionally, accuracy in language and symbol appeared to exert some impact on the validity and reliability of a specific grade system. The literature search revealed little definitive research on the validity and reliability of the more traditional and accepted forms of grading. The fact that the current grading system appeared to be disconnected from the actual assessment process would appear to support Hypothesis 1 and 2 that stated that students and instructors will neither express a strong preference for or against the traditional grading system.

Results of Research Question 2

Researchers appeared to agree that grade validity and reliability may be a function of the quality of the feedback received by the student and its application in the learning process (Blickensderfer & Jennison, 2006; Boud & Falchikov, 1989; Butler, 2004; Kohn, 1994; Stefani, 1998). The real purpose of any grading rubric should be to provide a meaningful path for communication between student and teacher with the goal of improving performance. The learner-centered approach attempts to define the requirements of the learner although limiting those aspects of the grading process that are counterproductive. Thus, an effective approach to student grading should be informed by the emotional, motivational, and physical aspects of the learning process. The learner-centered approach appeared to provide some advantages in this area.

The recognition of the importance of construct validity in the design of a student assessment process should produce a clear and easily understood set of symbols that communicate accurate and actionable information about the students' performances (Salvia & Ysseldyke, 2007). Because information about the students understanding of the task and performance is not solely the possession of the instructor, an assessment instrument should be carefully designed to extract the most accurate and timely information from student and the instructor. The FITS, LCG system appeared to achieve this goal.

However, it appeared that the concept of reliability in LCG had not been deeply explored in the conventional or aviation literature. Alternatively, the use of rubrics as a method to increase student and teacher communication and improve validity and reliability provided some useful guidance (Choinski et al., 2003; Lunsford & Melear, 2004). Thus, the design and application of objective, descriptive, and easily

communicated performance standards appeared to be a requirement to achieve acceptable grade validity and reliability.

The literature on learner-centered and collaborative grading appeared to support Hypothesis 3 and 4 that stated that the participants in the study will express a preference for grading systems that emphasize collaboration, cooperation, communication, and clarity in the symbols used. However, actual information on actual validity and reliability of LCG was explored in Research Questions 6 and 7.

Results of Research Question 3

Gopinath (2004) identified four threats to assessment validity that inform the construction of an assessment rubric. First, the meaning of the symbols used to define the specific grades can be interpreted differently by student and teacher. Second, the emotional and psychological impact of the grade can affect its award. Third, relative grading can compare students to other students, rather than to a standard. Finally, instructors can miss important information that might affect a specific grade due to the simple act of omission. Thus, validity in grading should be judged based on the existence of clear and understandable standards that accurately describe student performance in a language that is clearly understood by student and teacher. The research also identified those practices that increase grade validity.

Butler (2004) examined the role of feedback in a comments-only grading experiment. He found the accuracy and volume of communication appeared to increase grade validity. Schaffner et al. (2000) noted that grade validity should add value to the learning process and that the inclusion in the learning process adds validity to the grade. This relationship can be identified by the presence of connection, meaningfulness, and fairness in the grading system. Thus, grade validity appeared to be enhanced by active

communication and association with, rather than isolation from, the learning process.

Any attempt to measure grade validity should also test for the presence of concurrent and predictive validity (Messick, 1989). Does the grade schema give the student and the instructor a clear picture of the student's current performance while accurately predicting future performance? Thus, an inquiry into the validity of a grading system was considered to be more effective if it occurred over time and examined the short- and long-term accuracy of the grade.

Other researchers addressed the content of the actual survey instrument. Several previous research efforts identified in the literature linked the concepts of fairness, motivation, student involvement, and emotional impact to the validity and reliability of the grade (Schaffner et al., 2000; Wooley & Wooley, 1999). All appeared to contribute to the complex relationship between performance and evaluation. An instrument that explores this relationship should be designed to identify and measure their presence or absence. Wooley and Wooley identified the utility of grouping questions into functional categories. This approach appeared to yield more reliable data and was utilized by the researcher. The construct validity of a grading system can be measured by examining its clarity and accuracy in application over time. Additionally, the concepts of fairness, emotional impact, and relativism should be examined to determine their presence and impact. The research revealed that an assessment instrument designed to assess grade validity should reflect all of these concepts.

The related concept of grade reliability was viewed as validity measured over time (Gall et al., 2003). The longitudinal accuracy of grade systems is dependent on the ability of multiple graders to execute grade systems reliably. The characteristics one might expect of a reliable grading system were found to be clarity, location, and setting

stability. Rater objectivity and stable system design were also noted as important factors (Feldt & Brennan (1989).

Thus, researchers developing reliable grading systems should create clear, descriptive, and well thought-out symbols that describe student performance in easily understood terms. The literature on LCG referred to these as objective standards. The method of evaluation should also allow for a familiar and stress-free environment, and the assessment system should be designed to be self-explanatory. The more individual training is required to counteract complexity, the more chance for subjective standards to be introduced through the training process.

LCG takes on the subject of rater subjectivity by introducing a constant to the rating scenario, the student. So it was useful to know who had the more powerful voice in the assessment process, the student or the instructor or did both share the responsibility and power equally? Finally, it appeared to be difficult to measure grade validity and reliability in any single event, rather, an examination of the subject required some level of repetition of the grading exercise to let the participants learn how to use a new system and to reflect upon its effect on the grading process.

Results of Research Question 4

Qualitative and quantitative measures were utilized with success to identify the presence of grade validity and reliability. Because the grading process requires a subjective decision on the part of the rater, even if an objective standard is present, the opinion of the rater has significant value.

In the case of collaborative-grading techniques, student and teacher are asked to rate the students' performances. However, this may call into to question the accuracy and integrity of self-assessment. Shim et al. (2001) established the correlation between

student and teacher self-reporting of events that suggested the following. Student and teacher collaborative grading may be effective. Therefore, a single instrument developed and administered to the student and the teacher to measure their perceptions of a grading system may achieve useful results. The formative committee members reviewed the instrument prepared for this experiment and concurred. However, they asked that two versions of the instrument be created to express the different grammatical points of view possessed by student and instructor.

Thus, one instrument tailored to the different points of view and administered to student and teacher should yield significantly useful insights into the grading process. Pilot testing with university flight instructors of the proposed instrument supported the concept of grouping and identified questions and groups of questions that were redundant, irrelevant, or hard to understand. The pilot test group's written and oral feedback was concise and effective.

A review of the literature, the formative committee, and the pilot test group results supported the concept that an effective research design should include qualitative inquiry and quantitative methodology. It should measure a broad range of independent and dependant variables that affect the grading process, and it should measure the effect of change over time through a series of experiment repetitions as students and instructors are introduced to new processes and methodologies.

Results of Research Question 5

In this form of grading, the instructor assigns the student performance task grades using the traditional grading scale currently in use in the flight training department. Combined data for students and instructors as well as data for the subset of student participants are presented.

Combined student and instructor results. Thirty-four flight students and flight instructors (36 completed the pretest, and 2 were unable to complete the entire experiment) were administered a 30-question pretest designed to measure the validity and reliability of the particular grading system in use during this portion of the experiment. Upon completion of the pretest, each student and instructor team in this group completed five flight training device (FTD) lessons using the traditional grading scale. Immediately following the completion of these training periods, the participants were administered a posttest consisting of the same 30 questions administered in the pretest. The resulting data were manipulated through the use of a paired sample, two-tailed *t* test for significance.

The combined student and instructor pretest mean was 3.3876 (on a 5-point scale) as compared to a posttest mean of 3.3581 for a negative variance of 0.0295. In the student-only group, the pretest mean was 3.4429, and the posttest mean was 3.3407 for a negative variance of 0.1022. The means of the responses to Questions 6 and Question 7 reflected disagreement between instructors and students and appeared to account for the > .05 significance score in the combined student and instructor results. When student survey results were examined without the instructors, the disagreement disappeared, and significance was achieved at the < .05 level. This result appeared to support Hypothesis 1.

The survey was composed of positive statements designed to detect the presence or absence of grade validity and reliability. When only those questions were considered that made positive statements about grade validity and reliability, the results were as follows. The mean of the scores on the combined student and instructor group pretest was 3.4865 (on a 5-point scale) as compared to a mean of 3.4303 on the posttest for a

negative variance of 0.0562. In the student-only group, the pretest mean was 3.5238, and the posttest score was 3.3980 for a negative variance of 0.1258. This result appeared to support Hypothesis 1.

The mean scores of each question were determined and tested for significance using a paired sample, two-tailed *t* test. The mean scores of 11 of 30 questions on the posttest increased as compared to the pretest with the score for Question 7 achieving significance at the < .05 level. The mean scores of 19 of the 30 questions decreased from the pre- to the posttest with three of these decreases (Questions 6, 18, and 23) achieving significance at the < .05 level of significance and Question 19 approaching the < .05 level of significance.

These results appeared to provide support for the null hypothesis (see Table 5). The relatively strong, negative posttest results on Questions 18 and 23 (positive statements about grade system accuracy and grader consistency) after five repetitions of the traditional grading scale posed some specific questions for grade validity and reliability. However, the combined results for Group A appeared to support Hypothesis 2.

Student-only results. The following text describes a subset composed of the student participants in the experimental group. Twenty-four flight students (25 completed the pretest, and 1 was unable to complete the experiment) were administered a 30-question pretest designed to measure the validity and reliability of the particular grading system in use during this portion of the experiment. The procedures used were identical to those described for the instructors and student combined group. These results are described in Table 6. The mean of the scores in the twenty four pretests was 3.4073 (on a 5-point scale) as compared to a mean of 3.3537 on the posttest. The mean scores of each question were determined and tested for significance using a paired sample, two- tailed *t*

test.

Table 5

Combined Traditional Grading Individual Questions (N = 34)

Question	Pretest *M*	Posttest *M*
I believe my instructor is more critical of my performance than I am	2.8235	*2.4706
I believe I am more critical of my own performance than my instructor is	3.5588	*3.9412
I believe the grades I received were accurate	4.0000	*3.7059
I believe my instructor grades me consistently from lesson to lesson	3.9706	*3.6765

Note. Responses were made on a 5-point scale paired sample, two-tailed *t* test for significance (1 = *completely disagree* and 5 = *completely agree*); *p < .05.

The mean scores of 8 of 30 questions on the posttest increased as compared to the pretest. None of these score increases achieved the < .05 level of significance. Four questions neither increased nor decreased between the pre- and posttest. The mean scores of 18 of 30 questions decreased from the pre- to the posttest with 2 of these decreases (Questions 18 and 23) achieving significance at the < .05 level of significance and 3 questions (Question 6, 19, and 26) approaching the < .05 level of significance.

As in the Group A student and instructor results, the student-only results appeared to support Hypothesis 2. However, the student-only responses suggested some support for the null hypothesis. The relatively strong negative posttest results on Questions 18 and 23 that dealt directly with grade system accuracy and grader consistency after five repetitions of the traditional grading scale posed interesting questions for grade validity

and reliability.

Table 6

Student-Only Traditional Grading (N = 24)

Question	Pretest *M*	Posttest *M*
I believe the grades I received were accurate	3.9583	*3.5833
I believe my instructor grades me consistently from lesson to lesson	4.0417	*3.5833

Note. Responses were made on a 5-point scale paired sample, two-tailed *t* test for significance (1 = *completely disagree* and 5 = *completely agree*); *p < .05.

Results of Research Question 6

In this form of grading, the student self-assigned performance task grades using the traditional grading scale currently in use in the flight training department. These data helped determine if learner involvement in the grading methodology produced a separate effect from the actual grading scale used. Combined data for students and instructors as well as data for the subset of student participants are presented. Instructor only data are not presented due to the very low number for instructor participants.

Combined student and instructor results. Twenty-eight flight students and flight instructors (33 participants completed the pretest, and 5 were unable to complete the experiment) were administered a 32-question pretest designed to measure the validity and reliability of the particular grading system in use during this portion of the experiment. The remaining questions asked specific questions about the collaboration that took place during the experiment.

Upon completion of the pretest, each student and instructor team in this group

completed five FTD lessons using the traditional grading scale currently in use at the university. Additionally, this group of participants (instructors and students) collaborated on the final lesson grade by independently arriving at a proposed lesson grade and, then, discussing their individual grades prior to entry into the university training administration system. Immediately following the completion of these training periods, the participants were administered a posttest consisting of the same 30 questions administered in the pretest. The resulting data were manipulated through the use of a paired sample, two-tailed *t* test for significance.

The mean of the scores of the combined student and instructor pretest group was 3.4200 (on a 5-point scale) as compared to a mean of 3.6753 on the posttest for a positive variance of 0.2535. In the student-only group, the pretest score was 3.4498, and the posttest score was 3.6803 for a positive variance of 0.2305. The results represented a statistically significant increase in the mean among students and instructors who collaborated during the grading process. These data, when compared to the Group A data as well as the Group B presurvey, suggested a positive outcome for grade collaboration.

The survey was composed of positive statements designed to detect the presence or absence of grade validity and reliability. When only the questions were considered that made positive statements about grade validity and reliability, the results were as follows. The mean of the scores of the combined student and instructor group was 3.3541 (on a 5-point scale) as compared to a mean of 3.9285 on the posttest for a positive variance of 0.5944. In the student-only group the pretest score was 3.5919, and the posttest score was 3.9844 for a positive variance of 0.3925.

The mean scores of each individual question were determined and tested for significance using a paired sample, two-tailed *t* test. The mean scores of 27 of 32 paired

questions on the posttest increased as compared to the pretest. Nine of these score increases achieved < .05 level of significance (Questions 4, 5, 11, 12, 13, 17, 26, 36, and 37). One score increase (Question 30) approached the < .05 level of significance. The mean scores of 5 of 32 questions decreased from the pre- to the posttest. None of these decreases achieved significance at the < .05 level of significance. These data are depicted in Table 7.

A question-by-question analysis of these data revealed the following. Questions 5, 11, 12, and 13 were positive statements that spoke directly to student motivation as a product of the Group B grading system. Questions 4 and 26 were positive statements about increased instructor student feedback as a product of the Group B grading system. Questions 17, 36, and 37 were positive statements that asserted that the Group B grading system was fair and accurate and provided enough grading options.

Student-only results. The following text describes a subset composed of only the student participants in the experimental group. Nineteen flight students (23 completed the pretest, and 4 were unable to complete the experiment) were administered a 32-question pre- and postsurvey designed to measure the validity and reliability of the particular grading system in use during this portion of the experiment. The procedures used were identical to those described for the instructors and student combined group.

The mean of the scores in the 19 preexperiment surveys was 3.4492 (on a 5-point scale) as compared to a mean of 3.6799 on the posttest. The mean scores of each question were determined and tested for significance using a paired sample, two-tailed, *t* test. The mean scores of 24 of 32 questions on the posttest increased as compared to the pretest with Questions 5, 11, 12, 13, and 30 achieving significance at the < .05 level of significance and Questions 17 and 36 approaching the < .05 level of significance. The

mean scores of 8 of 32 questions decreased from the pre- to the posttest with none achieving the < .05 level of significance.

Table 7

Combined Collaborative Grading Individual Questions (N = 28)

Question	Pretest *M*	Posttest *M*
I believe the grade process provides feedback to help improve my performance	3.7143	*4.4286
I believe the grade process motivates me to improve my work	3.8571	**4.4643
I believe the grading system I used motivated me to work harder	3.3929	**3.8929
I believe the grading system I used made me feel more positive about my FTD lessons	3.0000	**3.6429
I believe the grading system I used motivated me to work harder when I received a low grade	3.3571	**4.2857
I believe the grades I received were fair	3.9643	*4.1176
I believe the way the lesson was graded improved the amount of feedback I get from my instructor	3.6429	*4.0714
I believe the grading scale (the actual grade) we used gives the grader an accurate way to describe student performance	3.0741	*3.7037
I believe the grading scale (the actual grade) we used gives the grader enough options to describe student performance	2.8519	*3.5556

Note. FTD = flight training device; responses were made on a 5-point scale paired sample, two-tailed *t* test for significance (1 = *completely disagree* and 5 = *completely agree*); *p < .05 and ** p < .01.

A question-by-question analysis of these data revealed the following. Questions 5, 11, 12, and 13 were positive statements that spoke directly to student motivation as a product of the Group B grading system. Question 30 was a positive statement about the

importance of all grades to the student. These data are depicted in Table 8.

Table 8

Student-Only Collaborative Grading Individual Questions (N = 19)

Question	Pretest *M*	Posttest *M*
I believe the grade process motivates me to improve my work	3.9474	**4.5789
I believe the grading system I used motivated me to work harder	3.4211	*3.8421
I believe the grading system I used made me feel more positive about my FTD lessons	3.1053	**3.7368
I believe the grading system I used motivated me to work harder when I received a low grade	3.5263	**4.3158
I believe all grades are important to me	3.7895	*4.3684

Note. FTD = flight training device; responses were made on a 5-point scale paired sample, two-tailed *t* test for significance (1 = *completely disagree* and 5 = *completely agree*); *p < .05 and ** p < .01.

The combined Group B data as well as the number of individual questions responses that showed significant increases appeared to support Hypothesis 3 that stated that the Group B postexperiment survey responses will express a significantly positive preference over the Group B preexperiment responses.

Results of Research Question 7

In this form of grading, the students self-assigned performance task grades using the objective performance grading developed by the FITS research team. Because the grading scale and the grading methodology were modified simultaneously, this question determined the combined effect. Combined data for students and instructors as well as data for the subset of student participants are presented. Instructor-only data are not

presented due to the very low number of instructor participants.

Combined student and instructor results. Thirty-four flight students and flight instructors (38 participants completed the pretest, and 4 were unable to complete the experiment) were administered a 32-question pretest designed to measure the validity and reliability of the particular grading system in use during this portion of the experiment. A total of 38 posttest questions were administered. However, only 32 were paired with pretest questions. The remainder asked specific questions about the collaboration that took place during the experiment. Upon completion of the pretest, each student and instructor team in this group completed five FTD lessons using the experimental FITS grading scale developed by the researcher. This group of participants also collaborated (as did Group B) on the final lesson grade by independently arriving at a proposed lesson grade and then discussing their individual grades prior to entry into the university training administration system. Immediately following the completion of these training periods, the participants were administered a posttest consisting of the same 30 questions administered in the pretest.

The mean of the scores on the combined student and instructor group was 3.3030 (on a 5-point scale) as compared to a mean of 3.6337 on the posttest for a positive variance of 0.3307. In the student-only group, the pretest score was 3.3659, and the posttest score was 3.6412 for a positive variance of 0.2753. The survey was composed of positive statements of belief that were designed to detect the presence or absence of grade validity and reliability.

When only the questions were considered that made positive statements about grade validity and reliability, the results were as follows. The mean of the scores on the combined student and instructor group was 3.3457 (on a 5-point scale) as compared to a

mean of 3.9271 on the posttest for a positive variance of 0.5814. In the student-only group, the pretest score was 3.3940, and the posttest score was 3.9442 for a positive variance of 0.5502.

The mean scores of each question were determined and tested for significance using a paired sample, two-tailed *t* test. The mean scores of 21 of 32 paired questions on the posttest increased as compared to the pretest. Fifteen of these score increases achieved a < .05 level of significance (Questions 4, 5, 11, 12, 14, 17, 19, 22, 24, 25, 26, 27, 30, 36, and 37). The mean scores of 11 of 32 questions decreased from the pre- to the posttest. The score increase for Question 21 approached the < .05 level of significance. None of the score decreases achieved the < .05 level of significance. The data are depicted in Table 9.

A question-by-question analysis of these data revealed the following. Questions 17, 19, 22, 36, and 37 were positive statements that supported grade validity as a product of the Group C grading system. Question 5, 11, 12, and 14 were positive statements that spoke directly to student motivation as a product of the Group C grading system.

Questions 24 and 25 were positive statements that supported grade reliability as a product of the Group C grading system. Questions 4 and 26 were positive statements about increased instructor student feedback as a product of the Group B grading system. Question 27 was a positive statement that the grading process improved the post-FTD briefing, and Question 30 stated that all grades were important.

Student-only results. The following text describes a subset composed of only the student participants in the experimental group. Twenty-one flight students (24 completed the pretest, and 3 were unable to complete the experiment) were administered a 32-question pretest designed to measure the validity and reliability of the particular grading

system in use during this portion of the experiment.

Table 9

Combined Learner-Centered Grading Individual Questions (N = 34)

Question	Pretest *M*	Posttest *M*
I believe the grade process provides feedback to help improve my performance	3.6765	**4.4706
I believe the grade process motivates me to improve my work	3.4118	**4.2941
I believe the grading system I used motivated me to work harder	3.2941	**3.9412
I believe the grading system I used made me feel more positive about my FTD lessons	2.8824	**3.8529
I believe the grading system I used motivated me to work harder when I received a high grade	2.8235	*3.2353
I believe the grades I received were fair	3.9118	*4.2941
I believe the grades I received were descriptive of my performance	3.3824	**4.0882
I believe the grades I received were consistent with my performance	3.7647	*4.1471
I believe different instructors grade me the same way	2.1765	*2.6765
I believe the grading process we used will help instructors grade all students more consistently	3.0000	**3.8824
I believe the way the lesson was graded improved the amount of feedback I get from my instructor	3.3235	*3.9412
I believe the grading process we used had a positive impact on the lesson post-FTD debriefing	3.4412	**4.0588
I believe all grade are important to me	3.7941	**4.2353
I believe the grading scale (the actual grade) we used gives the grader an accurate way to describe student performance	2.7353	**3.8529
I believe the grading scale (the actual grade) we used gives the grader enough options to describe student performance	2.6471	**3.7059

Note. FTD = flight training device; responses were made on a 5-point scale paired sample, two-tailed *t* test for significance (1 = *completely disagree* and 5 = *completely agree*); *p < .05 and ** p < .01.

The procedures used were identical to those described for the instructor and student combined group. The mean of the scores in the 21 pretests was 3.3348 (on a 5-point scale) as compared to a mean of 3.6725 on the posttest. The mean scores of each question were determined and tested for significance using a paired sample, two-tailed *t* test. The mean scores of 23 of 32 questions on the posttest increased as compared to the pretest with Questions 4, 5, 11, 12, 17, 19, 24, 25, 30, 36, and 37 achieving significance at the < .05 level of significance and Questions 14 and 22 approaching the < .05 level of significance. Two questions neither increased nor decreased between the pre- and posttest. The mean scores of 7 of 32 questions decreased from the pre- to the posttest with none achieving a < .05 level of significance. The data are depicted in Table 10.

A question-by-question analysis of these data revealed the following. Questions 17, 19, 36, and 37 were positive statements that supported grade validity as a product of the Group C grading system. Question 5, 11, and 12 were positive statements that spoke directly to student motivation as a product of the Group C grading system. Questions 24 and 25 were positive statements that supported grade reliability as a product of the Group C grading system. Question 4 was a positive statement about increased instructor-student feedback as a product of the Group B grading system. Question 30 stated that all grades were important. These data appeared to support Hypothesis 4 that stated that the Group C posttest results will express a significantly positive preference over the Group C pretest results. The individual question results provided even stronger support for Hypothesis 4.

Groups A, B, and C compared. When the all the data were compared, the differences of means between Groups A, B, and C appeared to support Hypothesis 5. Additionally, the significant increase in the number of individual questions that reflected improved grade validity and reliability (1 question in Group A, 9 questions in Group B,

and 15 questions in Group C) appeared to support Hypothesis 5. The number of individual questions that increased from *no opinion* to the *agree* column were noteworthy.

Table 10

Student-Only Learner-Centered Grading Individual Questions (N = 21)

Question	Pretest *M*	Posttest *M*
I believe the grade process provides feedback to help improve my performance	3.6667	**4.6190
I believe the grade process motivates me to improve my work	3.6667	**4.4762
I believe the grading system I used motivated me to work harder	3.3333	*3.9524
I believe the grading system I used made me feel more positive about my FTD lessons	2.9048	**3.9524
I believe the grades I received were fair	3.7619	*4.1905
I believe the grades I received were descriptive of my performance	3.2381	*4.0476
I believe different instructors grade me the same way	2.4286	*3.0952
I believe the grading process we used will help instructors grade all students more consistently	3.0000	**4.1429
I believe all grade are important to me	4.0000	*4.3810
I believe the grading scale (the actual grade) we used gives the grader an accurate way to describe student performance	2.9524	*3.8095
I believe the grading scale (the actual grade) we used gives the grader enough options to describe student performance	2.8095	**3.8095

Note. FTD = flight training device; responses were made on a 5-point scale paired sample, two-tailed *t* test for significance (1 = *completely disagree* and 5 = *completely agree*); *p < .05 and ** p < .01.

In Group A, a single response exceeded the agree criterion on the pre or posttest (4 out of 5 on the scale). Although in Group B, 5 of 9 responses exceeded this threshold, and, in Group C, 7 of 15 significant responses exceeded the criterion for *agree*. This support extended to the primary hypothesis that Group B posttest results will show a statistically significant increase over Group A. These results appeared to lend additional support to Hypothesis 5.

If only the questions that made positive statements about grade validity and reliability, the increase in means score was more dramatic (see Table 11). The increased variance between Group B and Group C was approximately 50% greater than the previous increase between Group A and Group B when the variance in the pretest mean was factored in.

Because Group B and Group C utilized student-instructor collaboration, much of the increase in Group C was attributed to the addition of the objective LCG scale. The combined effect of the complete LCG system appeared to be, in the opinion of the participants, the most valid and reliable of the three grading system tested. These data lent additional support to Hypothesis 5 that the Group C results would significantly increased over Group B. These data are depicted in Table 12.

These data (the data presented in support of Research Questions 5, 6, and 7) appeared to support Hypothesis 5 that participants in Group B will express a preference over Group A. The additional hypothesis that Group C will express a preference over Group B appeared to be supported as well.

Other Results

Each survey instrument contained two spaces in which students and instructors could write comments. All of the written comments were compiled, reviewed, and

evaluated. Comments were judged to be negative if they contained statements that questioned the validity and reliability of the grading system used by the particular group. Comments were judged to be positive if they contained statements that expressed satisfaction with the validity and reliability of the grading system used by the particular group.

Table 11

Group A, B, and C--Validity and Reliability Questions Only

Group	Pretest *M*	Posttest *M*	Variance
Combined student and instructor score			
A	3.4865	3.4303	-0.0562
B	3.5341	**3.9285	+0.3944
C	3.3457	**3.9271	+0.5814
Student-only score			
A	3.5238	*3.3980	-0.1258
B	3.5919	**3.9844	+0.3925
C	3.3940	**3.9442	+0.5502

Note. responses were made on a 5-point scale paired sample, two-tailed *t* test for significance (1 = *completely disagree* and 5 = *completely agree*); *p < .05 and ** p < .01.

Additionally, the number of comments of all types was compared as an anecdotal method to gauge the enthusiasm of participants about their particular grading system. The total number of pretest comments was compared to gauge the relative presurvey level of

agreement between the groups. Total number of postsurvey comments was compiled as an informal method of gauging the enthusiasm of the participants.

Table 12

Group A, B, and C--All Survey Questions

Group	Pretest *M*	Posttest *M*	Variance
Combined student and instructor score			
A	3.3876	3.3581	-0.0295
B	3.4200	**3.6753	+0.2553
C	3.3030	**3.6337	+0.3307
Student-only score			
A	3.4429	**3.3407	-0.1022
B	3.4498	**3.6803	+0.2305
C	3.3659	**3.6412	+0.2753

Note. responses were made on a 5-point scale paired sample, two-tailed *t* test for significance (1 = *completely disagree* and 5 = *completely agree*); ** $p < .01$.

The results, although not meant to be empirical, were nonetheless interesting. Although preexperiment survey comments between the three groups were uniformly negative and equally distributed, the total number of comments and the total number of positive comments increased rather steeply from the Group A postexperiment survey to the Group B and C postexperiment surveys.

Group A comments were generally focused on the lack of grade reliability

between different instructors and the lack of written comments and feedback inherent in the traditional system. Of note, one Group A instructor used the postsurvey comments to say that the use of the unsatisfactory grade during the presolo flight phase (a required grade if the student is unready to fly alone) was very de-motivating to the student.

Group B comments were mixed with seven participants making positive statements about the ability to collaborate with the instructor on lesson grading. However, an equal number of participants made negative comments on the postexperiment survey. These comments complained about the lack of use of certain grades (outstanding and marginal) and the overuse of the good and unsatisfactory grades.

Other comments spoke of the vague nature of the grades. Most of these comments were focused on the actual grading scale used, rather than on the collaborative technique used to arrive at the specific grade.

The Group C postexperiment comments were nearly all positive, doubling the Group B comments, and spoke of the validity; reliability; and, especially, the motivational aspects of the Group C grading system. Two of three negative comments were from a single instructor student pair.

The instructor did not understand or like the system and continued to dominate the grading discussion. The student noted this and made a negative comment about the instructor’s resistance to the experiment.

However, later in this comment, the student noted that he thought the new system would improve the grading process (this additional comment was not included in the positive comment tally). Although not empirical by any measure, these comments appeared to lend some anecdotal support to the hypotheses of the experiment. The data are depicted in Table 13.

Table 13

Group A, B, and C Anecdotal Written Survey Comments

Group	Presurvey negative	Postsurvey negative	Presurvey positive	Postsurvey positive
A	8	6	0	1
B	11	7	0	7
C	9	3	1	16

Note. Data presented are anecdotal and should not be considered statistically significant.

Chapter 5: Discussion

Overview of the Applied Dissertation

The purpose of this study was to evaluate student perception of the validity and reliability of the flight training lesson grading system in use at a major aeronautical university flight program. This research provided an increased understanding of the current assessment system in use as well as its effect upon the flight training program and student success. The study compared the current assessment system to a new form of flight training assessment that is soon to be adopted by the university. Students and their instructors were asked to evaluate three distinct assessment approaches to determine which system they perceived to be more valid and reliable. Sixty-four students of a population of approximately 250 and 22 flight instructors of a population of approximately 70 participated in the experiment. The results of the pre- and postexperiment survey formed the basis for this analysis.

The study asked seven research questions, and the data that supported the responses to these research questions are contained in the results. Research Questions 1 to 4 described the development of a methodology to evaluate validity and reliability as they related to traditional and LCG systems. The last three research questions utilized this methodology to examine three distinct grading systems and determine the perceived validity and reliability of each.

Relationship of Findings to the Literature

Research Questions 1 to 4. These research questions were composed of a review of the literature in an attempt to understand what it suggested about the validity and reliability of traditional and LCG procedures. The literature also provided insight into how best to develop the criteria and methodology to measure validity and reliability in

the practice of aviation education. The results suggested that the validity and reliability was a complex concept subject to multiple criteria.

Grade validity was identified by the presence of fairness, accuracy, clarity, and communication (Butler, 2004; Messick, 1989; Schaeffner et al., 2000). Collaboration and feedback between instructor and student were also identified by many researchers as strong contributors to grade validity as well as grade reliability (Blickensderfer & Jennison, 2005; Boud & Falchikov, 1989; Butler; Kohn, 1994; Stefani, 1998). Grade reliability appeared to be associated with the presence of clear and descriptive grade symbology, stable system design, and rater (and interrater) reliability and objectivity (Feldt & Brennan 1989). The presence of clearly definable standards and a grade system that took into account the emotional and motivational aspect of the grading process appeared to support the validity and reliability of grades (Davis et al., 2000; Schaeffner et al.). However, one would be wrong to assume that grade validity and reliability were isolated concepts. The symbiotic relationship between the two was present throughout the literature. The most accurate description of grade reliability appeared to be grade validity measured over time and among raters.

In addition to identifying the appropriate criteria, the research identified threats to grade reliability and validity. Failure to understand and apply grade criteria across the curriculum and a lack of basic understanding of the meaning of grades between students and teachers were noted by many as threats (Gopinath, 2004; Holmes & Smith, 2003). The greatest threat appeared to be the failure to understand that the grade is a valuable part of the learning process (Boud & Falchikov, 1989; Kohn, 1994; Stefani, 1998). The learner-centered approach favored by the FITS research appeared to understand fully this weakness and addressed it in the LCG methodology examined in this study. These data

were used to assist in the creation of a survey instrument and the study methodology.

With the assistance of a formative committee and a pilot test group, a pretest-posttest design containing 38 questions (30 were administered to Group A, and 38 were administered to Group B and Group C) were developed to administer to the participants (Gall et al., 2003). This design appeared to control effectively for internal validity problems and provided statistical data to indicate change in the participants' perceptions and opinions. Additionally, two grading instruments were developed to support the experiment, one based on a traditional grading system and one based on the FITS methodology soon to be adopted by the university (Connolly et al., 2005). The results of the pre and posttest survey appeared to support the validity and reliability of the survey instrument, and meaningful results were obtained. The survey instrument that was the primary product of Research Questions 1 to 4 appeared to be quite useful in other settings, outside of aviation, to measure grade validity and reliability.

The summative committee provided valuable insight into the survey instruments as well as the grading instruments. For example, the summative committee noted that the grade descriptor *regressing*, developed to indicate substandard performance, did not agree with the written description provided. Specifically, the grade description noted the students' failure to make progress, whereas the grade symbol described the students' performances in terms of a loss of already learned skills and knowledge. This literal disagreement should be addressed prior to the deployment of the proposed modification to the LCG grading system

Elaboration and Interpretation of Results

Research Question 5. Participants were given the opportunity to rate the traditional grading scale and procedures twice: prior to the beginning of the study and

after five repetitions of the grade procedure. Their responses on the pretest expressed a neutral perception of the grading system in use. However, all groups responded very positively prior to any discussion of collaboration or LCG to the concept of collaboration between the student and the instructor. This desire for increased communication and collaboration appeared to explain the results achieved by the experimental groups later in the experiment. The participants also believed, to a much lesser extent, that the traditional grading scale was fair, accurate, and consistent. In retrospect, this pretest result may have been more a commentary on the professionalism of the flight instructors, rather than an endorsement of the traditional grade descriptors.

The posttest, conducted after five repetitions of the same traditional grading scale, showed similar strong endorsement of collaboration between student and instructor. However, significantly, the results on the posttest were driven by the students' and instructors' negative impressions of the validity and reliability of the grade system. It appeared that, upon reflection, the participants decided that the traditional grade system was not as accurate and consistent as previously thought. Additionally, during the posttest, students and instructors differed sharply over who was more critical of the student's performance, the instructor or the student. Although some information about the experimental groups may have been discovered by Group A, it appeared more likely that the dialog between student and instructor, created by the experiment, caused the participants to change their responses.

Participant comments were also significant. Students and instructors in Group A continued to express negative opinions about the traditional grading scale on the pre- and posttest. Having nothing better or worse to compare it to, this cohort still found a variety of problems with regard to student motivation, grade accuracy, and grade fairness that did

not show up in the survey. These comments, although anecdotal, were representative of a general displeasure with traditional grading systems present in the literature.

Statistically, the Group A participants expressed little preference for or against the traditional grading scale. Additionally, the Group A pre- and postmean scores were consistent with the preexperiment scores of Groups B and C, lending credence to the results of the two experimental groups. These results fully supported Hypotheses 1 and 2.

However, the most interesting outcome of the preexperiment survey was the strong preference of participants in all groups for collaboration between student and instructor. In each group, this response approached 4.5 on a scale of 5.0 and remained the highest scoring response in each group on the pre- and posttests. This desire for collaboration and cooperation appeared to endorse the LCG approach and lend at least anecdotal support to Hypotheses 3 and 4. Further research supported this initial assumption.

Research Question 6. The literature review created an expectation that the introduction of collaboration would create an increase in participant satisfaction with the grade process. This experiment was carried out by Group B. In this group, student and instructor were able to grade the lesson independently from each other and, then, compare their results prior to the final grading. As hypothesized, this collaboration produced statistically significant increases in the overall experiment mean for Group B and, specifically, for 9 of 32 questions asked pre- and postexperiment to Group B.

Analysis of these questions revealed that each supported the idea that student-instructor collaboration significantly increased the participants' perception of student motivation, instructor feedback, grade accuracy, and fairness. Participant-written comments also supported these concepts. As the literature revealed, learning in

partnership, especially among adult students, resulted in a perception of increased motivation, shared control of the learning process, and learning effectiveness. This result supported research Hypothesis 3 and suggested that increased collaboration between student and instructor improved the grading process and may have improved the learning process as well. However, the participants in this group did not score as highly as Group C who also used collaboration and introduced objective learner-centered grade symbols.

One plausible explanation is that the presence of the traditional grading scale made it easier for instructors and students to fall back on previous habits as the experiment progressed. The five iterations of the grading experiment were designed to let the participants learn the new system and, then, personalize it as a habit. Written comments showed that this personalization took a variety of forms that may explain the greater variety of positive and negative results for Group B when compared to Group C. A second and more likely explanation is that the grade descriptors still did not provide enough valid options for the collaborative graders.

Research Question 7. Group C introduced all the elements of LCG collaboration as well as learner-centered objective grading criteria. The participants were asked to grade independently, then, collaborate on the final result. Additionally, the experiment utilized an objective grading scale that described student progress in terms of student performance against a known performance standard. Just as in the previous groups, participants were asked to accomplish five repetitions of the grading in conjunction with their FTD lesson. This allowed student and instructor to learn and become comfortable with the new system and standards.

The survey results as well as the anecdotal written comments supported the literature's assertion that LCG improved the participants' perception of the validity and

reliability of lesson grading. As expected, this produced statistically significant increases in the overall experiment mean for Group C and for 15 of 32 questions asked pre- and postexperiment. Analysis of these questions revealed that they supported the concept that collaboration significantly increased the participants' perceptions of grade validity (fairness, accuracy, and consistency), student motivation, grade and grader reliability, and the value of the debriefing process. This result fully supported Hypothesis 4 and suggested that increased collaboration between student and instructor improved the grading process and may have improved the learning process as well. The improvement of the Group C results over the Group B results also supported Hypothesis 5 that stated that Group C participants will find the results of the experiment statistically more significant than the participants in Group B.

The results of this study suggested that a learner-centered approach to aviation training assessment, in this case, an adaptation of the FITS methodology, provided a welcome and positive change to how flight training grading is conducted (Connolly et al., 2005; Knowles et al., 1998). The most striking result was the positive response to collaboration, communication, and feedback in the learning process (Boud & Falchikov, 1989). Before the experiment began (as noted on the presurvey), the students and instructors expressed a desire for collaboration that was apparently not present in the current system. Although this experiment was focused on assessment, one could make the assumption that the desire for collaborative learning was not limited to this narrow area. Once given the chance to participate and collaborate, the participants showed an immediate affinity for the process and appeared to have embraced it enthusiastically. More to the point, the increased level and accuracy of the communication, especially in Group C who had access to more precise and well-written grade symbols, appeared to

increase the participants' perceptions of fairness, accuracy, repeatability, and reliability. The presence of a perceived increase in assessment validity and reliability might not necessarily translate into an increase in actual student performance. However, the literature and the survey results suggested that the increase in student motivation and communication, due to improvements in the learning system, should have a significant positive impact on student performance (Blickensderfer & Jensen, 2006; Crocker et al., 2003).

Discussion of Conclusions

The purpose of this study was to evaluate the students' perceptions of the validity and reliability of LCG as it is applied to a university flight training environment. The study concluded that the insertion of formalized collaboration between instructor and student and the addition of objective LCG criteria had a significant effect upon the students' and flight instructor's perceptions of grade validity and reliability. Additionally, the study concluded that the addition of student and flight instructor collaboration without an improved grading scale exerted a lesser, but nonetheless significant, effect upon the students' and flight instructor's perceptions of grade validity and reliability

Research Questions 1 through 4 led to the development of an effective instrument to determine the perception of grade validity and reliability. The literature provided guidance on the elements that compose grade validity, including fairness, accuracy, clarity, and communication (Butler, 2004; Messick, 1989; Schaeffner et al., 2000). The literature also identified the presence of clear and descriptive grade symbology, a stable system design, and rater (and interrater) reliability and objectivity (Feldt & Brennan 1989; Gall et al, 2003). The researcher concluded that these elements could be used to develop a survey instrument and a pre- and posttest experimental design that could

identify the presence of these elements through participant opinion (Gall et al.). The survey was utilized to produce the results in Research Questions 5, 6, and 7.

Research Question 5 produced evidence that students had little strong positive or negative perception about traditional grading systems. This may have been a product of the omnipresence of these types of systems from kindergarten to the university. This fact should not be taken as an endorsement of traditional grading, rather as a symptom of the desensitization that common usage creates. Ultimately, the participants did question the validity and reliability of the traditional grading system during the posttest survey and their written comments.

Research Question 6 produced statistically significant evidence that the addition of student-instructor collaboration improved the perception of the validity and reliability of the traditional grading system. This was the same grading symbology that produced no significant result during Research Question 5. Of 16 questions on the survey that dealt directly with validity and reliability, the participants scored 9 of them significantly higher. The researcher concluded that the way the grading was conducted, in this case through collaboration, may have been as important as the actual grade used or received. These results lent support to those authors in the literature who advocated for no grades at all (Kohn, 1994). The participants' comments, although mixed, supported the conclusion that increased collaboration between student and instructor will increase the perceived effectiveness of the grading system.

Research Question 7 produced significant evidence that the addition of clearer and more descriptive grade symbols, when combined with a collaborative grading system, will increase the perceived validity and reliability of the grades produced. Of the 16 questions on the survey that dealt directly with validity and reliability, the participants

scored 15 of them significantly higher. The research indicated that the addition of more descriptive grade options significantly increased student morale and motivation. This appeared to have a positive impact on student performance. Additionally, participants noted significant increases in feedback, communication, fairness, accuracy, and reliability. The combination of collaboration and the objective LCG-grading symbols appeared to eliminate the majority of the negative opinions expressed by participants about the traditional grading scale present in Research Questions 5 and 6. The increased grading options provided by the LCG grades as well as the positive and descriptive nature of the grades appeared to have made a significant difference in student perception.

Implications of Findings

The primary implication of this study was that the traditional grading system in place in the university flight training department appeared to have little positive or negative effect upon the student learning process. However, the addition of increased student-instructor collaboration and more objective and clearly defined LCG grade symbols appeared to promise increased student motivation and student instructor communication, trust, and confidence. The goal of these techniques was to increase student participation in their own training and, thus, increase the effectiveness of the learning process. LCG appeared to support this goal. There may be broader implications as well.

The literature used to develop this research was derived almost entirely from research completed in grade, middle, and high school settings. Thus, one might consider that the conclusions reached in this study might have broader applications than the field of flight instruction. LCG is based on the concepts of adult education (Brookfield, 1986; Knowles et al., 1998). Thus, the concepts of collaboration and LCG symbols might have

applications beyond aviation, especially in those fields in which students more readily exhibit the characteristics of adult learners.

Recommendations

The following four recommendations for further action have been made to the university to increase the effectiveness of the university flight training program:

1. The researcher recommends that the university adopt a collaborative grading system beginning in the fall of 2008. This will require the development of additional computer software to allow the student and instructor to enter grades simultaneously into the university flight training management system.

2. The researcher recommends that the university adopt the objective LCG symbols, developed for the study beginning in the fall of 2008. The grading symbols should be modified in accordance with the recommendation of the summative committee. This change to the university grading system will not require software modifications and can be accomplished by simply changing the grade descriptors in the university flight training management system. This study did not test these grade symbols without the presence of grade collaboration. However, based on the broad support found in the literature, the researcher recommends that these changes be made, even if the software changes required to introduce collaboration cannot be made in an expeditious manner.

3. The researcher recommends that the university develop a training program to introduce students and instructors to the concepts of collaboration and objective LCG symbols. This training program should be a part of the larger training envisioned as the university transitions to the FITS training methodology.

4. The researcher recommends that the university conduct a longitudinal study of the students who begin training in the fall of 2008 to determine the actual effect upon

training validity and reliability brought about by the inclusion of collaboration and objective LCG symbols in the flight training curriculum.

Limitations of the Applied Dissertation

The chief limitation of this research was the relatively small number present in each experimental group. Practical and financial limitations limited each group to approximately 10% of the available participants. Attrition during the experiment reduced Group B to approximately 9%. The procedures used to select and distribute the samples as the commonality of the pretest results and selected posttest results appeared to minimize this limitation. One could make the case that the small sample size may have contributed to the relatively robust results achieved. If the procedures resulted in a convenience sample, rather than a more representative group of participants, a more cohesive and less random group of participants might have induced error into the data (Gall et al., 2003).

However, the results of the participant selection procedures did not appear to support this conclusion. The participants in the experiment accurately reflected the demographics of the overall population. The arithmetic means of age, ethnicity, and gender of the participants fell within the demographic for the specific instrument flight course population. The random selection process ensured that this representative sample was evenly distributed across the three test groups. Once selected and surveyed, these data indicated strong preexperiment agreement between three diverse groups. These indicators appeared to support the representative nature and randomness of the samples as well as the statistical inference of the sample to the population.

The other significant limitation of this study was that it measured student and instructor perception of validity and reliability, rather than actual validity and reliability.

A longitudinal study of the effectiveness of LCG might reveal more about the actual impact of this new grading strategy on student learning. One probable result of this study is the proposed adoption of LCG by the entire flight department in the fall of 2008. This department-wide change presents an opportunity to examine actual validity and reliability as well the effect of LCG on the entire population.

Recommendations for Further Research

As previously noted, this research indicated a need for more rigorous research on the actual learning effectiveness of LCG. A longitudinal study of participants in the university flight training program compared to the data available in the university flight management software provided answers to this next and most important question: How effective is LCG in regard to student learning?

The proposed study might take two forms. First, a researcher might measure the actual validity and reliability of LCG on a larger sample. Second, the researcher might examine the larger question of actual impact upon student learning. Both questions might utilize a similar participant selection process. The entire student population might be divided up by grading practice with roughly half of all classes utilizing LCG and the other half utilizing the traditional grading scale. This would allow for the study of two large samples, each roughly 50% of the population and containing nearly 500 students per sample.

Validity of the actual grading practice might be measured by comparing actual student performance on required end-of-course examinations and check rides with the pattern of grades leading up to these events. Reliability could be examined by comparing the actual results of multiple student-instructor pairs over time, looking for rater reliability as well as interrater reliability. Based on the results to date, one would expect

these data to support the relatively robust results achieved in the current study. However, attributing increase student learning to LCG may be more difficult.

The number of variables that impact student learning appears to be significantly greater than those affecting grade validity and reliability. A researcher might establish milestones and metrics for speed and accuracy of student learning that could be applied to the same student and instructor population described above. The researcher would need to identify the specific impact of grading practice from among a host of variables present in the learning process. Careful work to isolate preexisting student aptitude, instructor ability, environmental factors, and other variables as yet unknown would need to be accomplished prior to undertaking an experiment of this scope. The resulting data would allow the researcher to measure the actual short-term effect of the increased communication, collaboration, and standardization of the grading process on the student learning. One might expect these data to be less robust than the results achieved to date due to the presence of additional variables that impact the overall learning process.

If accomplished, this study would build on this research through the development of instruments to measure actual grade validity, grade reliability, and learning effectiveness. The study might examine the progress of a cohort of students as they progress through an entire course or curriculum using LCG and compare them to a similar group using traditional grading. Learning effectiveness could be examined through a variety of measures designed to identify validity and reliability through actual student performance. The instrument and the methodology developed for this follow-up study could be applied to grading in other forms of education.

References

Anderson, R. A. (1998). Why talk about different ways to grade? The shift from traditional assessment to alternative assessment. In R. S. Anderson & B. W. Speck (Eds.), *New directions for teaching and learning: Changing the way we grade student performance. Classroom assessment and the new learning paradigm* (pp. 5-16). San Francisco: Jossey-Bass.

Baines, L., & Stanley, G. (2004). No more shopping for grades at the B-Mart: Reestablishing grades as indicators of academic performance. *The Clearing House, 77,* 101-104.

Blickensderfer, B., & Jennison, J. (2006). *Empirical investigation of the learner-centered grading debriefing approach* (FY 2005 FITS Instructor Education Research Report No. 4). Retrieved August 20, 2006, from http://www.faa.gov/education_research/training/fits/training/generic/media/course_developers.pdf

Bloom, B., Hastings J., & Madeus, G. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw Hill.

Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education, 18,* 529-549.

Brookfield, S. (1986). *Understanding and facilitating adult learning.* San Francisco: Jossey-Bass.

Butler, S. (2004). Question: When is a comment not worth the paper it's written on? Answer: When it's accompanied by a level, grade, or mark. *Teaching History, 115*, 37-42.

Byrnes, K. L. (2007) *Flight instructor orientation handbook*. Daytona Beach, FL: Embry Riddle.

Caffarella, R. S. (2002). *Planning programs for adult learners: A practical guide for educators, trainers, and staff developers*. San Francisco: Jossey-Bass.

Choinski, E., Mark, A., & Murphey, M. (2003). Assessment with rubrics: An efficient and objective means of assessing student outcomes in an information resources class. *Libraries and the Academy*, *3*, 563-575.

Connolly, T., Summers, M., & Ayers F. (2005). *FAA/Industry Training Standards scenario-based training course developers guide*. Retrieved August 20, 2006, from http://www.faa.gov/education_research/training/fits/training/generic/media/course_developers.pdf

Crocker, J., Quinn, M., Karpinski, A., & Chase, S. (2003). When grades determine self-worth: Consequences of contingent self-worth for male and female engineering

and psychology majors. *Journal of Personality and Social Psychology, 85*, 507-515.

Davis, W., Fedor, D., Parsons, C., & Herold, D. (2000). The development of self-efficacy during aviation training. *Journal of Organizational Behavior, 21*, 857-866.

Department of Transportation. (1999). *Aviation instructor's handbook*. Retrieved August 20, 2006, from http://www.sportpilot.info/sp/FAA-H-8083-9 _Aviation_Instructors_Handbook.pdf

Duffy, T. M., & Jonassen, D. H. (1992). *Constructivism and the technology of instruction: A conversation*. Hillsdale, NJ: Lawrence Erlbaum and Associates.

Ennels, A. (2002). The Wright stuff: Pilot training at America's first civilian flying school. *Air Power History, 49*(4), 22-32.

Embry Riddle. (2007). *Tuition, room and board, and mandatory fees*. Retrieved August 20, 2007, from http://www.erau.edu/er/costs.html

Federal Aviation Administration. (2003). *FAA/Industry Training Standards program plan*. Retrieved August 20, 2006, from http://www.faa.gov/education_research/ training/ fits/media/program%20plan.doc

Feldt, L., & Brennan, R. (1989). Reliability. In R. Linn (Ed.), *Educational measurement* (pp.105-146). New York: McMillan.

Flight Standards Service. (2002). *Private pilot practical test standard for airplanes*. Retrieved August 20, 2006, from http://www.faa.gov/education_research/testing/ airmen/test_standards

Gagne, R. M., Briggs, L. J., & Wager, W. (1992). *Principles of instructional design*. New York: Harcourt, Brace, and Jovanovich.

Gall, M. D., Borg, W. R., & Gall, J. P. (2003). *Educational research: An introduction* (7th ed.). White Plains, NY: Longman.

Gopinath, C. (2004). Exploring effects of criteria and multiple graders on case grading. *Journal of Education for Business, 79*, 317-322.

Hendrickson, J., Gable, R., & Manning, M. (1999). Can everyone make the grade? Some thoughts on student grading in the contemporary classroom. *The High School Journal, 82*, 248-253

Holmes, L., & Smith, L. (2003). Student evaluations of faculty grading methods. *Journal of Education for Business, 78*, 318-323.

Knowles, M., Holton, E., & Swanson, R. (1998). *The adult learner: The definitive classic*

in adult education and human resource development. Wobern, MA: Butterworth-Heinemann.

Kohn, A. (1994). *Grading: The issue is not how but why*. Retrieved August 12, 2006, from http://www.Alfiekohn.org/teaching/grading.htm

Kuisma, R. (1999). Criteria-referenced marking of written assignments. *Assessment and Evaluation in Higher Education, 24*(1), 27-40.

Lunsford, E., & Melear, C. (2004). Using scoring rubrics to evaluate inquiry. *Journal of College Science and Teaching, 34*(1), 34-38.

Merrill, J. (2003). Record your ensemble for better learning. *Teaching Music, 11*(3), 34-36.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (pp. 105-146). New York: McMillan.

Michaels, J. (1976). A simple view of the grading issue. *Teaching Sociology, 3*(2), 198-203.

Milton, O., & Edgerly, J. W. (1976). *The testing and grading of students*. Stanford, CA: Carnegie Foundation.

Salvia, J., & Ysseldyke, J. E. (2007). *Assessment in special and inclusive education* (10th ed.). Boston: Houghton Mifflin.

Schaffner, M., Burry-Stock, J., Cho, G., Boney, T., & Hamilton, G. (2000, April). *What do kids think when their teachers grade*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Shaw, J. (2004). Demystifying the evaluation process for parents: Rubrics for marking student research projects. *Teacher Librarian, 32*(2), 16-20.

Shim, M., Felner, R., Shim, E., & Noonan, N. (2001, April). *Multidimensional assessment of classroom internal practice: A validity study of the classroom instructional practice scale*. Paper presented at the annual meeting of the American Educational Research Society, Seattle, WA.

Speck, B. W. (1998). Unveiling some of the mystery of professional judgment in classroom assessment. In R. S. Anderson & B. W. Speck (Eds.), *New directions for teaching and learning: Changing the way we grade student performance. Classroom assessment and the new learning paradigm* (pp.17-32). San Francisco: Jossey-Bass.

Stefani, L. (1998). Assessment in partnership with learners. *Assessment and Evaluation in Higher Education, 23*, 339-350.

Wooley, S., & Wooley, A. (1999, April).*Can we change teachers beliefs? A survey about constructivist and behaviorist approaches*. Paper presented at the annual meeting of the American Educational Research Society, Montreal, Quebec, Canada.

World Educational Services. (2007). *WES grade conversion guide*. Retrieved June 3, 2007, from http://www.wes.org/gradeconversionguide/index.asp

Wright, R. A. (2002). *Changes in general aviation flight operations and their impact on systems safety and flight training*. Unpublished manuscript.

Appendix A

Student Survey

Student Survey

Participant Survey ID# _____

Demographic Information

Age (circle one):

(18–25) (26-35) (36-45) (46-55) (56 or older)

Gender (circle one):

(female) (male)

I am a (circle one):

(freshman) (sophomore) (junior) (senior) (graduate student)

The flight ratings I hold are (circle any or all applicable):

(private) (instrument) (commercial) (multiengine) (CFI) (CFI/II) (ATP)

How many flight instructors have you had since you began to fly at the university?

(1) (2) (3) (4) (5) (6) (7) (8) (9) (10 or more)

Answer the following questions based on the grading process used during the previous five Flight Training Device lessons.

Please circle the number that corresponding to the response that best indicates your agreement with the statement listed below.

	Strongly Disagree	Disagree	No Opinion	Agree	Strongly Agree
Purpose of the lesson grading process					
1. I believe the grade process improves my instructor's authority over his/her students.	1	2	3	4	5
2. I believe the grade process compares me to other students by how proficient we are.	1	2	3	4	5
3. I believe the grade process compares me to a published and easy to understand standard.	1	2	3	4	5
4. I believe the grade process provides feedback to help improve my performance	1	2	3	4	5
5. I believe the grade process motivates me to improve my work.	1	2	3	4	5
Collaboration and participation					
6. I believe my instructor is more critical of my performance than I am.	1	2	3	4	5
7. I believe I am more critical of my own performance than my instructor is.	1	2	3	4	5
8. I believe it is important that the instructor decide what we do and how we do it.	1	2	3	4	5
9. I believe it is important that I decide what we do and how we do it.	1	2	3	4	5
10. I believe it is important that the instructor and I work together to decide what we do and how we do it.	1	2	3	4	5
Emotional and self-esteem impact of the grade					
11. I believe the grading system I used motivated me to work harder.	1	2	3	4	5
12. I believe the grading system I used made me feel more positive about my FTD lessons.	1	2	3	4	5
13. I believe the grading system I used motivated me to work harder when I received a low grade.	1	2	3	4	5

14. I believe the grading system I used motivated me to work harder when I received a high grade.	1	2	3	4	5
15. I believe the lesson grades I received reflected my instructor's good or bad attitude about me.	1	2	3	4	5
16. I believe the lesson grades I receive reflected my good or bad attitude about my instructor.	1	2	3	4	5

Validity of the grade process

17. I believe the grades I received were fair.	1	2	3	4	5
18. I believe the grades I received were accurate.	1	2	3	4	5
19. I believe the grades I received were descriptive of my performance	1	2	3	4	5
20. I believe I only receive a low grade when my instructor needs to justify an need an extra lesson (XT) or I have to repeat a lesson.	1	2	3	4	5
21. I believe the lesson grade I received reflected my performance as compared to my instructor's other students.	1	2	3	4	5

Reliability of the grade process

22. I believe the grades I received were consistent with my performance.	1	2	3	4	5
23. I believe my instructor grades me consistently from lesson to lesson	1	2	3	4	5
24. I believe different instructors grade me the same way.	1	2	3	4	5
25. I believe grading process we used will help instructors grade all students more consistently	1	2	3	4	5

Impact on the learning process

26. I believe the way the lesson was graded improved the amount of feedback I get from my instructor.	1	2	3	4	5
27. I believe the grading process we used had a positive impact on the lesson post-FTD debriefing.	1	2	3	4	5

Importance of the grading process

28. I believe individual task grades are the most important to me.	1	2	3	4	5

29. I believe the overall lesson grade is the most important to me.	1	2	3	4	5
30. I believe all grades are important to me.	1	2	3	4	5

Please add any additional comments, questions, or suggestions in the space provided below. Reference each comment with the specific survey question number. Thank you!

Group 2 and 3 only	Strongly Disagree	Disagree	No Opinion	Agree	Strongly Agree
1. I believe that participating in the lesson grading increased the accuracy of my grades.	1	2	3	4	5
2. I believe that participating in the lesson grading increased the fairness of my grades.	1	2	3	4	5
3. I believe that participating in the lesson grading increased the consistency of my grades.	1	2	3	4	5
4. I believe that my participation in the lesson grading helped my instructor understand me better.	1	2	3	4	5
5. I believe that my participation in the lesson grading helped me understand my instructor better.	1	2	3	4	5
6. I believe the grading scale (the actual grade) we used gives the grader an accurate way to describe student performance.	1	2	3	4	5
7. I believe the grading scale (the actual grade) we used gives the grader enough options (grades) to describe student performance.	1	2	3	4	5
8. I believe that the grade the instructor and I arrived at together is more accurate than the individual grades we arrived at.	1	2	3	4	5

Please add any additional comments, questions, or suggestions in the space provided below. Reference each comment with the specific survey question number. Thank you!

Appendix B

Flight Instructor Survey

Flight Instructor Survey

Participant Survey ID# _____

Demographic Information

Age (circle one):

(18-25) (26-35) (36-45) (46-55) (56 or older)

Gender (circle one):

(female) (male)

I am a (circle one):

(freshman) (sophomore) (junior) (senior) (graduate student) (part-time instructor) (full-time instructor)

The flight ratings I hold are (circle any or all applicable):

(private) (instrument) (commercial) (multiengine) (CFI) (CFI/II) (ATP)

How many students do you have assigned?

(1) (2) (3) (4) (5) (6) (7) (8) (9) (10 or more)

Answer the following questions based on the grading process used during the previous five Flight Training Device lessons.

Please circle the number that corresponding to the response that best indicates your agreement with the statement listed below.

Purpose of the lesson grading process	Strongly Disagree	Disagree	No Opinion	Agree	Strongly Agree
1. I believe the grade process improves an instructors authority over his/her students.	1	2	3	4	5
2. I believe the grade process compares my students to other students I fly with.	1	2	3	4	5
3. I believe the grade process compares my students to a published standard.	1	2	3	4	5
4. I believe the grade process provides feedback to help improve my students' performance.	1	2	3	4	5
5. I believe the grade process motivates my students to improve.	1	2	3	4	5
Collaboration and participation					
6. I believe my students are more critical of their performances than I am.	1	2	3	4	5
7. I believe I am more critical of my students' performance than they are.	1	2	3	4	5
8. I believe it is important that the instructor decide what we do and how we do it.	1	2	3	4	5
9. I believe it is important that the students decide what we do and how we do it.	1	2	3	4	5
10. I believe it is important that the students and I work together to decide what we do and how we do it.	1	2	3	4	5
Emotional and self-esteem impact of the grade					
11. I believe the grading system I used motivated my students to work harder.	1	2	3	4	5
12. I believe the grading system I used made my students feel more positive about my FTD lessons.	1	2	3	4	5
13. I believe the grading system I used motivated My students to work harder when they received a low grade.	1	2	3	4	5

14. I believe the grading system I used motivated my students to work harder when they received a high grade.	1	2	3	4	5
15. I believe the lesson grades I give reflect my students' good or bad attitudes.	1	2	3	4	5
16. I believe the lesson grades I give reflect my good or bad attitude about my students.	1	2	3	4	5

Validity of the grade process

17. I believe the grades I awarded were fair.	1	2	3	4	5
18. I believe the grades I awarded were accurate.	1	2	3	4	5
19. I believe the grades I awarded were descriptive of my students' performances.	1	2	3	4	5
20. I believe I only award a low grade when I need to justify an need an extra lesson (XT) or I have to repeat a lesson.	1	2	3	4	5
21. I believe the lesson grades I award reflect my students' performances as compared to my other assigned students.	1	2	3	4	5

Reliability of the grade process

22. I believe the grades I awarded were consistent with my students' performances.	1	2	3	4	5
23. I believe I graded my students consistently from lesson to lesson	1	2	3	4	5
24. I believe different instructors grade all students the same way.	1	2	3	4	5
25. I believe the grading process we used will help instructors grade all students more consistently.	1	2	3	4	5

Impact on the learning process

26. I believe the way the lesson was graded improved the amount of feedback I get from my students.	1	2	3	4	5
27. I believe the grading process we used had a positive impact on the lesson post-FTD debriefing.	1	2	3	4	5

Importance of the grading process

28. I believe individual task grades are the most important to my students.	1	2	3	4	5

29. I believe the overall lesson grade is the most important to my students.	1	2	3	4	5
30. I believe all grades are important to my students.	1	2	3	4	5

Please add any additional comments, questions, or suggestions in the space provided below. Reference each comment with the specific survey question number. Thank you!

Group 2 and 3 only	Strongly Disagree	Disagree	No Opinion	Agree	Strongly Agree
1. I believe that participating in the lesson grading increased the accuracy of my grades.	1	2	3	4	5
2. I believe that participating in the lesson grading increased the fairness of my grades.	1	2	3	4	5
3. I believe that participating in the lesson grading increased the consistency of my grades.	1	2	3	4	5
4. I believe that my participation in the lesson grading helped me understand my students.	1	2	3	4	5
5. I believe that my participation in the lesson grading helped me understand my students better.	1	2	3	4	5
6. I believe the grading scale (the actual grade) we used gives the grader an accurate way to describe student performance.	1	2	3	4	5
7. I believe the grading scale (the actual grade) we used gives the grader enough options (grades) to describe student performance.	1	2	3	4	5
8. I believe that the grade the instructor and I arrived at together is more accurate than the individual grades we arrived at separately.	1	2	3	4	5

Please add any additional comments, questions, or suggestions in the space provided below. Reference each comment with the specific survey question number. Thank you!

Appendix C

Grading Instrument 1

Grading Instrument 1

Participant ID Number_____Date_____ FTD Lesson Number _____

Grading Procedure

Immediately following the FTD lesson the instructor and student will independently fill out the grade sheet for the FTD activity. Once complete, the instructor and student will compare the grades they have written down and decide upon a final grade for the lesson to be entered in to ETA. These grade sheets will be placed in the box at the FTD dispatch desk upon completion of the activity. If a question arises, please call Frank Ayers at 386-437-5211 at any time. Thanks for your participation.

Grading Scale

Outstanding (O)
Good (G)
Minimum (M)
Unsatisfactory (U)
Incomplete (I)

COCKPIT PROCEDURES	Circle One
1. Cockpit Management	O G M U I
2. Use of Checklists	O G M U I
3. Engine Starting	O G M U I
4. Instrument Cockpit Check	O G M U I
5. Checking Instruments and Equipment	O G M U I
BASIC INSTRUMENT FLIGHT MANUEVERS	
6. Basic Instrument Flight Maneuvers	O G M U I
7. Maneuvers During Slow Flight	O G M U I
8. Power – On Stalls	O G M U I
9. Power off Stalls	O G M U I
10. BAI Pattern “A”	O G M U I
11. BAI Pattern “B”	O G M U I
VOR PROCEDURES	
12. Intercepting and Tracking VOR	O G M U I
13. Holding Procedures VOR	O G M U I
14. VOR Approach procedures	O G M U I
15. VOR Holding	O G M U I
16. Missed Approach	O G M U I
17. Non precision Approach Procedures	O G M U I
GPS PROCEDURES	
18. Intercepting and Tracking GPS	O G M U I
19. Holding procedures GPS	O G M U I

20. GPS Approach Procedures	O	G	M	U	I
21. GPS Holding	O	G	M	U	I
22. Missed Approach	O	G	M	U	I
23. Non precision Approach Procedures	O	G	M	U	I
LOCALIZER PROCEDURES					
24. Intercepting and Tracking Localizer	O	G	M	U	I
25. Holding Procedures Localizer	O	G	M	U	I
26. Localizer Holding	O	G	M	U	I
27. Localizer Approach Procedure	O	G	M	U	I
28. Missed Approach	O	G	M	U	I
29. Non precision Approach Procedures	O	G	M	U	I
30. Localizer Back Course Procedure	O	G	M	U	I
ILS PROCEDURES					
31. ILS Approach Procedure	O	G	M	U	I
32. Missed Approach	O	G	M	U	I
33. Precision Approach Procedures	O	G	M	U	I
DME PROCEDURES					
34. Intercepting and tracking DME Arc	O	G	M	U	I
35. Holding procedures DME	O	G	M	U	I
CROSS COUNTRY AND ATC CLEARANCE					
36. Air Traffic Control Clearances	O	G	M	U	I
37. IFR Cross-Country Procedures	O	G	M	U	I
38. Compliance with Departure, En Route and Arrival	O	G	M	U	I
Procedures and Clearances	O	G	M	U	I
EMERGENCY PROCEDURES					
39. Loss of primary Flight Instrument Indicators	O	G	M	U	I
40. Magnetic Compass Turns	O	G	M	U	I
41. Timed Turns to Magnetic Compass headings	O	G	M	U	I
42. System and Equipment Malfunctions	O	G	M	U	I
43. Recovery fro Unusual Flight Attitudes	O	G	M	U	I
44. Approach with Loss of primary Flight Instruments	O	G	M	U	I
45. Loss of Communication	O	G	M	U	I
46. Approach with Loss of Primary Flight Inst Indicators	O	G	M	U	I
47. Emergency Procedures	O	G	M	U	I
LANDING					
48. Landing From a Straight –in or Circling Approach	O	G	M	U	I
MISC PROCEDURES					
49. Holding procedures Intersection	O	G	M	U	I
50. Circling Approach	O	G	M	U	I

Appendix D

Grading Instrument 2

Grading Instrument 2

Participant ID Number_____Date_____ FTD Lesson Number _____

Grading Procedure

Immediately following the FTD lesson the instructor and student will independently fill out the grade sheet for the FTD activity. Once complete, the instructor and student will compare the grades they have written down and decide upon a final grade for the lesson to be entered into ETA. These grade sheets will be placed in the box at the FTD dispatch desk upon completion of the activity. If a question arises, please call Frank Ayers at 386-437-5211 at any time. Thanks for your participation.

Grading Scale

Performing (P1)
Practicing (P2)
Learning (L)
Regressing (R)
Incomplete (I)

Performing (P1)	At the completion of the lesson, the student will be able to perform the without assistance from the instructor. Errors and deviations will be identified and corrected by the student in an expeditious manner. The student meets the PTS standard. At no time will the successful completion of the activity be in doubt
Practicing (P2)	At the completion of the lesson the student will be able to practice the activity with input from the instructor. The student, with coaching or assistance from the instructor, will quickly correct minor deviations and errors. The student deviates from the PTS standard from time to time.
Learning (L)	At the completion of the lesson the student has been recently introduced to a task or maneuver and requires significant help from the instructor to complete it. The student is making good progress toward the practicing level. The student is working toward the PTS standard and will achieve this level of performance when the syllabus requires it.
Regressing (R)	At the completion of the task the student and instructor agree that the student does not fully understand or needs more practice to make progress. This grade requires the student and instructor to discuss the plan for the next lessons and may require additional training. The student will not meet the PTS standard when the syllabus requires.
Incomplete (I)	The scheduled task or maneuver was not performed and will need to be made up in future lessons.

COCKPIT PROCEDURES		Circle One
1.	Cockpit Management	P1 P2 L R I
2.	Use of Checklists	P1 P2 L R I
3.	Engine Starting	P1 P2 L R I
4.	Instrument Cockpit Check	P1 P2 L R I
5.	Checking Instruments and Equipment	P1 P2 L R I
BASIC INSTRUMENT FLIGHT MANUEVERS		
6.	Basic Instrument Flight Maneuvers	P1 P2 L R I
7.	Maneuvers During Slow Flight	P1 P2 L R I
8.	Power – On Stalls	P1 P2 L R I
9.	Power off Stalls	P1 P2 L R I
10.	BAI Pattern “A”	P1 P2 L R I
11.	BAI Pattern “B”	P1 P2 L R I
VOR PROCEDURES		
12.	Intercepting and Tracking VOR	P1 P2 L R I
13.	Holding Procedures VOR	P1 P2 L R I
14.	VOR Approach procedures	P1 P2 L R I
15.	VOR Holding	P1 P2 L R I
16.	Missed Approach	P1 P2 L R I
17.	Non precision Approach Procedures	P1 P2 L R I
GPS PROCEDURES		
18.	Intercepting and Tracking GPS	P1 P2 L R I
19.	Holding procedures GPS	P1 P2 L R I
20.	GPS Approach Procedures	P1 P2 L R I
21.	GPS Holding	P1 P2 L R I
22.	Missed Approach	P1 P2 L R I
23.	Non precision Approach Procedures	P1 P2 L R I
LOCALIZER PROCEDURES		
24.	Intercepting and Tracking Localizer	P1 P2 L R I
25.	Holding Procedures Localizer	P1 P2 L R I
26.	Localizer Holding	P1 P2 L R I
27.	Localizer Approach Procedure	P1 P2 L R I
28.	Missed Approach	P1 P2 L R I
29.	Non precision Approach Procedures	P1 P2 L R I
30.	Localizer Back Course Procedure	P1 P2 L R I
ILS PROCEDURES		
31.	ILS Approach Procedure	P1 P2 L R I
32.	Missed Approach	P1 P2 L R I
33.	Precision Approach Procedures	P1 P2 L R I
DME PROCEDURES		
34.	Intercepting and tracking DME Arc	P1 P2 L R I

35.	Holding procedures DME	P1	P2	L	R	I

CROSS COUNTRY AND ATC CLEARANCE

36.	Air Traffic Control Clearances	P1	P2	L	R	I
37.	IFR Cross-Country Procedures	P1	P2	L	R	I
38.	Compliance with Departure, En Route and Arrival	P1	P2	L	R	I
39.	Procedures and Clearances	P1	P2	L	R	I

EMERGENCY PROCEDURES

40.	Loss of primary Flight Instrument Indicators	P1	P2	L	R	I
41.	Magnetic Compass Turns	P1	P2	L	R	I
42.	Timed Turns to Magnetic Compass headings	P1	P2	L	R	I
43.	System and Equipment Malfunctions	P1	P2	L	R	I
44.	Recovery fro Unusual Flight Attitudes	P1	P2	L	R	I
45.	Approach with Loss of primary Flight Instruments	P1	P2	L	R	I
46.	Loss of Communication	P1	P2	L	R	I
47.	Approach with Loss of Primary Flight Inst Indicators	P1	P2	L	R	I
48.	Emergency Procedures	P1	P2	L	R	I

LANDING

49.	Landing From a Straight –in or Circling Approach	P1	P2	L	R	I

MISC PROCEDURES

50.	Holding procedures Intersection	P1	P2	L	R	I
51.	Circling Approach	P1	P2	L	R	I