A Study of Student Perception of the Validity and Reliability in University Flight Training Assessment

Francis H. Ayers, Jr.
Embry Riddle Aeronautical University

ABSTRACT

This paper examines 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 seven research questions, this paper sought to determine the student perception of the validity and reliability of each portion of the new grading symbols as well as the collaborative grading technique. 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 four major recommendations. The primary recommendation was that the university adopt the learner-centered grading system described in the study.

INTRODUCTION

This paper examines 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. This paper 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.

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 (teacher centered) approach to flight training to an androgogical (learner centered) 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).

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 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 Aeronautical Science, a 4-year degree program. Flight training students flew approximately 200 hours in small, single, and multiengine aircraft as well as flight simulators and earned FAA approved pilot proficiency ratings.

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. 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 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.

Research Questions

This study was guided by the following research questions:

1. What does the literature suggest about the validity and reliability of traditional and/or LCG grading procedures in aviation or other more conventional classroom education programs?
2. How should the perceived validity and reliability of flight student assessment programs be evaluated?
3. How do the participants (instructors and students) in the study rate the validity and reliability of traditional grading techniques?
4. 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.
5. 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.

REVIEW OF THE LITERATURE

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 and understanding (Gall, Borg, & Gall, 2003).

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.

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.

CURRENT APPROACHES

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.

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 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, co-inventor 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 a 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.

The university flight grading system represented a traditional approach. 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.

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.

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, 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.

The grades of outstanding, good, and marginal 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.

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 (Figure 2) 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. 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.

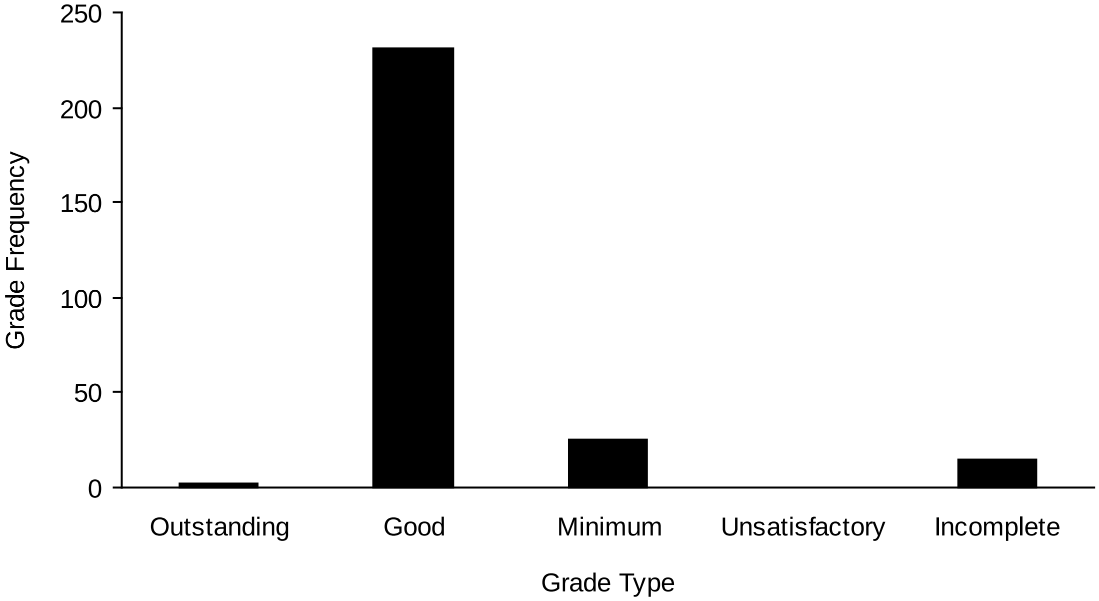

Figure 1. A graphic depiction of a small sample of student grades.

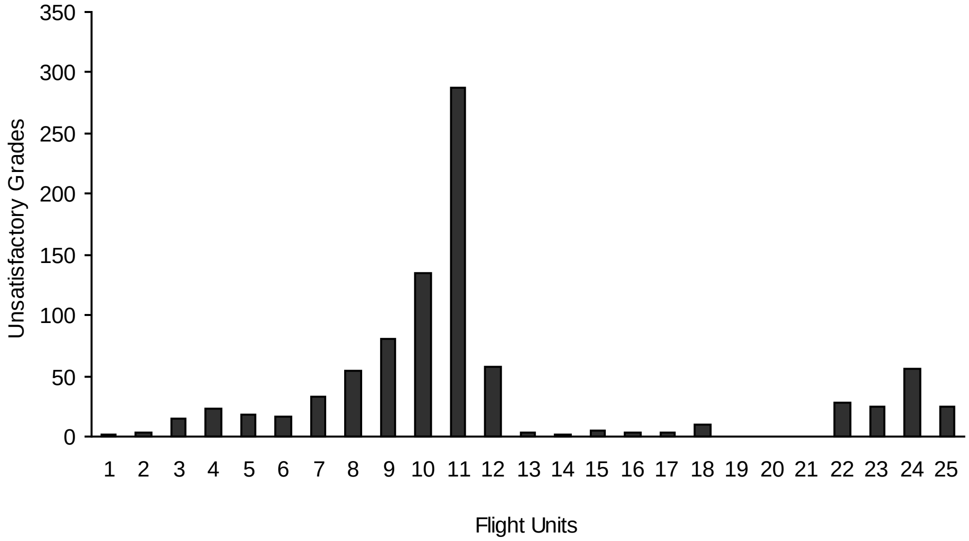

Figure 2. The number of individual lessons graded unsatisfactory by flight unit.

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 & 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.

The approach to grading 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.

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.

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.

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.

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?

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 of 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.

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 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.

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.

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 simulator 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. 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.

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).

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.

METHODOLOGY AND PROCEDURES

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. 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.

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. The following table represents the experimental LCG grading scale utilized in the study.

Hypothesis

The primary hypothesis of the study was that students and instructors would demonstrate a statistically significant difference in grade validity and reliability, in the presence of grade collaboration as well as the LCG grading scale. Since collaboration was utilized during two experimental iterations, the first with traditional grade descriptors only, and the second in concert with the LCG grade descriptors, students and instructor preference would be assumed to increase as each new element was added to the experiment. The null hypothesis stated that there will be no significant difference between the perceived validity and reliability of the grading systems utilizing collaboration and/or LCG grading descriptors (Table 4).

Table 4. *Experimental Learner-Centered Grading (LCG) 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.

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.

RESULTS

The researcher of this study sought to determine the most appropriate form of lesson grading consistent with the desired university approach to flight training.

Results of the Control Group (Traditional Grading, Group A)

In this experiment, the instructor assigns the student performance task grades using the traditional grading scale currently in use in the flight training department. Thirty-four flight students and flight instructors (36 completed the pretest, and 2 were unable to complete the entire experiment) participated in the control group.

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.

Table 5. *Group A, 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.

These results appeared to provide support for the null hypothesis (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. Specific areas of pre/post test disagreement are noted in the following table.

These data appear to support the notion that students, when given the opportunity to reflect (five iterations) upon a traditional grading system, do not express a strong preference for, and do express at least some negative preferences in the areas of grade validity and reliability.

Results of Grade Collaboration in the Presence of Traditional Grading (Group B).

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 (Table 6). Instructor only data are not presented due to the very low number for instructor participants. The mean of the scores of the combined student and instructor 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 control group data as well as the grade collaboration group presurvey, suggested a positive outcome for grade collaboration.

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 6.

Results of Grade Collaboration and LCG Combined (Group C).

In this form of grading (Table 7), the students self-assigned 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.

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.

Table 6. *Group B, 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.

Table 7. *Group C, 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.

When only the questions were considered that made positive statements about grade validity and reliability (Table 8), 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).

Table 8. *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.

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. Table 7 depicts these data

Table 9. *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.

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.

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 (Table 10). The total number of pretest comments was compared to gauge the relative pre survey level of agreement between the groups. Total number of post survey comments was compiled as an informal method of gauging the enthusiasm of the participants. The results, although not meant to be empirical, were nonetheless interesting.

Although pre experiment 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 post experiment survey to the Group B and C post experiment surveys.

Table 10. *Group A, B, and C Anecdotal Written Survey Comments*

Group	Pre Survey Negative	Post Survey Negative	Pre Survey Positive	Post Survey 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.

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 post survey comments to say that the use of the unsatisfactory grade during the pre solo 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 post experiment 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 post experiment 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.

DISCUSSION

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 four major recommendations. The primary recommendation was that the university adopt the learner-centered grading system described in the study.

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.

Discussion of Conclusions

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.

Of note, the group C data 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.

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. 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. 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.

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 will provide 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 increased 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 April 3, 2009, 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.

Connolly, T., Summers, M., & Ayers F. (2005). *FAA/Industry Training Standards scenario-based training course developers guide*. Retrieved April 3, 2009, from http://www.faa.gov/training_testing/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 April 3, 2009, from http://www.faa.gov/library/manuals/aviation/media/FAA-H-8083-9.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.

Federal Aviation Administration. (2003). *FAA/Industry Training Standards program plan*. Retrieved April 3, 2009, 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. (2006). *Private pilot practical test standard for airplanes.* Retrieved April 3, 2009, 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.

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

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

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.

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.

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

APPENDIX A

Flight Instructor Survey Questions

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		No	Strongly	
	Disagree	Disagree	Opinion	Agree	Agree
1. I believe the grade process improves an Instructor’s 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!