

5-2021

Gold Standards Training and Evaluator Calibration of Pilot School Check Instructors

Paul M. Cairns
Embry-Riddle Aeronautical University, cairnsp@erau.edu

Follow this and additional works at: <https://commons.erau.edu/edt>



Part of the [Aviation and Space Education Commons](#)

Scholarly Commons Citation

Cairns, Paul M., "Gold Standards Training and Evaluator Calibration of Pilot School Check Instructors" (2021). *PhD Dissertations and Master's Theses*. 568.
<https://commons.erau.edu/edt/568>

This Thesis - Open Access is brought to you for free and open access by Scholarly Commons. It has been accepted for inclusion in PhD Dissertations and Master's Theses by an authorized administrator of Scholarly Commons. For more information, please contact commons@erau.edu.

**Gold Standards Training and Evaluator Calibration of Pilot School Check
Instructors**

Paul M. Cairns

Thesis Submitted to the College of Aviation in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Aeronautics

Embry-Riddle Aeronautical University

Daytona Beach, Florida

May 2021

© 2021 Paul M. Cairns

All Rights Reserved.

Gold Standards Training and Evaluator Calibration of Pilot School Check

Instructors

Paul M. Cairns

This thesis was prepared under the direction of the candidate's Thesis Committee Chair, Dr. Andrew R. Dattel, and has been approved by the members of the thesis committee.

It was submitted to the College of Aviation and was accepted in partial fulfillment of the requirements for the Degree of Master of Science in Aeronautics.



Andrew R. Dattel, Ph.D.
Committee Chair



Kenneth P. Byrnes, Ph.D.
Committee Member



Donald S. Metscher, D.B.A.
Master of Science in Aeronautics
Program Coordinator



Steven Hampton, Ed.D.
Associate Dean, School of Graduate
Studies, College of Aviation



Alan J. Stolzer, Ph.D.
Dean, College of Aviation

Christopher D. Grant, Ph.D.
Associate Provost of Academic Support

April 9, 2021

Date

Abstract

Researcher: Paul M. Cairns

Title: Gold Standards Training and Evaluator Calibration of Pilot School Check
Instructors

Institution: Embry-Riddle Aeronautical University

Degree: Master of Science in Aeronautics

Year: 2021

A key component of air carrier advanced qualification programs is the calibration and training of instructors and evaluators and assurance of reliable and valid data in support of such programs. A significant amount of research is available concerning the calibration of air carrier evaluators, but no research exists regarding the calibration of pilot school check instructors. This study was designed to determine if pilot school check instructors can be calibrated against a gold standard to perform reliable and accurate evaluations. Calibration followed the principles and theories of andragogy and adult learning and teaching, including emphasis on the cognitive domain of learning, learner-centered instruction, and human resource development. These in combination with methods commonly used in aviation instruction aimed to increase the effectiveness of the calibration. Discussion of these combinations is included. A specific method for delivery of the calibration was provided along with a complete lesson plan. This study used a one group pretest-posttest design. A group of 10 pilot school check instructors were measured before and after receiving rater calibration training. Statistical measures included raw inter- and referent-rater agreement percentages, Cohen's kappa and kappa-like statistics for inter- and referent-rater reliability, Pearson product-moment correlations for

sensitivity to true changes in pilot performance, and a standardized mean absolute difference for grading accuracy. Improvement in all the measurements from pretest to posttest was expected, but actual results were mixed. However, a holistic interpretation of the results combined with feedback from the check instructors showed promise in calibration training for pilot school check instructors. Thorough discussion of the limitations and lessons learned from the study, recommendations for pilot schools, and recommendations for future research is included.

Keywords: behavioral indicator, calibration, check instructor, competency, gold standard, pilot school, rater reliability

Table of Contents

	Page
Signature Page	iii
Abstract	iv
List of Tables	x
List of Figures	xi
Chapter I: Introduction.....	1
Statement of the Problem.....	1
Purpose Statement.....	2
Significance of the Study	3
Research Question	4
Delimitations.....	4
Limitations and Assumptions	5
Summary	5
Definitions of Terms	6
List of Acronyms	9
Chapter II: Review of the Relevant Literature.....	11
Referents and Gold Standards.....	11
Calibration Studies.....	12
Statistical Measures	13
Competencies.....	15
Learning Theories Applicable to Calibration	16
Teaching Methods Applicable to Calibration.....	19
Human Resource Development	21

Gaps in the Literature.....	23
Theoretical Framework.....	23
Research Model	24
Summary.....	24
Chapter III: Methodology	26
Research Method Selection.....	26
Population/Sample	26
Population and Sampling Frame.....	27
Sample Size.....	27
Sampling Strategy.....	28
Confederates	28
Calibration Facilitator	29
Data Collection Process	29
Design and Procedures.....	30
Apparatus and Materials	35
Sources of the Data	41
Ethical Considerations	41
Measurement Instrument	41
Variables and Scales	41
Instrument Reliability	42
Instrument Validity	42
Data Analysis Approach	43
Participant Demographics.....	44

Reliability Assessment Method	44
Validity Assessment Method	44
Data Analysis Process.....	45
Summary.....	45
Chapter IV: Results.....	47
Demographics Results	47
Descriptive Statistics.....	47
Reliability and Validity Testing Results	49
Quantitative Data Analysis Results	50
Agreement Results	50
Reliability Results.....	52
Sensitivity Results.....	57
Accuracy Results	59
Summary.....	62
Chapter V: Discussion, Conclusions, and Recommendations	63
Discussion.....	63
Inter-Rater Reliability	63
Referent-Rater Reliability.....	65
Sensitivity and Accuracy	67
Check Instructor Feedback and Discussion Notes.....	71
Conclusions.....	79
Theoretical Contributions	81
Practical Contributions.....	81

Limitations of the Findings.....	82
Recommendations.....	84
Recommendations for Pilot Schools.....	84
Recommendations for Future Research.....	85
References.....	87
Appendices.....	90
A Permission to Conduct Research.....	90
B Basic Competencies and Behavioral Indicators.....	96
C Maneuver Evaluation Grade Sheets.....	98
D Slide Show Used for Grading System Training.....	105
E Lesson Plan for Check Instructor Calibration.....	127
F Video Scenarios and Associated Aeronautical Information.....	135

List of Tables

Table	Page
1 Five-Point Grading Scale	39
2 Participant Demographic Descriptive Statistics	47
3 Summary of Gold Standard and Rater Scores Across All Graded Items	49
4 Percentages of Agreement Across All Graded Items	52
5 Pretest Pairwise Comparisons of Inter-Rater Reliability	53
6 Posttest Pairwise Comparisons of Inter-Rater Reliability	54
7 Average Inter-Rater Reliability Across All Rater Pairs	55
8 Pretest to Posttest Pairwise Comparisons of Intra-Rater Reliability	56
9 Referent-Rater Reliability Across All Graded Items	57
10 Sensitivity Correlations With The Gold Standard	59
11 Accuracy Results Across All Maneuvers Segments	61

List of Figures

Figure		Page
1	Calibration Procedure.....	30
2	Example of Pre-Recorded Check Ride Maneuver Segment Videos	38
3	Pretest SMAD Measurements with Color Coding	70
4	Pratt's Four-Quadrant Model Applied to Check Instructor Calibration.....	79

Chapter I: Introduction

For decades, many large, U.S.-based air carriers have used a voluntary, alternative training program called advanced qualification program (AQP) to train their pilots, instructors, and evaluators. AQP uses proficiency-based training and evaluation centered around the concepts of crew resource management (CRM) (Air Carrier Operations Branch, 2017). The CRM behaviors that are trained and evaluated are done so through the use of line-operational simulations (LOS) that replicate the real-life environments and situations pilots might encounter during actual flight operations. Evaluation scenarios used in LOS are called line-operational evaluations (LOE) and are developed to solicit specific, observable, and measurable behaviors from pilots based on the training and evaluation data collected through each air carrier's AQP and other data-driven programs. One of the key components of AQP is the calibration of evaluators in observing and grading the CRM behaviors that are required of the specific LOE. AQP requires that rater reliability training be provided to evaluators to ensure the AQP data remain reliable and valid (Air Carrier Operations Branch, 2017).

Statement of the Problem

Aviation universities are important in training future air carrier pilots. Flight training organizations called pilot schools are certificated by the Federal Aviation Administration (FAA) under Title 14 of the Code of Federal Regulations (CFR), Part 141. Training at university pilot schools includes extensive academic study paired with strict flight training guidelines. The strict flight training guidelines are based on the combination of a minimum number of flight hours and specific training course outlines (TCO), at the completion of which proficiency must be demonstrated to an FAA-

authorized evaluator or check instructor. Check instructors are designated by pilot school chief instructors and approved by the FAA to conduct end-of-course (EOC) proficiency tests. Proficiency is judged by the check instructor using the FAA's Airman Certification Standards (ACS) or Practical Test Standards (PTS).

However, unlike under AQP, there is no formalized standardization program approved by the FAA that teaches check instructors how to observe and judge piloting proficiency. Individual pilot schools may have developed their own standardization programs. However, it is unknown if they possess the data collection and validation processes to ensure evaluations are being done in a consistent and reliable manner. In the absence of a formal standardization program, individual check instructors may rely on their own experiences and biases when conducting their evaluations, especially when they are required to make decisions about how to interpret or apply a specific proficiency standard listed in the relevant ACS or PTS.

Furthermore, specific training methods used for pilot school check instructor training programs may not be sufficiently planned or detailed for maximum effectiveness. Most aviation instruction follows the general guidelines of the FAA's Aviation Instructor's Handbook. However, these guidelines are written for the purpose of achieving individual learner outcomes (Airman Testing Standards Branch, 2020). The guidelines do not address the goals and outcomes an organization may seek to achieve in improving the output of its workforce.

Purpose Statement

The purpose of this study was to determine applicability of the concepts of AQP to pilot school evaluation activities. Although air carrier flight training and evaluation are

mostly done in qualified, full flight simulators (FFS), pilot school flight training and evaluation are primarily accomplished during actual flight operations with one exception—a significant percentage of flight training and evaluation for the purpose of obtaining an instrument airplane rating is done in qualified flight training devices (FTD). The check instructor calibration study this paper details was built around the evaluation activities for the instrument airplane rating conducted in qualified FTDs. To enhance the effectiveness of check instructor calibration, appropriate application of learning and teaching theories was necessary. This paper also details the learning facilitation principles, adult teaching methods, and human resource development (HRD) methods that were used in an effort to yield such enhancement.

Significance of the Study

The results of this study are important in advancing flight training and evaluation methods at pilot schools, especially those that hold examining authority. Examining authority may be granted by the FAA to pilot schools that establish and maintain enough activity and quality of training, as measured by practical test pass rate, that they can conduct their own certification proficiency tests using their own check instructors. Pilot schools without examining authority must not certificate their graduates and can only recommend them to take and pass practical tests conducted by FAA designated pilot examiners. Large flight training organizations with teams of flight instructors and check instructors may benefit from the study through increased standardization, improved human resource development, and implementation of data-collections streams. In general, any flight training organization may find the results of this study useful in making nominal improvements to their processes and procedures.

Research Question

This study attempted to answer the following research question: Can pilot school check instructors be calibrated against a gold standard to increase reliability and accuracy in their evaluations? To answer that question, three prerequisites had to be in place, which are also detailed in this paper. Those prerequisites were: (a) a competency and behavioral indicator system based on the requirements of the ACS that translates to performance levels for various maneuver segments or event sets, (b) a grading system that allows evaluators to score piloting proficiency on a scale rather than as just pass or fail, and (c) development of a gold standard against which check instructors will be calibrated.

Delimitations

This study purposefully limited the number of participants. Limiting the number of participants improved ease of access to them because actual, on-the-job training took place. However, data analysis is based upon the number of maneuver segments and individual graded tasks that were scored by each check instructor. Therefore, the number of check instructors was less critical than the number of maneuver segments and graded tasks included during the calibration sessions and data analysis processes.

This study was also limited to only the evaluation activity that takes place in FAA-qualified FTDs involving pilots applying for an instrument airplane rating. Because digital video recordings of simulated EOC tests were needed for the calibration session, this limitation was necessary to set up recording equipment on a semi-permanent basis, control the environment within which the recordings were produced, eliminate variables associated with actual flight operations or actual evaluation activities, and to be able to

re-record as necessary to ensure having enough and quality recordings for use during the calibration sessions. This limitation was intended to increase internal validity of the study.

Limitations and Assumptions

Lack of random sampling of participants was the primary limitation of this study. Participants identified for the calibration sessions were chosen from an already established team of pilot school check instructors. The team of check instructors had already received highly standardized training relating to pilot evaluation, which may have been a factor limiting the significance of any behavioral changes. Additional volunteers involved during the video recording process were randomly chosen from a larger group of pilots, but that group was limited geographically and demographically to pilots at one pilot school. These limitations may have caused difficulty in recording a range of piloting proficiency levels and may have limited realism and external validity.

Summary

Both air carrier flight training and pilot school flight training follow strictly prescribed standards for curricula and pilot proficiency. However, using AQP, air carrier flight training additionally benefits from a calibrated instructional and evaluator workforce. While large, university pilot schools may have training and standardization programs in place for their instructors, the concept of evaluator calibration like under AQP may further improve individual and organization performance. This study attempted to answer the question about whether pilot school check instructors can successfully be calibrated against a gold standard. This paper details the process about how such calibration took place.

Definitions of Terms

Accuracy	The difference between the score awarded by a check instructor and the gold standard.
Agreement	A score awarded by a check instructor that matches the gold standard score.
Andragogy	Self-directed or facilitated learning (Knowles et al., 2012)
Behavioral indicator	An action or a statement performed or made by a pilot that indicates how a job is being handled (International Civil Aviation Organization [ICAO], 2013)
Calibration	The process of increasing check instructor accuracy, agreement, reliability, or sensitivity.
Check instructor	An evaluator designated by the chief flight instructor of a pilot school and approved by the FAA to conduct EOC tests.
Check ride	A practical test conducted by an FAA designated pilot examiner or an EOC test conducted by a pilot school check instructor.
Competency	A combination of knowledge, skills, and attitudes required to perform a complex task to a specified standard (ICAO, 2013).

Core competency	Groups of related behavioral indicators that describe how to proficiently perform a job (ICAO, 2013).
End-of-course test	A check ride conducted at a pilot school for the purpose of assessing piloting proficiency and determining eligibility for graduation from an approved course of training.
Evidence-based training	Training for and the assessment of the competencies that lead to successful completion of a task (International Air Transport Association, 2013)
Examining authority	A pilot school authorized by the FAA to recommend a graduate from an approved course of training for FAA certification without further practical testing.
Flight training device	A stationary flight simulation training device qualified by the FAA for the purpose of flight training and testing.
Gold standard	The true performance of a pilot as determined by a group of expert evaluators.
Guided discussion	An instructor-controlled learning and teaching process that places the instructor in the role of a facilitator (Airman Testing Standards Branch,

2020) and requires the instructor to carefully guide learners toward the learning objectives (Department of the Air Force, 2003).

Maneuver segment	An ACS task or group of tasks that are normally performed as part of an EOC testing scenario.
Pilot school	A flight training organization certificated by the FAA to conduct pilot training and testing in accordance with 14 CFR Part 141.
Practical test	A check ride conducted at an FAA designated pilot examiner for the purpose of assessing piloting proficiency and determining eligibility for FAA certification.
Proficiency	Performance of an element or a task within the standards prescribed by the ACS.
Reliability	The ability of a check instructor to agree with the gold standard by more than chance alone.
Sensitivity	The ability of a check instructor to identify changes in piloting proficiency.
Standardization	The process of training pilot school check instructors to perform EOC tests and other job functions in a similar matter.

List of Acronyms

ACS	Airman certification standards
ADM	Aeronautical decision making
AQP	Advanced qualification program
CFR	Code of Federal Regulations
CRM	Crew resource management
EBT	Evidence-based training
EOC	End-of-course
FAA	Federal Aviation Administration
FFS	Full flight simulator
FTD	Flight training device
HOTS	Higher order thinking skills
HRD	Human resource development
IATA	International Air Transport Association
ICAO	International Civil Aviation Organization
IRA	Inter-rater agreement
IRR	Inter-rater reliability
KSAs	Knowledge, skills, and attitudes
LLC	Line/LOS checklist
LOS	Line-operational simulation
LOE	Line-operational evaluation
PTS	Practical test standards
RRA	Referent-rater agreement

RRR	Referent-rater reliability
SBT	Scenario-based training
SMAD	Standardized mean absolute difference
SRM	Single-pilot resource management
TCO	Training course outline

Chapter II: Review of the Relevant Literature

A review of relevant literature revealed significant amounts of research regarding the training, standardization, and calibration of air carrier evaluators. However, similar research is limited for pilot school check instructors. Literature related to learning and teaching methods that may be applicable to the structure of evaluator calibration was also completed. Specific focus was placed on adult learning and teaching methods and human resource development.

Referents and Gold Standards

Much of the research done regarding air carrier evaluator calibration centered around inter-rater reliability (IRR) and IRR training. IRR is used to analyze the consistency of an evaluator across items and to analyze the agreement between evaluators (Holt et al., 1997). However, there is a potential pitfall in calibrating individual evaluators to the group. That is, if the group is wrong, then the individual evaluator could unintentionally be calibrated to the wrong referent (Holt et al., 1997). Training to the wrong referent can be avoided by using what is known as gold standards training. Gold standards training uses an external referent as the basis of comparison for individual evaluators. Gold standards training is possible in aviation because clearly defined standards of performance have already been established for a majority of the skills and behaviors that pilots must possess, but the downside is the significant amount of work that must be done in developing the gold standard (Holt et al., 1997). Despite the amount of work necessary, gold standards training is believed to be the most suitable method of calibration for evaluators because it accomplishes the training using the desirable

characteristics of well-studied frame-of-reference training and lesser-studied behavioral observation training (Baker, 2002).

Baker and Dismukes (2002) summarized the overall process for developing gold standards training. Gold standards training under AQP involves the creation or evaluation of LOE such that the overall scenario incorporates several event sets designed to solicit specific, observable CRM behaviors from the flight crew. These behaviors translate to overall performance ratings. LOE worksheets are designed to aid evaluators making the translation and enhance the debriefing and feedback provided to flight crews (Holt et al., 2002). Event sets are triggered by a condition, such as an abnormal indication on the flight deck, which sets the event into motion and requires the flight crew to work through the event in real time to its logical conclusion. The LOE are recorded on video so they can be presented to evaluators-in-training to practice observing and grading (Baker & Dismukes, 2002).

Calibration Studies

Only two examples of AQP-like training at pilot schools or equivalent flight training organizations were found in the literature. The United Arab Emirates (UAE) instituted an ab-initio pilot training program based on AQP and the FAA's then-active FAA-industry training standards. However, the program did not address calibration of its evaluators (Al-Romaithi, 2006). Western Michigan University attempted gold standards maneuver training and calibration with a group of its flight instructors, but not check instructors. However, the study used a pretest-posttest design in which the flight instructors were shown the exact same set of videos after receiving calibration training as those shown before (Beaudin-Seiler & Seiler, 2015). The results of the study may have

been more informative had the posttest videos not been exactly the same so as to minimize testing effects and counterbalance the within-subjects design.

The limitations of these studies may highlight a challenge to researchers in that a pilot's early flight training rests in the development of technical knowledge and maneuvering skills rather than the complex behavioral skills of CRM that are translatable to overall performance levels. Therefore, there may be limited ability to create enough varying scenarios and event sets that can be used to avoid testing effects. However, the ACS, a currently evolving replacement to the PTS, places more emphasis on decision-making and risk-management skills. These skills are subsets of the concept of single pilot resource management (SRM). SRM and CRM share these and other similar skill subsets. Furthermore, the majority of CRM behavioral indicator systems in use at air carriers are based on the well-known Line/LOS Checklist (LLC) (Flin & Martin, 2001). In addition, O'Connor and Long (2011) were successful at adapting the LLC and other systems in order to create a prototype system used in the training of U.S. Navy officers. Consequently, perhaps there is now the ability to define behavioral indicators for maneuvers-based flight training using the standards set forth in the ACS in connection with previously developed systems.

Statistical Measures

Multiple methods to measure reliability, accuracy, and agreement among evaluators under AQP have been studied. One or more of these measures may be usable for pilot school check instructor calibration. Goldsmith and Johnson (2002) described using a Pearson product-moment correlation to determine IRR and referent-rater reliability (RRR). The RRR was of particular importance because it was a measure of an

evaluator's sensitivity to true changes in performance as noted by the referent, or gold standard, although both measures in combination revealed more information than either alone (Goldsmith & Johnson, 2002). Goldsmith and Johnson (2002) also described using a standardized mean absolute deviation (SMAD) coefficient to measure the accuracy of the rater against the gold standard. At Western Michigan University, researchers used a Cohen's Kappa coefficient to measure pretest and posttest levels of agreement among its flight instructors (Beaudin-Seiler & Seiler, 2015). Mulqueen et al. (2002) described using a multifacet 1-parameter item response theory model, or Rasch model, to analyze evaluator leniency or severity in grading, the complexity of grade sheets, the skill of flight crews, and the interaction among all these variables.

While calibrating evaluators and improving IRR is a worthwhile goal unto itself, understanding the complexities and reasons why IRR may be poor prior to calibration or why an attempt at calibration may not yield desired results is just as important. Gontar and Hoermann (2015) explained four themes that influence IRR: (a) target- or pilot-related influences such as level of experience compared to that of the evaluator, (b) scenario- and task-related influences involving the interaction of both pilots in a two-pilot crew, (c) measurement-related influences based on the grading system used, and (d) rater-related influences such as personal interpretations and motivation of the evaluator.

Gontar and Hoermann's (2015) third theme, measurement-related influences, was of critical concern to this study because of the current method of evaluating pilot performance as either pass or fail. Pass-fail grading mechanisms showed lower agreement among evaluators relative to comparable four- or five-point scales. Therefore, use of an appropriate grading system was necessary to show changes in IRR following calibration.

Gontar and Hoermann's fourth theme, rater-related influences, was important to consider as well, because evaluator experience levels varied. Kim et al. (2015) showed that when dental students were tasked to score their peers, significant differences were found between third year students, fourth year students, and faculty regarding the scores awarded—higher scores were awarded by less experienced evaluators. Similar variances of pilot school check instructor experience levels exist.

Competencies

Discussion of behavioral indicators enters the realm of evidence-based training (EBT). A greater evolution of the ACS may be to derive evidence-based standards that detail the behavioral indicators necessary to demonstrate proficiency. The International Air Transport Association (IATA) (2013) explains that the basis of EBT is the assessment of competencies that lead to completion of a task rather than measurement of the task outcomes alone. The International Civil Aviation Organization (ICAO) (2013) defines competency as a combination of knowledge, skills, and attitudes (KSAs) required to perform a complex task to a specified standard. Because the FAA ACS already contain what could be considered KSAs for each task, evolving or distilling them into a fundamental set of competencies may be possible. ICAO (2013) further explains that core competencies are groups of related behavioral indicators that describe how to proficiently perform a job and that a behavioral indicator is an action or statement performed or made by a pilot that indicates how the job is being handled.

Although developing a specific set of core competencies is necessary for a true, AQP-like approach to training and evaluation, doing so would require a significant amount of work involving a job-task analysis specific to the organization and that

correlates to the ACS. That work was beyond the scope of this study. Instead, a basic list of already-developed competencies, along with their behavioral indicators, learning levels, and performance levels sufficed. Embry-Riddle Aeronautical University (ERAU) in Daytona Beach, Florida developed and is implementing such a list of competencies along with a 5-point grading system designed to measure progress toward achieving the standards set forth in the FAA ACS. Because ERAU's primary goal is flight education and training, the behavioral indicators used are evidentiary measures of achieving various learning levels, indirectly correlating to the outcome-based standards in the FAA ACS (ERAU, 2021).

Learning Theories Applicable to Calibration

To be most effective, the process of calibrating pilot school check instructors should rely on theories of learning and teaching. Pilot school check instructors should already possess a background in evaluation, assessment, and critique of learner pilots. It is important for the person who facilitates the calibration to use appropriate learning theories and teaching methods as applied in aviation.

Since 1885, numerous literary works proposed different learning theories and many researchers interpreted those works differently, making organization of such theories difficult (Knowles et al., 2012). However, it is generally accepted to loosely categorize the many theories into two groups—behaviorism and cognitive learning theories. Of the two categories, cognitive learning theories are generally accepted as the most effective in aviation education and flight training and are widely used. Of specific importance to evaluator calibration is scenario-based training (SBT), which is based on active interaction with the environment (Airman Testing Standards Branch, 2020).

Although cognitive theories are most used in aviation education and flight training, use of behaviorism theories is applicable, and combining the two yields the most thorough results (Airman Testing Standards Branch, 2020). Behaviorism centers on the idea that specific behaviors can be observed and measured in response to environmental or external stimuli (Airman Testing Standards Branch, 2020). The Department of the Air Force (2003) explains to its classroom instructors that, “We need to realize the importance of controlling learning experiences by manipulating the classroom environment (stimuli) which gives our students a chance to behave or perform (respond) in the way we desire” (p. 24). Applying this realization to check instructor calibration, the stimuli is the calibration training that the check instructors receive and the change in behavior is measured from before the training to after the training. Behavior, in this context, should not be confused with the behavioral indicators the check instructors are ultimately calibrated to identify when evaluating piloting proficiency.

The three domains of learning are the cognitive, affective, and psychomotor domains. Bloom’s taxonomy of the cognitive domain is the basis for many aviation and flight instruction methods. The taxonomy includes six major classes—knowledge, comprehension, application, analysis, synthesis, and evaluation (Bloom et al., 1956). While 90% of what is taught in Air Force schools is appropriately in the lower three levels of the cognitive domain (Department of the Air Force, 2003), successful piloting abilities rely on the higher three levels of the cognitive domain to form what are called higher order thinking skills (HOTS) and are the basis of aeronautical decision making (ADM) (Airman Testing Standards Branch, 2020). However, even within the knowledge level of the cognitive domain, behaviors expected of the learner progress from specific to

abstract (Bloom et al., 1956). The aviation industry focuses on SBT as the primary method of developing learner pilots' HOTS and in turn ADM skills.

The idea that much of cognitive learning involves the lower levels of the cognitive domain seems at odds with the concept of HOTS, ADM, and the SBT required of aviation training programs. The same could be true of evaluator training. While evaluators assess complex educational objectives and abstract behaviors, limited real-world experience or practice during evaluator training is available to new check instructors. Instead, their evaluative abilities are assumed to be satisfactory because of their experience as flight instructors. By providing SBT to check instructors, improved evaluative abilities may be possible. The intent of the check instructor calibration sessions, in essence SBT, is to help the check instructors identify the behaviors that yield the outcomes required in the ACS.

As explained by the Airman Testing Standards Branch (2020), the three domains of learning translate, like ICAO core competencies, to the KSAs necessary for pilot training and certification. The various FAA ACS further translate attitudes to risk management skills and, as a result, contain a complete set of standards, or proficiency elements, covering each domain of learning. However, those standards are written in the form of outcomes. Check instructor calibration is designed to train and calibrate check instructors toward identifying the relationships between the domains of learning and KSAs, thereby allowing the evaluation of the observable learning behaviors that lead to the performance outcomes stated in the ACS.

Teaching Methods Applicable to Calibration

In addition to understanding what and how evaluators will learn during calibration training, the use of appropriate teaching methods to enhance their learning is also important. A dramatization is a teaching presentation method that involves indirect discourse that is seen and heard by the learners (Department of the Air Force, 2003). Dramatization is the primary tool in employing SBT during evaluator calibration. The dramatizations are check ride maneuver segments (analogous to event sets in AQP) that are pre-recorded in video format. Both the Department of the Air Force (2003) and the Airman Testing Standards Branch (2020) explain that video aids should be used only to supplement what the instructor presents and teaches, and that the duration of the videos should be kept short. However, the videos used for evaluator calibration contain the scenarios by which SBT takes place, so they serve more of a primary role and should be longer in duration. In addition to the videos, teaching lecture and guided discussion methods are used during evaluator calibration.

The teaching lecture is a form of lecture that allows some active participation by the learners but otherwise is used primarily for the instructor to convey general understanding of a topic (Airman Testing Standards Branch, 2020). The amount of learner participation involved as well as the class size can differentiate the teaching lecture either as a formal lecture (no or very little participation) or an informal lecture (greater participation) (Department of the Air Force, 2003). Less structured than a teaching lecture is the guided discussion. The guided discussion is an instructor-controlled process that places the instructor in the role of a facilitator (Airman Testing Standards Branch, 2020) and requires the instructor to carefully guide the learners toward

the learning objectives (Department of the Air Force, 2003). The teaching lecture, whether formal or informal, naturally sets the stage for a guided discussion if the learners do not already possess requisite knowledge.

SBT is a learner-centered approach that uses constructivism learning theory as its basis (Airman Testing Standards Branch, 2020). Several of the principles of constructivism, such as learner ownership of the learning process, experiential learning, and problem-based learning, share those with andragogy (Knowles et al., 2012). Andragogy is simply described as self-directed or facilitated learning (Knowles et al., 2012). According to IATA (2013), facilitation is a key instructional framework that an EBT instructor should follow. Although normally associated with adult education, andragogy becomes increasingly appropriate over pedagogy beginning at a very young age and especially during adolescence (Knowles et al., 2012). There are six principles of andragogy, which are: (a) learners must have a need to know something, (b) learners feel responsible for their own learning, (c) learners' range of experiences affect how they learn, (d) learners are ready to learn only if the material can be applied in life situations, (e) learners are task and problem oriented, and (f) learners source motivation internally (Knowles et al., 2012).

Based on the principles of andragogy, a process for teaching adults was developed. The process includes seven steps. The steps are: (a) set a cooperative learning environment; (b) create mechanisms for mutual planning; (c) diagnose learner needs and interests; (d) structure learner objectives around learner needs and interests; (e) design sequential activities for achieving the objectives; (f) conduct the activities by selecting appropriate methods, materials, and resources; and (g) evaluate the quality of the learning

experience and re-diagnose additional learner needs (Carlson, 1989). The fundamentals of aviation instruction as detailed by the Airman Testing Standards Branch (2020) incorporate all the principles of andragogy and generally follow the adult teaching process. Like flight instruction, check instructor calibration should follow the same process.

Human Resource Development

Specific adult learning methods are applicable to evaluator calibration. The first method is human resource development (HRD), which primarily focuses on performance improvement. The process and methods used to achieve the performance improvement balance organizational control and needs with individual control and needs (Knowles et al., 2012). Because the goal of evaluator calibration is improving the performance of both the individual and the organization, HRD seems to be a particularly important method. HRD places individual performance improvement within the context of and in agreement with organizational performance improvement. HRD also provides a data stream to the organization about individual and team performance. AQP is a form of HRD.

The second method is Pratt's Four-Quadrant Model, which was developed to try to show and explain the variability in adults' readiness to learn (Knowles et al., 2012). Readiness to learn centers around the life situations adults face that create the need for learning, but these situations expose adults' level of competence, commitment, and confidence and therefore create variance in adults' required level of direction and support (Knowles et al., 2012). Questionnaires or surveys can serve as the basis for applying Pratt's Four-Quadrant Model. Such tools can assist with tailoring the teaching lecture and guided discussion that is part of check instructor calibration.

A third method is the Whole-Part-Whole Learning Model (WPW). Knowles et al. (2012) explain that the WPW Learning Model is useful because it can be adapted to learning experiences of varying length including very short experiences, it is simple enough for learners to use on their own, subject matter experts are not required to have a deep understanding of learning theory to share knowledge, and it is a practical tool for education professionals. The calibration session loosely follows the WPW Learning Model. The overview at the beginning of a calibration session (whole) is supported by the specific learning activities during the videos and guided discussions (parts) that in turn are drawn together during the debriefing and conclusion of the session (whole).

Debriefing is an important part of skill development and performance improvement in aviation. The Airman Testing Standards Branch (2020) suggests that debriefing should happen after flight training events but provides little guidance on how to structure the debriefing. Instead, the Airman Testing Standards Branch (2020) focuses on the interaction between the instructor and the learner by describing various forms of critique following an evaluation or assessment of a learner pilot and emphasizes that critiques should be used to enhance learner-centered training. Critique is perhaps a component of a debriefing event, and so a structure for the debriefing is necessary. Gardner (2013) summarized the process, goal, and a tool for debriefing as applied in simulation-based medical education. The process involves three steps—reaction, understanding, and summary. The goal is to use results to work backward in uncovering actions and frames of mind of the person being evaluated. A common tool is the plus-delta tool, which categorizes the events of the lesson or situation into what specifically went well and what specifically should change to improve during the next lesson or

training period. The debriefing portion is an important part of the calibration session and should be much more involved than simply summarizing the day's activities. Specific focus on the method, structure, and process of the debrief may help to solidify the concepts of calibration and gold standards training to ensure maximum effectiveness.

Gaps in the Literature

As previously described, only two examples of AQP-like training at pilot schools or equivalent flight training organizations were found in the literature. In one case, the calibration of evaluators was not addressed (Al-Romaithi, 2006). In the other, calibration of flight instructors, not check instructors was conducted (Beaudin-Seiler & Seiler, 2015).

In addition to the lack of research on check instructor calibration, no guidance was found on how to develop a competency system for use in primary flight training and evaluation. However, the Instrument Rating Airplane ACS mentions the words competency or competence several times in relation to the SRM and CRM behaviors that are similar to the core competencies of AQP and explains that evaluation of SRM and CRM may be subjective in nature (Airman Testing Standards Branch, 2019).

This study attempts to contribute to the body of literature by detailing the calibration of pilot school check instructors that other studies did not. In addition, explanation and use of an already-developed basic competency system may invoke other researchers' desire to propagate similar research and further develop such a system for broader use.

Theoretical Framework

Based on the literature, the use of evaluator calibration methods as a means of improving IRR and RRR for evaluation of pilot performance has been shown to be

effective. The same techniques should be translatable to pilot school check instructor calibration. However, lack of awareness and experience using competencies as evidence of satisfactory piloting performance jeopardizes the success of pilot school check instructor calibration. To substitute for this lack of awareness and experience, the use of learning theories and teaching methods, as applied in aviation, in delivering the calibration training further guided the development and execution of this study.

Research Model

Four statistical measures were determined for this study: (a) IRR, (b) RRR, (c) Pearson product-moment correlation, and (d) SMAD. The most important of these was RRR and SMAD because they measured the reliability and accuracy of each check instructor against the gold standard. As used in evaluator calibration, these measurements are within-subjects measurements. Changes in these measurements showed the level of effectiveness of the calibration. Therefore, a within-subjects, pretest-posttest design was used as the basis for the design and data collection. Naturally, a more carefully constructed and delivered calibration training should yield a greater change in these four measurements. Therefore, the development of the training and calibration around proven instructional and human resource development methods was important.

Summary

The literature regarding AQP, gold standards training, and previous pilot school check instructor calibration attempts gives the appropriate background and considerations for developing a training program that may be effective in calibrating pilot school check instructors against a gold standard. Doing so involved the use of competencies, behavioral indicators, and a grading system that evidence proficiency and allowed pilot

school check instructors to discriminately grade different levels of proficiency. A calibration session rooted in appropriate learning theories and teaching methods, with focus on facilitation, was designed to make the session more meaningful and effective for the check instructors involved. The learner-centered focus of SBT, informal lecture, and guided discussion teaching methods are consistent with andragogy and theories of adult learning and teaching. Additional consideration of HRD allowed the calibration session to have meaningfulness at an organizational level.

Chapter III: Methodology

Research Method Selection

The following research question guided this study: Can pilot school check instructors be calibrated against a gold standard to increase reliability and accuracy in their evaluations? To answer the research question, check instructors received training on gold standards, calibration, grading scales, and competencies. The study was designed to measure changes in the check instructors' grading from before the calibration training to after the calibration training. Measurements analyzed were raw agreement percentage, IRR and RRR using a Cohen's kappa statistic or kappa-like statistic, grading sensitivity using a Pearson product-moment correlation coefficient, and grading accuracy using a SMAD coefficient. This general procedure supported the selection of a within-subject design that used a pretest-posttest analysis. The analysis was based on the scores the check instructors awarded for each graded task associated with several pre-recorded check ride maneuver segments, which are part of instrument airplane end-of-course tests.

Population/Sample

The target population was the group of evaluators and check instructors at any Part 141 pilot school that have been granted examining authority by the FAA. The check instructors at these schools range in age, gender, ethnicity, and aviation experience.

The sample population was the team of check instructors at one university pilot school. The sample of check instructors ranged in age, gender, ethnicity, and aviation experience similar to the target population.

Population and Sampling Frame

Specifically, participants were FAA-designated check instructors from Embry-Riddle Aeronautical University (ERAU) in Daytona Beach, Florida. The pilot school has one team of check instructors whose full-time job is evaluation of student and flight instructor piloting proficiency. The team normally consists of 15 check instructors. Ten check instructors were chosen from this team. Experience in evaluation duties varied among the check instructors. Experience ranged from several months to several years. Overall flying experience also varied. Flying experience ranged from a few years to more than a decade. Results of the study should be generalizable to any other sample or the larger population of pilot school check instructors.

Sample Size

Statistical power was a function of the number of maneuver segments and individual graded tasks rather than the number of check instructors. Based on Beaudin-Seiler and Seiler (2015), a Cohen's kappa statistic of .3 for the initial level rater reliability was expected prior to calibration taking place. Bujang and Baharum (2017) explained how to determine the minimum sample size of graded items when using Cohen's kappa as a measure of rater reliability, in this case when each item was graded on a 5-point scale. For a 5 x 5 pairwise contingency table with an equal number of agreements in each category, an increase in the Cohen's kappa statistic to .7, and a statistical power of 80% at $\alpha = .05$, 18 ratings are necessary. Therefore, 18 maneuver segments were created to achieve this statistical power.

However, an overall agreement and reliability measurement for each maneuver segment was not determined. Instead, agreement and reliability were based on individual

graded tasks. Furthermore, pairwise contingency tables for the 5-point grading scale were not likely to have equal marginal frequencies, so, it was necessary to have a larger number of the individual graded items (Bujang & Baharum, 2017).

The Instrument Rating Airplane ACS contains 12 tasks that were appropriate for use in this study. Specificity was added to the tasks to create eight additional tasks. For example, the task titled Non-Precision Instrument Approach was turned into six tasks by specifying the type of navigational aid and transition to the final approach course, such as Non-Precision Instrument Approach (VOR with Procedure Turn). The 20 resulting tasks were arranged in various combinations to form the basis of each of the scenarios that the pre-recorded check ride maneuver segments depicted. Each maneuver segment contained from one to four individual graded tasks. The resulting arrangement provided for 58 individual grading opportunities for each check instructor.

For a 5 x 5 pairwise contingency table with 80% of the agreements in one category and 5% of the agreements in each of the other four categories, an increase in the Cohen's kappa statistic from .3 to .7, and a statistical power of 80% at $\alpha = .05$, 50 ratings are necessary (Bujang & Baharum, 2017), so the 58 individual grading opportunities in this study exceeded the minimum required.

Sampling Strategy

A nonprobability, convenience sampling strategy was used. This strategy improved ease of access to the participants because actual, on-the-job training took place.

Confederates

To prevent the participants from learning about the pre-recorded maneuver segments in advance, one flight standards evaluator was selected to help create the

recordings and three additional expert evaluators were used to determine the gold standard scores for each task on each maneuver segment.

Nine different pilots were chosen whose performance was recorded. The pilots were selected from among volunteers in the pilot population ERAU in Daytona Beach, Florida. Volunteers were required to be within four modules of beginning the instrument airplane EOC or to have recently completed the instrument airplane EOC within approximately one month prior to the recording taking place.

Calibration Facilitator

The calibration facilitator was the assistant chief flight instructor and manager of flight standards at ERAU in Daytona Beach, who had been in that role for over 9 years. The calibration facilitator was designated by the FAA as being responsible, under the direction of the chief flight instructor, for the proficiency testing and designation of the pilot school's team of check instructors. The calibration facilitator had over 20 years of flying experience, over 16 years of professional flight instruction experience, and over 13 years of evaluating experience as a pilot school check instructor.

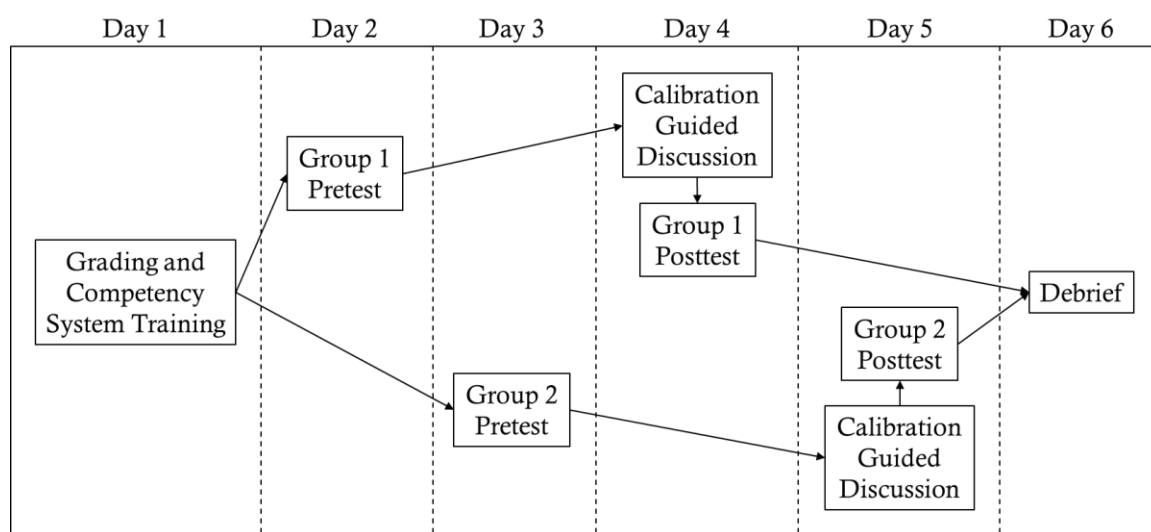
Data Collection Process

The data collection process involved a within-subjects, pretest-posttest design. Procedural elements involved the determination of gold standard scores, participant in-briefing and questionnaire completion, a grading system training session, and the check instructor calibration session. The participant in-briefing and questionnaire completion along with the grading system training was presented to the entire group of 10 check instructors on Day 1. The check instructors were then split into two groups of five check instructors. The calibration session was delivered twice—once to each of the two smaller

groups of check instructors. Each of the check instructor calibration sessions occurred across 3 days—the first group on Day 2, Day 3, and Day 6 and the second group on Day 4, Day 5, and Day 6. All 10 check instructors met on Day 6 to discuss the final results of the calibration sessions. The overall calibration procedure is shown in Figure 1.

Figure 1

Calibration Procedure



Design and Procedures

This was a causal study that used a within-subjects, pretest-posttest design. The 10 check instructor participants were split into two groups of five so the calibration facilitator could present the pre-recorded maneuver segment videos in different combinations to counterbalance the within-subjects design. The videos were grouped in pools to help organize the process. For example, Pool C then Pool B followed by Pool A then Pool D were presented to the first group of check instructors, but Pool D then Pool A followed by Pool B then Pool C was presented to the second group of check instructors. For both groups of check instructors, videos were presented in random order within each

pool. The videos were individually numbered within each pool. Videos having the same number were not repeated in consecutive pools, and each pool was used only once for each group of check instructors.

Determination of Gold Standard Scores. The three expert evaluators selected to determine the gold standard scores watched each of the pre-recorded maneuver segments, discussed each pilot's performance during each maneuver segment, and come to a unanimous agreement for the score of each task of the maneuver segment. Those scores were used as the gold standard. Failure to achieve unanimous agreement would have required re-recording of the maneuver segment and re-evaluation by the expert evaluators until unanimous agreement was reached. The expert evaluators were able to reach agreement for each task, so re-recording and re-evaluation was not necessary. However, the three expert evaluators did require multiple viewing attempts and some extra time to deliberate and agree on the score for some of the tasks.

In order for the calibration facilitator to have the necessary information to support the guided discussion during the calibration sessions, the three expert evaluators prepared a written justification for the score awarded to each task of each maneuver segment. The justification for each score was broken down by competency and detailed the behavioral indicators or levels demonstrated by the pilot in the video. In the case of unsatisfactory performance by the pilot, the justification document also included the ACS code representing the proficiency element in the ACS that was below standard.

Participant In-Briefing and Questionnaire Completion. On Day 1, each participant completed an informed consent and answered a participant questionnaire to capture background information regarding flying and evaluating experience. The

information was used by the calibration facilitator to tailor guided discussions during the calibration session, following Pratt's Four-Quadrant Model of HRD. The calibration facilitator briefed all the participants simultaneously. The calibration facilitator described the general purpose of and the procedure used in the study. Using a teaching lecture method supported by a brief electronic presentation, the calibration facilitator explained the background information regarding AQP and evaluator calibration. To minimize rater bias, the measurements and expected results of the study were not shared. The calibration facilitator instructed the participants not to discuss the recordings or help each other score the tasks during the calibration session.

Grading System Training Session. Before the calibration session took place, it was necessary to train the check instructors on the grading system. The check instructors were familiar with evaluating and grading piloting proficiency in comparison to the Instrument Rating Airplane ACS using a binary pass or fail grade. The check instructors were not familiar with grading on the 5-point scale that was used during calibration. The grading system training was delivered using a guided discussion teaching method supported by a Microsoft PowerPoint presentation. A copy of the presentation is included in Appendix D. The grading system training session occurred on Day 1 and lasted approximately 60 minutes.

Check Instructor Calibration Session. The calibration facilitator began the calibration session on Day 2 for the first group of five check instructors by selecting nine pre-recorded maneuver segments for viewing from one of the four pools of videos. The segments each featured a different pilot. Each of the check instructors scored the tasks for each maneuver segment using the provided maneuver evaluation grade sheet, basing their

scores on the Basic Competencies and Behavioral Indicators in Appendix B. After a short break, the calibration facilitator selected another nine pre-recorded maneuver segments for viewing from a different pool of videos but that did not have the same video numbers. The segments may or may not have featured the same pilots as in the first pool of videos, but each video did have a different maneuver segment than in the first pool. Each of the check instructors again scored the tasks for each maneuver segment using the provided maneuver evaluation grade sheet, basing their scores on the Basic Competencies and Behavioral Indicators in Appendix B.

Between Day 2 and Day 4, the calibration facilitator conducted a statistical analysis of each of the check instructor's scores. A raw agreement percentage, Cohen's kappa coefficients for both IRR and RRR, Pearson product-moment correlation, and SMAD were determined. These values were used to compare each of the check instructor's scores to that of the group's and each of the check instructor's scores to that of the gold standard to determine the initial level of rater agreement, reliability, sensitivity, and accuracy. It was possible that the check instructors showed a high initial level of rater agreement, reliability, sensitivity, and accuracy. In that case, the check instructor calibration session would still have been used to attempt an increase in agreement, reliability, sensitivity, and accuracy.

At the beginning of Day 4, each check instructor was provided with individual feedback about his or her scores and how they compared to the group and to the gold standard. Focusing on the maneuver segments and tasks with the lowest agreement, reliability, and accuracy first, the calibration facilitator explained the gold standard for the tasks in those maneuver segments and why the group differed from the gold standard.

The gold standard justification document prepared by the three expert evaluators was used as an aid.

The calibration facilitator then facilitated a group discussion with the specific purpose of facilitating learning and emphasizing methods for more reliable and accurate observation of the associated behavioral indicators. The guided discussion focused on the use and interpretation of the Basic Competencies and Behavioral Indicators in Appendix B and their relationship to the Instrument Rating Airplane ACS proficiency elements. The guided discussion followed a cause-effect organization as the check instructors linked observed behaviors and proficiency outcomes as shown in the videos. This process was expected to be effective at helping the check instructors understand their grading in comparison to the gold standard. The guided discussion teaching method was the manipulated stimuli designed to affect a change in rater agreement, reliability, sensitivity, and accuracy. The calibration facilitator facilitated the discussion using the three-step process for effective debriefing (reaction, understanding, and summary) described by Gardner (2013).

Following the guided discussion portion of the calibration session, the calibration facilitator selected another 18 videos for viewing from the two remaining pools of videos and the check instructors again scored the tasks for each of the maneuver segments. It is important to note here that the same videos used in the beginning of the calibration session on Day 2 were not used during this portion. Instead, different videos showing the same maneuver segments were shown to limit testing effects. The pilots and their levels of proficiency may or may not have been the same as in the videos selected prior to the guided discussion portion of the calibration session.

The same procedure that was used on Day 2 and Day 4 was repeated for the second group of five check instructors on Day 3 and Day 5. However, the video pools were presented in a different combination to counterbalance the within-subjects design.

Between Day 5 and Day 6, after both groups of check completed the process, the calibration facilitator again conducted a statistical analysis of all the check instructors' scores from the the calibration sessions on Day 3 and Day 5. A raw agreement percentage, Cohen's kappa coefficients for both IRR and RRR, Pearson product-moment correlation, and SMAD were determined. These values were compared to those determined on Day 2 and Day 4 prior to the calibration session in order to determine a change in the level of rater agreement, reliability, sensitivity, and accuracy.

On Day 6, all 10 check instructors met together and the calibration facilitator facilitated a second group discussion following the same three-step process as the first. The calibration facilitator completed the calibration by drawing conclusions about the group's change in performance. The plus-delta debriefing tool described by Gardner (2013) was used in combination with Pratt's Four-Quadrant Model of HRD to focus individual calibration results within the context of possible future organization needs.

Apparatus and Materials

Pre-Recorded Check Ride Maneuver Segments. A maneuver segment was operationally defined as an ACS task or group of tasks that are normally performed as part of an EOC testing scenario. Pre-recorded check ride maneuver segments were made of instrument airplane EOC tasks. Recordings of actual EOC tests were not used. The recordings were made during fabrications of the FTD portion of the EOC test. This

contrived setting allowed greater control and mitigation of any confounding variables that might have impacted the pilots' performance.

The same flight standards evaluator was used in each recording to give a consistent EOC test. By using the same flight standards evaluator, confounding variables associated with different evaluation methods or techniques that might have impacted the pilots' performances were minimized. The flight standards evaluator created the specific scenarios used to complete each of the maneuver segments and archived the instrument approach procedures and notices to airman that were applicable at the time of the recording. Because instrument approach procedures change or are removed by the FAA on a regular basis and notices to airman change regularly, the check instructors participating in the calibration needed to be able to reference what the flight standards evaluator and pilots in the videos used to complete the maneuver segments. The scenarios, instrument approach procedures, and notices to airman are included in Appendix F.

The nine pilots each performed four different maneuver segments. A total of 36 videos were recorded, with each maneuver segment being performed twice, but by different pilots. By choosing pilots who were nearing the completion of the instrument airplane training or who recently completed the training and EOC test, authentic performances were expected and likely to mimic performances seen by the check instructors during actual EOC tests.

The recordings were organized into four pools of nine. The pools were labeled Pool A through Pool D. Pool A videos were numbered with odd numbers from Video 1 through Video 17. Pool B videos were numbered with even numbers from Video 2

through Video 18. Together, Pool A and Pool B contained 18 videos numbered from Video 1 through Video 18, with each video representing one maneuver segment. Similarly, Pool C videos were numbered with odd numbers from Video 1 through Video 17. Pool D videos were numbered with even numbers from Video 2 through Video 18. Together, Pool C and Pool D contained 18 videos numbered from Video 1 through Video 18, with each video representing one maneuver segment but each featuring a different pilot than the same video number in Pool A and Pool B. Appendix F shows how the videos were arranged. The video numbers matched the video number labels on the maneuver evaluation grade sheets to ensure the correct grade sheet was used by the check instructors to evaluate the correct video. Organizing the videos in this fashion allowed the calibration facilitator to present them in different combinations to counterbalance the lack of random sampling of the check instructors.

The FTD that was used was housed in ERAU's Advanced Flight Simulation Center. The FTD was a replica of the Cessna 172S Nav III flight deck and instrument panel and mimic all operations of the real airplane. The FTD included two pilot seats and had an open back between the flight deck and instructor operating station, which was positioned directly behind the pilot seats. The design of the FTD facilitated the use of recording equipment (e.g., microphones and video recorders).

Video and audio recording were made in a digital format. The recordings showed the pilots' manipulation of the flight controls the flight instrument indications presented to the pilots on the primary flight display, multi-function display, and standby flight instruments. The pilots typically use their laps to lay their checklist and electronic flight bag for use during flight, so the videos also showed the pilots' lap area for the check

instructors to be able to view the pilots' use of those materials. The pilots' lap area was shown in an inset in the lower left corner of each video. To avoid rater bias, the identities of the flight standards evaluator and pilots in the videos were not shown. Because visual maneuvers are not tested in the FTD portion of instrument airplane EOC tests, it was not necessary to record the pilots' visual references outside of the flight deck. A still image of one of the videos is shown in Figure 2 as an example of what all the videos looked like to the check instructors during calibration.

Figure 2

Example of Pre-Recorded Check Ride Maneuver Segment Videos



The audio portion of the recordings contained the communications between the flight standards evaluator and the pilots. These communications include the instructions, oral questions, and simulated air traffic control communications given by the flight standards evaluator to the pilots.

Maneuver Evaluation Grade Sheets. The maneuver evaluation grades sheets were designed to show an objective and completion standard for each maneuver segment similar to how objective and completion standard statements might look on a grade sheet for a real EOC test. The maneuver evaluation grade sheets also presented each task of the maneuver segment together with a 5-point grading scale that allowed the check instructors to score each task as they viewed the pre-recorded check ride maneuver segment videos. The maneuver evaluation grade sheets are included in Appendix C and the 5-point grading scale is shown in Table 1. A complete explanation of the 5-point grading scale is the Microsoft PowerPoint presentation included in Appendix D.

Table 1

Five-Point Grading Scale

Score	Meaning
Inc.	Incomplete; task not performed, attempted, demonstrated, or discussed
1	Requisite knowledge is demonstrated; deviations from the prescribed task standards occur that are not recognized or corrected
2	Deviations from the prescribed task standards can be explained but not corrected
3	Deviations from the prescribed task standards occur that are recognized and corrected
4	Performance remains within the prescribed task standards
5	Performance remains within the prescribed task standards; cognitive abilities are exemplary

Note. Adapted from (ERAU, 2021).

The Basic Competencies and Behavioral Indicators are included in Appendix B. The list of competencies is a simplified listing of the ICAO core competencies. However, the behavioral indicators are the descriptors and evidentiary examples of the various levels of learning and grading rubrics described in the FAA's Aviation Instructor's Handbook (Airman Testing Standards Branch, 2020). The competency system was developed by ERAU in Daytona Beach to support the 5-point grading scale described in Table 1. By merging competencies, learning levels, and grading rubrics, a system was created with the goal of improved analysis of learner pilot training progression from start to finish, including the final EOC test (ERAU, 2021).

Lesson plan for check instructor calibration. A lesson plan was used to aid the calibration facilitator during the calibration session. The lesson plan detailed specific teaching methods, organizational patterns, references, module durations, learning objectives, and associated samples of behavior expected of the check instructors. The lesson plan also included calibration facilitator and check instructor actions. Of importance is the strategy statement that assisted the calibration facilitator in delivery of the lesson. A lesson introduction section with specific attention, motivation, and overview guidance as well as a lesson conclusion section with specific summary, remotivation, and closure guidance was included. The Department of the Air Force (2003) provided guidance on lesson plan formats, which was followed for the development of the calibration lesson plan. Appendix E contains the Lesson Plan for Check Instructor Calibration.

Sources of the Data

The primary source of data were the scores recorded for each task by the check instructors on the maneuver evaluation grade sheets. Secondary sources of data were the gold standard scores agreed upon by the three expert evaluators prior to the calibration session and the demographic data collected about the check instructors on the participant questionnaire.

Ethical Considerations

Each check instructor participant was presented with and required to complete an informed consent prior to participation. The check instructors' participation was voluntary and confidential. Names or other identifying information were not asked for, collected, or recorded. Demographic data collected on the participant questionnaire was not associated with the individual completing it. Check instructor participants were not exposed to any harm or adverse conditions. The setting for the calibration was a typical classroom or conference room setting that the check instructors regularly use for their normal day-to-day work and educational activities. The ERAU Institutional Review Board granted approval for the study, which is included in Appendix A.

Measurement Instrument

Variables and Scales

The measurement instrument was the maneuver evaluation grade sheet. It was used to collect the check instructors' scores for each task on each pre-recorded check ride maneuver segment video. As shown in Table 1, the 5-point grading scale used for each task was as an ordinal scale representing five distinct levels of piloting proficiency.

Instrument Reliability

To increase instrument reliability, carefully presented training on the maneuver evaluation grade sheets, the 5-point grading scale, and the competency system upon which everything was based was accomplished before beginning the calibration session. Failure of the check instructors to achieve a thorough understanding of these items, as perceived by the calibration facilitator, would have precluded the calibration from taking place. In that case, additional training on these items would have been necessary. However, it was found that the check instructors' understanding of these items was sufficient, but that the calibration results were mixed, so the instrument's reliability was called into question. Additional training on the maneuver evaluation grade sheets, the 5-point grading scale, and the competency system upon which everything was based along with additional calibration sessions was necessary to better understand the instrument's reliability, but such additional training and calibration was outside the scope and approval for this study.

Instrument Validity

The maneuver evaluation grade sheet had been used and evaluated by flight instructors during actual flight training operations at ERAU in Daytona Beach. Its validity was supported by the flight instructors' positive anecdotal feedback. However, the maneuver evaluation grade sheet had not yet been used for check instructor calibration or during actual EOC testing. Also, as shown in the literature, grade sheet complexity, piloting ability, evaluator skill, and evaluator leniency or severity preference compromise the maneuver evaluation grade sheet's validity (Mulqueen et al., 2002).

Data Analysis Approach

Data analyses included:

- A raw agreement percentage across all graded tasks among all check instructors to determine the group's inter-rater agreement,
- A raw agreement percentage across all graded tasks for each check instructor to determine individual referent-rater agreement,
- A raw agreement percentage across all graded tasks across all check instructors to determine the group's referent-rater agreement,
- A Cohen's Kappa coefficient across all graded tasks between each check instructor and every other check instructor averaged across all pairwise comparisons to determine the group's inter-rater reliability,
- A Cohen's Kappa coefficient across all graded tasks between each check instructor and the gold standard averaged across all check instructors to determine the group's referent-rater reliability,
- A Cohen's Kappa coefficient across all graded tasks between each check instructor and the gold standard to determine individual referent-rater reliability,
- A Pearson product-moment correlation coefficient across all graded tasks between the mean score across all check instructors for each task and the gold standard to determine the group's sensitivity to changes in performance,
- A Pearson product-moment correlation coefficient across all graded tasks between each check instructor and every other check instructor averaged

across all check instructors to determine the group's sensitivity to changes in performance,

- A Pearson product-moment correlation coefficient across all graded tasks between each check instructor and the gold standard to determine individual sensitivity to changes in performance,
- A standardized mean absolute difference across all videos between each check instructor and the gold standard to determine individual accuracy, and
- A standardized mean absolute difference across all videos between each check instructor and the gold standard averaged across all check instructors to determine the group's accuracy.

Participant Demographics

Descriptive statistics were determined about the participants. Check instructor ages, flight hours, years of flying experience, and years of check instructor experience were summarized using means, standard deviations, medians, minimums, and maximums.

Reliability Assessment Method

Reliability was inherent in the statistical analyses that were performed for the collected data. Kappa and kappa-like statistics, correlation statistics, and the SMAD have all been previously used as measures of rater reliability (Beaudin-Seiler & Seiler, 2015; Goldsmith & Johnson, 2002; Holt et al. 1997).

Validity Assessment Method

Validity was largely based on face and content validity. The maneuver evaluation grade sheet was used for an unrelated project but that included evaluation of the Basic

Competencies and Behavioral Indicators in Appendix B. Validity was also supported by use of an ordinal grading scale similar to that used in air carrier training environments.

Data Analysis Process

Data analyses took place after Day 2, Day 3, Day 4, and Day 5. A combination of the software programs Microsoft Excel and IBM SPSS was used to enter all the scores and conduct the analyses. The software programs were pre-configured with the appropriate general organization, data entry fields, and formulas ahead of time to increase efficiency in conducting the analyses and returning the appropriate results to support the calibration training. IBM SPSS was also used to calculate descriptive statistics for the participant demographics and all scores from the maneuver evaluation grade sheets. IBM SPSS was also used to evaluate levels of significance for the individual pairwise Cohen's kappa and Pearson product-moment correlation coefficient results. The alpha level used for all analyses was .05, when applicable.

Summary

The research methodology for this study revolved around a within-subjects, pretest-posttest design. A sample of pilot school check instructors evaluated piloting proficiency recorded on digital video by completing a maneuver evaluation grade sheet that corresponded to each video. A total of 18 maneuver segments and 58 individual tasks were available for viewing and evaluating both before the calibration guided discussion and after, providing the statistical power necessary, as related to Cohen's kappa, to show the changes in check instructor performance from pretest to posttest. Changes in raw agreement percentage, inter- and referent-rater reliability, sensitivity, and accuracy showed the effectiveness of the calibration training and helped answer the research

question about whether pilot school check instructors be calibrated against a gold standard to increase reliability and accuracy in their evaluations.

Quantitative results are presented and discussed in the following chapters. Discussion of the results and suggestions and recommendations about future research is presented. The discussion also addresses the validity of the maneuver evaluation grade sheet and the Basic Competencies and Behavioral Indicators in Appendix B as an effective means of describing and evaluating piloting performance.

Chapter IV: Results

Demographics Results

The sample of 10 check instructors from Embry-Riddle Aeronautical University (ERAU) in Daytona Beach, Florida ranged in age, gender, ethnicity, and aviation experience. The participant questionnaire was used to collect specific information about their age, hours of flying experience, years of flying experience, and years of check instructor experience. The descriptive statistics of these data are presented in Table 2. Instead of taking time to collect actual logbook data and because some participants' logbooks may not have been up to date, for efficiency, only estimates of their hours of flying experience were asked to be supplied.

Table 2

Participant Demographic Descriptive Statistics

Variable	<i>n</i>	<i>M (SD)</i>	Median	Min.	Max.
Age	10	29.4 (8.9)	26.0	24.0	53.0
Hours	10	3,140.0 (2,096.7)	2,100.0	2,000.0	7,200.0
Fly Years	10	8.6 (2.5)	8.0	6.0	15.0
Chk Years	10	1.9 (1.2)	1.8	.5	5.0

Note. Hours = total flight hours; Fly Years = number of years of flying experience; Chk Years = number of years of check instructor experience.

Descriptive Statistics

Of the 18 pre-recorded check ride maneuver segment videos that were available, 15 were viewed by each smaller group of five check instructors prior to the calibration guided discussion. Following the calibration guided discussion another 15 videos were viewed. The same maneuver segments were viewed both pretest (before the calibration

guided discussion) and posttest (after the calibration guided discussion), but from different video pools. The two smaller groups did not view the same 15 videos. In order to combine and make the final analysis of the two smaller groups' data, only the commonly-viewed videos were considered. The combining process resulted in 12 remaining videos. Of those 12 videos, 36 individual tasks were graded by the check instructors. However, Video Number 16 in Pool B had different gold standard scores for three of the four tasks than those for Video Number 16 in Pool D, so those tasks were excluded, leaving one common task for both videos. The result was 33 individual graded tasks that all check instructors graded and for which the gold standard score remained the same from pretest to posttest. The descriptive statistics for these 33 graded tasks are shown in Table 3.

Table 3*Summary of Gold Standard and Rater Scores Across All Graded Items*

Rater	N	Pretest	Posttest
		M (SD)	M (SD)
Gold Std	33	3.97 (.394)	3.97 (.394)
Rater 1	33	3.70 (.585)	3.79 (.485)
Rater 2	33	4.00 (.354)	3.82 (.528)
Rater 3	33	3.97 (.174)	3.88 (.545)
Rater 4	33	3.48 (.939)	3.79 (.485)
Rater 5	33	3.79 (.415)	3.91 (.292)
Rater 6	33	3.76 (.561)	4.12 (.893)
Rater 7	33	3.55 (.833)	3.61 (.556)
Rater 8	33	3.94 (.242)	4.00 (.000)
Rater 9	33	3.82 (.465)	3.85 (.364)
Rater 10	33	3.79 (.485)	3.79 (.485)
Average	33	3.78 (.330) 3.78 (.337) ^a	3.85 (.302) 3.82 (.275) ^a

Note. Gold Std = gold standard score; Average = mean score of all raters for each task averaged across all tasks.

^a Results excluding Rater 6.

Reliability and Validity Testing Results

Specific tests for reliability of the data were not performed because reliability was inherent in quantitative data analysis results. Kappa and kappa-like statistics, correlation statistics, and the SMAD have all been previously used as measures of rater reliability (Beaudin-Seiler & Seiler, 2015; Goldsmith & Johnson, 2002; Holt et al. 1997).

Similarly, specific tests for validity were not performed. However, qualitative and anecdotal check instructor feedback about the maneuver evaluation grade sheets and the

Basic Competencies and Behavioral Indicators in Appendix B matched the feedback from flight instructors who used the grading system in an unrelated project.

Quantitative Data Analysis Results

Quantitative data analysis was performed for the common 33 individual graded tasks that all 10 check instructors scored across the common 12 maneuver segments. A combination of Microsoft Excel and IBM SPSS software was used. Some manual calculations were also made.

One check instructor, Rater 6, seemingly changed the grade awarded for 13 tasks to a 5 after the calibration guided discussion. Rater 6 did not grade any task a 5 prior to the calibration discussion. This change was unexpected and did not seem to match the changes made by any other check instructor. Considering Rater 6 as an outlier is supported by the check instructor's mean posttest score of 4.12. No other check instructor's mean score was more than 4.00 either pretest or posttest. As a result, all combined group results will be shown for all 10 check instructors and again for nine check instructors excluding Rater 6. The overall results, discussion, and recommendations in Chapter V are considered without Rater 6.

Agreement Results

A group raw agreement percentage was used as a baseline indication of inter-rater agreement (IRA) among all the check instructors. The method used differs than the common method described by Hallgren (2012) in which the mean agreement between rater pairs is determined for each graded task and then averaged across all tasks. Instead, for each graded task, the scores from all the check instructors were compared. If all 10 scores matched for a given task, the task was marked as an agreement. If one or more of

the 10 scores differed from the others for a given task, the task was marked as a disagreement. The total number of agreements was divided by the total number of tasks to determine an agreement percentage.

Similarly, a group raw agreement percentage was used as a baseline indication of the group's referent-rater agreement (RRA). For each graded task, the scores from all the check instructors were compared to the gold standard score. If all 10 scores matched the gold standard score for a given task, the task was marked as an agreement. If one or more of the 10 scores differed from the gold standard for a given task, the task was marked as a disagreement. The total number of agreements was divided by the total number of tasks to determine an agreement percentage.

Also, individual raw agreement percentages were used as a baseline indication of RRA for each check instructor. For each graded task, the score from a given check instructor was compared to the gold standard score and marked as an agreement if it matched or a disagreement if it did not. The total number of agreements was divided by the total number of tasks to determine an agreement percentage.

Microsoft Excel was used to aid the agreement-marking process. The agreement percentages for both pretest and posttest are shown in Table 4.

Table 4*Percentages of Agreement Across All Graded Items*

Rater	N	Pretest		Posttest	
		IRA	RRA	IRA	RRA
Group	33	27.27 30.30 ^a	24.24 27.27 ^a	12.12 39.39 ^a	12.12 36.36 ^a
Rater 1	33		72.73		78.79
Rater 2	33		84.85		87.88
Rater 3	33		90.91		75.76
Rater 4	33		51.52		81.82
Rater 5	33		75.76		87.88
Rater 6	33		78.79		45.45
Rater 7	33		60.61		60.61
Rater 8	33		87.88		93.94
Rater 9	33		84.85		81.82
Rater 10	33		78.79		78.79
Average	10 9 ^a		76.67 (12.38) 76.43 (13.10) ^a		77.27 (14.30) 80.81 (9.46)

Note. Group IRA = percentage of tasks with 100% inter-rater agreement; group RRA = percentage of tasks with 100% referent-rater agreement; Average = mean (standard deviation) across all check instructors of individual RRA.

^a Results excluding Rater 6.

Reliability Results

A Cohen's kappa statistic was used to determine the group's inter-rater reliability (IRR). Being that Cohen's kappa is appropriate for the comparison of only two raters, an averaging method was used for determining each check instructor's IRR with all the other check instructors, as suggested by Fleiss (1971). The Cohen's kappa statistic was calculated manually for every check instructor pair and then averaged across all pairwise

comparisons to determine the group's IRR. Formula 1 shows the equation used for calculating the statistic manually, as explained in Privitera (2017).

$$k = \frac{P_A - P_E}{1 - P_E} \quad (1)$$

where:

P_A = Percentage of agreement between one pair of two check instructors.

P_E = Percentage of expected error between one pair of two check instructors.

Table 5 and Table 6 report the results of the pretest and posttest pairwise comparisons, respectively. Table 7 reports the averages of those pairwise results. Microsoft Excel was used to aid the calculation process. IBM SPSS was used to verify the results and determine levels of significance.

Table 5

Pretest Pairwise Comparisons of Inter-Rater Reliability

Rater	Rater								
	2	3	4	5	6	7	8	9	10
1	.250*	-.049	.194	.668***	.316*	.105	.017	.000	.482***
2		-.038	.258***	.280**	.336**	.075	-.065	.019	.331**
3			.020	.208	.104	-.046	-.042	.302**	-.050
4				.236*	.220*	.059	.100	.206*	.338***
5					.448**	.198	.141	.306*	.438**
6						.053	.189	.267	.414**
7							.144	.228*	.318**
8								.535***	.318*
9									.260

Note.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 6*Posttest Pairwise Comparisons of Inter-Rater Reliability*

Rater	Rater								
	2	3	4	5	6	7	8	9	10
1	.214	.279*	.012	.500***	-.021	.079	a	.250	.210
2		.086	.214	.225	.102	.276**	a	.258	.214
3			.098	.333**	.061	.072	a	.516***	.189
4				.250	.165*	.079	a	.357*	.605***
5					.132*	.148	a	.436**	.250
6						-.087	a	.104	.072
7							a	.038	-.062
8								a	a
9									.464**

Note.

^a All scores awarded by Rater 8 were the same; unable to compute k with a constant.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 7*Average Inter-Rater Reliability Across All Rater Pairs*

Rater	<i>N</i>	Pretest	Posttest
Rater 1	9	.220 .208 ^a	.190 .221 ^a
Rater 2	9	.161 .139 ^a	.199 .212 ^a
Rater 3	9	.045 .038 ^a	.204 .225 ^a
Rater 4	9	.181 .176 ^a	.223 .231 ^a
Rater 5	9	.325 .309 ^a	.284 .306 ^a
Rater 6	9	.261	.066
Rater 7	9	.126 .135 ^a	.068 .090 ^a
Rater 8	9	.149 .144 ^a	b b
Rater 9	9	.236 .232 ^a	.303 .331 ^a
Rater 10	9	.317 .304 ^a	.243 .267 ^a
Group	c	.202 .187 ^a	.198 .235 ^a

Note.^a Results excluding Rater 6.^b All scores awarded by Rater 8 were the same; unable to compute *k* with a constant.^c Pretest: *N* = 45 including Rater 6; *N* = 36 excluding Rater 6. Posttest: *N* = 36 including Rater 6 but excluding Rater 8; *N* = 28 excluding Rater 6 and Rater 8.

Also of interest was the intra-rater reliability. The analysis was done by determining a Cohen's kappa statistic between the pretest and posttest pairs for each check instructor. IBM SPSS was used to make the analysis and the results are reported in Table 8.

Table 8*Pretest to Posttest Pairwise Comparisons of Intra-Rater Reliability*

Rater	<i>k</i>
1	.136
2	.053
3	.217*
4	.388***
5	.542***
6	.029
7	.012
8	a
9	.421**
10	.802***

Note.

^a All posttest scores awarded by Rater 8 were the same; unable to compute *k* with a constant.

* $p < .05$. ** $p < .01$. *** $p < .001$.

A Cohen's kappa statistic was also used to determine each individual check instructor's referent-rater reliability (RRR) and the group's RRR. In the case of individual RRR, the Cohen's kappa statistic was evaluated directly between each check instructor and the gold standard. The group's RRR was determined by averaging the individual RRR's across all check instructors, similar to how the group's IRR was calculated. Microsoft Excel was used to aid the calculation process and IBM SPSS was used to verify the individual pairwise comparisons and evaluate the level of significance for each individual check instructor's RRR. The results are reported in Table 9.

Table 9*Referent-Rater Reliability Across All Graded Items*

Rater	<i>N</i>	Pretest	Posttest
Rater 1	33	.048	.080
Rater 2	33	.122	.298**
Rater 3	33	-.021	.057
Rater 4	33	.044	.211**
Rater 5	33	.067	.170*
Rater 6	33	.076	.115*
Rater 7	33	.012	.018
Rater 8	33	-.031	.000
Rater 9	33	.250**	.104
Rater 10	33	.080	.080
Group	10 9 ^a	.065 .063 ^a	.113 .113 ^a

Note. Group = mean RRR across all check instructors.

^a Results excluding Rater 6.

* $p < .05$. ** $p < .01$.

Sensitivity Results

A Pearson product-moment correlation was used to evaluate individual and group sensitivity to changes in performance. The correlations were determined between each individual check instructor and the gold standard across all graded tasks. The correlation statistic for each check instructor was then averaged across all check instructors to determine the group's sensitivity. A second method used to determine a group correlation

statistic was to determine the mean score awarded across all check instructors for each graded task and then compare the resulting means to the gold standard across all graded tasks. Microsoft Excel was used to aid the calculation process and IBM SPSS was used to verify the results and evaluate the level of significance for each individual comparison.

All the correlation results are reported in Table 10.

Table 10*Sensitivity Correlations with The Gold Standard*

Rater	Pretest		Posttest	
	<i>N</i>	<i>r</i>	<i>N</i>	<i>r</i>
Rater 1	33	.230	33	.293
Rater 2	33	.449**	33	.574***
Rater 3	33	-.014	33	.273
Rater 4	33	.294	33	.620***
Rater 5	33	.342	33	.519**
Rater 6	33	.249	33	.455**
Rater 7	33	.528**	33	.229
Rater 8	33	-.020	b	b
Rater 9	33	.652***	33	.403*
Rater 10	33	.293	33	.293
Average	10 9 ^a	.300 .306 ^a	9 8 ^a	.406 .400
Group	33 33 ^a	.524** .523 ^a **	33 33 ^a	.618*** .590 ^a ***

Note. Average = mean *r* across all check instructors; Group = correlation between mean scores for each task and the gold standard.

^a Results excluding Rater 6.

^b All scores awarded were the same; unable to compute *r* with a constant.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Accuracy Results

An accuracy statistic was determined by using a standardized mean absolute difference (SMAD) as described by Goldsmith and Johnson (2002). Because the number

of graded tasks for each pre-recorded check ride segment varied, it seemed more appropriate to calculate the SMAD across maneuver segments rather than across graded tasks. To calculate SMAD, the absolute value of the difference between a check instructor's score and the gold standard score averaged across the tasks for a given maneuvers segment was subtracted from one and then divided by the maximum difference possible between any given score and the gold standard score, which, for the grading scale used in this study, was four. The equation is shown in Formula 2.

$$SMAD = \frac{1 - (S_1 - S_g)}{S_d} \quad (2)$$

where:

S_1 = Score awarded by the check instructor.

S_g = Gold standard score

S_d = Maximum possible differential score

After the SMAD for each maneuver segment and each check instructor were calculated, the values were averaged across all maneuver segments for each check instructor. The average SMAD determined the level of accuracy for each check instructor relative to the gold standard. Finally, these final accuracy measurements were averaged across all check instructors to determine the group's accuracy. The results of all SMAD accuracy measurements are reported in Table 11.

Table 11*Accuracy Results Across All Maneuvers Segments*

Rater	Pretest	Posttest
Rater 1	.913	.927
Rater 2	.962	.938
Rater 3	.984	.955
Rater 4	.828	.944
Rater 5	.948	.969
Rater 6	.925	.840
Rater 7	.872	.913
Rater 8	.951	.979
Rater 9	.951	.972
Rater 10	.929	.913
Group	.926 .927 ^a	.935 .946 ^a

^a Results excluding Rater 6.

Summary

The data collected during the overall calibration process allowed the ability to compute a wide variety of statistical analyses. Raw agreement percentages, Cohen's kappa and kappa-like coefficients for inter-, intra-, and referent-rater reliabilities, Pearson product-moment correlation coefficients for sensitivity, and standardized mean absolute difference calculations to determine accuracy all generated lengthy and insightful discussions with the check instructors during the calibration guided discussion periods and the post-calibration debriefing period. The data and statistical analyses were useful in

evaluating the effectiveness of the calibration training. Discussion of these results and of qualitative and anecdotal feedback from the check instructors follows in Chapter V.

Chapter V: Discussion, Conclusions, and Recommendations

Discussion

The statistical measures used show that this study had mixed results. However, the qualitative and anecdotal feedback from the check instructors collected during the calibration guided discussions and the post-calibration debrief session support the importance of furthering this research. In many respects, and when all the statistical measures are considered together, the calibration was somewhat effective. The limitations of this study, many of which were discovered only during the calibration guided discussions, may have contributed to the mixed results. As explained in Chapter IV and unless discussed otherwise, the following is considered excluding Rater 6.

Inter-Rater Reliability

The study was designed around a Cohen's kappa as the primary measure of inter-rater and referent-rater reliability (IRR and RRR respectively). Pretest measurement of IRR showed that 21 out of 45, or 46.7% of the individual pairwise comparisons were significant to at least the $p < .05$ level. Posttest measurement of IRR showed that only 11 out of 36, or 30.6% of the valid pairwise comparisons were significant to at least the $p < .05$ level. However, averaging the kappa measurements using the method described by Hallgren (2012) and then considering the method of categorizing the strength of the agreements as suggested by Landis and Koch (1977) offered more utility and insight, especially when using the measurement in conjunction with other measures and feedback.

The average pretest kappa across all pairs was $\bar{k} = .187$, which showed there was poor agreement, but the average posttest kappa across all pairs was $\bar{k} = .235$, which showed a slight improvement to fair agreement. The change in kappa seemed to be in

alignment with the change in raw inter-rater agreement percentage of 30.30% pretest to 39.39% posttest (see Table 4).

What was interesting to note regarding the kappa measurement was the change from pretest to posttest within each of the smaller groups of five check instructors. For the group including Raters 1 through 5, $\bar{k} = .203$ pretest, which showed poor agreement, and improved very little to $\bar{k} = .221$ posttest, which showed fair agreement. However, for the group including Raters 6 through 10, $\bar{k} = .273$ pretest, which showed fair agreement, and decreased substantially to $\bar{k} = .088$, which showed poor agreement. Excluding Rater 6, the second group changed from $\bar{k} = .301$ to $\bar{k} = .147$, which also showed a change from fair to poor agreement.

The counterbalancing method used in the research design was likely the cause of these changes in the kappa statistic in each of the smaller groups. The group with Raters 1 through 5 viewed video Pool D then Pool A pretest and Pool B then Pool C posttest, whereas the group with Raters 6 through 10 viewed video Pool C then Pool B pretest and Pool A then Pool D posttest. So, there may have been a problem with one or more of the pretest-posttest video pairs (same video number in different pools). The exact problem remained unknown but could have been related to the complexity of the maneuver segment, the number of graded tasks, how each of the graded tasks were represented and performed in the video, or the scenario chosen by the flight standards evaluator differing from that which may have been chosen by the check instructors. However, this was not completely unexpected. As discussed in Chapter 2, Gontar and Hoermann's second theme (2015) of scenario-related influences and fourth theme of rater-related influences, as they relate to the complexity of IRR, might explain these measurements.

In addition, the amount of time allocated and approved for the study was not sufficient for both groups to watch all 18 videos. Both groups only watched 15 videos pretest and posttest, but the two groups did not view the same 15 videos. It is possible that, because the same videos were not viewed by each group, unequal distribution of the maneuver segments and tasks that were scored caused different changes in the kappa statistic.

Referent-Rater Reliability

Although the IRR measurements were informative and generally showed a positive change following the calibration guided discussion, the objective of the research was to calibrate the check instructors against a gold standard. So, the RRR results were of greatest importance to this study. Like IRR, a Cohen's kappa statistic was used to determine RRR, the results of which are reported in Table 9. Overall RRR was low both pretest and posttest, both on an individual basis and on a group basis.

Individual pretest RRR measurements ranged from $k = -.031, p = .711$ for Rater 8 to $k = .250, p = .003$ for Rater 9 with only Rater 9 showing a significant measurement to at least $p < .01$. In other words, only Rater 9 showed a fair agreement to the gold standard that can likely be explained by other than chance alone. The group RRR was only $\bar{k} = .063$, which showed poor agreement.

Individual posttest RRR measurements ranged from $k = .000, p = 1.00$ for Rater 8 and $k = .298, p = .004$ for Rater 2. However, no raters showed disagreement (a negative k), two raters showed fair agreement, seven raters increased their RRR, two raters showed significant agreement at the $p < .05$ level, and two raters showed significant agreement at the $p < .01$ level. That is, although RRR remained poor overall, agreements that were

made were less likely to occur by chance alone than the agreements made prior to calibration. The group RRR increased to $\bar{k} = .113$, which continued to show poor agreement overall.

For RRR, Gontar and Hoermann's (2015) third theme of measurement-related influences may explain the results in addition to their second theme of scenario-related influences and fourth theme of rater-related influences. Measurement-related influences, stemming from the grading system and the maneuver evaluation grade sheets themselves, were not unexpected and likely had the greatest effect for two reasons.

The first reason is that the grading system was new, so the check instructors lacked experience in using it. The second reason is that the three expert evaluators who determined the gold standard scores for each graded task had no more training on or experience with the grading system than the check instructors. The only advantage the expert evaluators had was the ability to discuss the scores with each other in order to come to a unanimous consensus for each score. However, their lack of experience with the grading system meant that any one of gold standard scores could have been wrong.

There was evidence of at least one incorrect gold standard score. Video 3 in both Pool A and Pool C had a gold standard score of 5 for the task Instrument Flight. Every check instructor disagreed with the score by instead grading it a 4 both pretest and posttest. Interestingly, because all the check instructors graded the task a 4, it showed perfect inter-rater agreement. Also interesting, because Rater 6 changed many of the pretest scores to a 5 posttest, that rater agreed with the gold standard posttest, although having had the lowest raw referent-rater agreement (RRA) percentage, which was more than 2 standard deviations from the mean RRA (see Table 4).

Based on feedback collected during the calibration guided discussion and the post-calibration debrief, the possible error with the gold standard score for Instrument Flight seemed to be the result of the specific meaning and identification of the application versus the correlation levels of knowledge. Referencing the Basic Competencies and Behavioral Indicators in Appendix B, the competency Use of Knowledge was the culprit. In Video 3, the flight standards evaluator asked the pilot to maintain the best rate of climb up to the assigned cruise altitude. The expert evaluators who determined the gold standard agreed that the pilot correlated knowledge of aircraft performance to achieve the best rate of climb. However, the check instructors all agreed that the pilot only applied knowledge rather than correlating knowledge because the flight standards evaluator directly asked for a best rate of climb instead of indirectly asking for it, such as with an instruction to expedite the climb.

In that specific case, the flight standards evaluator's solicitation of a specific behavior was appropriate for rater calibration and gold standards training (Baker & Dismukes, 2002; Air Carrier Operations Branch, 2017). However, the check instructors' interpretation and use of the Basic Competencies and Behavioral Indicators and the Maneuver Evaluation Grade Sheet showed measurement-related influences that caused poor reliability in identifying and evaluating it (Gontar & Hoermann, 2015).

Sensitivity and Accuracy

A Pearson product-moment correlation was used to measure the sensitivity of the check instructors' scores to changes in true performance of the pilot. True performance was represented by the gold standard scores for each of the graded tasks. Therefore, measuring sensitivity was done by determining the correlation coefficient between each

check instructor and the gold standard. Table 10 shows the results and reveals that the pretest correlations ranged from $r = -.020, p = .913$ for Rater 8 and $r = .652, p < .001$ for Rater 9. A total of three check instructors had significant correlations to at least the $p < .01$ level. Posttest correlations ranged from $r = .273, p = .124$ for Rater 3 to $r = .620, p < .001$ for Rater 4. While the maximum correlation coefficient did not change much, the minimum did and showed that all the check instructors had a positive correlation posttest. In addition, the number of significant correlations increased from three check instructors to five check instructors to at least the $p < .05$ level. This result was perhaps the most drastic of all the measurements collected.

Furthermore, the group correlation coefficients showed the calibration had an effect. Two methods were used to determine the group correlation coefficient. The first method was simply to average the individual correlation coefficients as described by Goldsmith and Johnson (2002). The resulting group correlation was $\bar{r} = .306$ pretest and $\bar{r} = .400$ posttest. The second method was to determine the mean score awarded across all check instructors for each graded task and then compare the resulting means to the gold standard across all graded tasks. The resulting group correlation was $r = .523, p = .002$ pretest and $r = .590, p < .001$ posttest.

The standardized mean absolute difference (SMAD) statistic was also insightful. The SMAD statistic was used to measure each check instructor's accuracy in matching the gold standard. SMAD, when combined with other measurements, was helpful in explaining other results, in particular RRR results, as explained by Goldsmith and Johnson (2002). As shown in Tables 9 and 11, although RRR was very low, the SMAD

was relatively high, indicating that while the check instructors' scores did not match the gold standard, they were not far off.

Unlike the other statistics, however, SMAD was calculated across videos, or entire check ride maneuver segments, rather than across individual graded tasks. Doing so gave good insight into which videos lacked grading accuracy and provided a starting point for the calibration guided discussions. To further simplify the identification of poorly graded videos, color-coded data was used as shown in Figure 3. The lowest individual SMAD was .500 and the highest possible was 1.000, so gradient coloring was set with a range between those two values, where red represents a SMAD of .500 and white represents a SMAD of 1.000.

Figure 3*Pretest SMAD Measurements with Color Coding*

Video	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10
1	1.000	1.000	1.000	1.000	1.000	0.500	1.000	1.000	1.000	1.000
3	0.875	0.875	0.875	0.875	0.875	0.875	0.625	0.875	0.875	0.875
5	1.000	1.000	1.000	1.000	1.000	1.000	0.833	1.000	1.000	1.000
7	1.000	1.000	1.000	0.625	1.000	1.000	1.000	1.000	1.000	1.000
9	0.938	1.000	0.938	0.750	0.875	0.875	0.813	0.875	0.938	0.938
11										
13										
15										
17	1.000	1.000	1.000	0.833	1.000	0.917	0.583	0.917	0.917	0.833
2	1.000	0.917	1.000	0.917	1.000	1.000	1.000	1.000	1.000	1.000
4	0.833	1.000	1.000	1.000	1.000	1.000	0.917	1.000	1.000	1.000
6										
8	1.000	1.000	1.000	0.938	1.000	1.000	0.938	1.000	0.938	1.000
10	0.813	1.000	1.000	0.750	0.875	0.938	1.000	1.000	1.000	1.000
12										
14										
16	0.500	1.000	1.000	0.500	0.750	1.000	0.750	0.750	0.750	0.500
18	1.000	0.750	1.000	0.750	1.000	1.000	1.000	1.000	1.000	1.000

Note. The figure shows the individual pretest SMAD measurements. The color coding shows that Videos 3, 9, and 16 were the least accurately graded videos and Raters 1, 4, 6, and 7 were the least accurate check instructors.

The group SMAD measurement, which was the average of all individual SMAD measurements, showed that the calibration did have an effect (see Table 11). The mean SMAD increased from .927 pretest to .946 posttest.

The SMAD is more useful than simply comparing the mean scores, especially when the grading scale differs, say from different flight courses or different flight schools, but the tasks, maneuver segments, or competencies remain consistent, or if the grading scale changes or evolves over time, such as changing from a 5-point scale to a 4-

point scale (Goldsmith & Johnson, 2002). In this study, however, the SMAD provided the same insight as and validated the mean scores and standard deviations, which showed two check instructors deviated more than one standard deviation from the mean gold standard pretest, no check instructors deviated more than one standard deviation from the mean gold standard posttest, and the group mean and standard deviation improved from pretest to posttest (see Table 3).

Regarding Rater 6, it was interesting to note that inclusion of the posttest scores awarded by that check instructor caused the group's agreement and reliability measurements to suffer. Neither the raw agreement percentage nor the reliability statistic showed much of any change from pretest to posttest (see Tables 4 and 7). In fact, the pretest to posttest intra-rater reliability for Rater 6 was the second lowest at $k = .029$, $p = .668$. However, the group's sensitivity measurements and significance improved when the scores from Rater 6 were included in the results (see Table 10). Although Rater 6 had less of an ability to identify the correct score, the check instructor was able to identify and account for changes in performance. Rater 6 also had the lowest posttest SMAD of any check instructor, supporting this conclusion, although the check instructor's mean score remained within one standard deviation from the gold standard.

Check Instructor Feedback and Discussion Notes

During each of the calibration guided discussions and the post-calibration debriefing with the entire group of check instructors, written notes were recorded to summarize the check instructors' feedback. The discussion and debriefing used the three-stage process of debriefing in combination with the plus-delta tool described by Gardner (2013). Pratt's Four-Quadrant Model gave focus to the concepts of human resource

development as they were blended into the discussions and debriefing, identifying areas of organizational need in addition to individual check instructor performance and need (Knowles et al., 2012). The feedback received from the check instructors during the calibration guided discussion revolved around a few key themes: (a) the proper use and interpretation of the Basic Competencies and Behavioral Indicators in Appendix B, (b) limitations of the videos and the passive involvement of the check instructors in evaluating the pilots' performance, and (c) using the scoring matrix and determining an accurate score for each graded task.

Basic Competencies and Behavioral Indicators. Beginning with the proper use and interpretation of the Basic Competencies and Behavioral Indicators in Appendix B, the calibration guided discussions helped to reframe the check instructors' understanding of exactly what each of the competencies meant. As previously discussed, the confusion about the Use of Knowledge competency resulted in complete disagreement with the gold standard score for the task Instrument Flight in Video 3. The disagreement was the result of differences in a fundamental understanding between the application level of knowledge and the correlation level of knowledge. However, there were differences in understanding between the competencies themselves. For example, understanding the difference between Use of Knowledge and Adherence to Standard Operating Procedures caused some consternation in determining the appropriate score for some of the graded tasks.

Video 16 was a good case study for illustrating the difference between these two competencies and the difficulty the check instructors had at using them. The maneuver segment for Video 16 involved the tasks Compliance with Air Traffic Control

Procedures; Holding Procedures; Departure, En Route, and Arrival Operations; and Precision Instrument Approach (ILS with Course Reversal). In the scenario, the pilot was instructed to fly direct to the initial approach fix CALOO, perform the published holding pattern course reversal in lieu of a procedure turn, and then complete the ILS approach to Runway 5 at Paige Field in Fort Myers, Florida (KFMY). In one instance, the pilot performed a standard procedure turn instead of the published holding pattern course reversal in lieu of a procedure turn. Many of the check instructors attributed the incorrect course reversal as improper use of knowledge about holding patterns, but the calibration guided discussion centered on the idea that the incorrect course reversal could have instead been a failure to adhere to standard operating procedures. If the former was true, then the task Holding Procedures was correctly awarded low scores. However, if the latter was true, then the task Departure, En Route, and Arrival Operations deserved the low scores. Referencing the FAA Instrument Rating Airplane ACS, the former would have been a failure of knowledge element IR.III.B.K1 whereas the latter would have been a failure of skill element IR.V.B.S6 (Airman Testing Standards Branch, 2019).

It was clear from the discussions that the check instructors tended to weigh knowledge, and therefore attribute poor performance to lack of knowledge, more heavily than adherence to standard operating procedures. They also tended to evaluate only the proficiency elements directly listed for the task being performed rather than considering more appropriate proficiency elements in other tasks related to the overall maneuver segment. In other words, they confused the two competencies and as a result pinpointed the wrong proficiency element in the ACS as the source of the failure. Following the calibration guided discussions, the check instructors evaluated competencies with less

attenuation toward other tasks in the scenario and tended to consider a wider range of proficiency elements in their evaluations, as was observed by greater use of the Basic Competency and Behavioral Indicators handout and by reference to the Instrument Rating Airplane ACS rather than recalling it from memory.

Passive Check Instructor Involvement. Continuing with the limitations of the videos and the passive involvement of the check instructors, each video limited the check instructors' abilities to fully evaluate the pilots' performances as they would normally be able to do. These factors may have also attributed to some of the grading inaccuracies. While the videos were of high quality and resolution and fully displayed all the flight instruments, aircraft controls, and the pilots' lap area where they normally have their electronic flight bag and checklist, the check instructors expressed that some of the information they needed was missing. The missing information was in the form of their active participation during the evaluation. During a normal check ride, each check instructor crafts his or her own scenario to solicit specific behaviors from the pilot. Frequently, the scenario evolves based on the performance of the pilot. Each check instructor also views the entire performance for all tasks and the EOC test from start to finish during a normal check ride.

However, for this study, the check instructors were not able to create their own scenarios and instead were forced to evaluate the pilots based on the scenarios created by the flight standards evaluator. While the flight standards evaluator allowed each scenario to progress in a certain direction, the check instructors may or may not have guided the scenarios in the same direction had they been involved. Some of the comments and discussion about passive check instructor involvement centered around the idea that

based on any particular observed behavior or lack of behavior, oral questioning, different task ordering, or different air traffic control instructions may have been used by the check instructors. These personal approaches and involvement in the conduct of the evaluations appear to be an important part of the evaluation process, but not detrimental to the goal of calibration. Whereas calibration serves the goal of consistently, reliably, and accurately grading or judging proficiency, the method, technique, and preference in assessing the performance varies. In fact, the FAA clearly differentiates judgment and assessment in similar terms (Airman Testing Standards Branch, 2020).

As for the videos themselves, the check instructors found it difficult to remember air traffic control instructions early in the videos for use later in the videos. The check instructors also commented on the fact that each video lacked the context of an overall EOC test during which the pilots' performance is evaluated from start to finish and judgment of questionable tasks or proficiency elements can be withheld until related tasks or repeated proficiency elements occur later during the evaluation.

Using Video 16 again as an example case, the task Departure, En Route, and Arrival Procedures was unsatisfactory because of a failure of the proficiency element IR.V.B.S6, which is "comply with all applicable charted procedures" (Airman Testing Standards Branch, 2019, p. 14). On a normal EOC test, that proficiency element would occur at least three times because a minimum of three instrument approaches is required to be flown and each approach is preceded by arrival procedures. If the pilot makes a mistake one time but the other two are performed without error, the check instructor may find the pilot satisfactory at that proficiency element, especially if the safety of flight was never in question. This type of decision making agrees with the leeway afforded by the

ACS when it states unsatisfactory performance includes, among other things, consistently exceeding the tolerances specified in the skill elements of a task (Airman Testing Standards Branch, 2019, p. A-9). Colloquially, evaluators of all types refer to this as “looking at the big picture,” but in the case of the videos used in this study, the “picture” lacked context.

Scoring Accuracy. Finally, with regard to using the scoring matrix and determining an accurate score for each graded task, the check instructors tended to focus too much on the matrix itself and focused less on the Basic Competencies and Behavioral Indicators in Appendix B. The calibration guided discussions attempted to correct such tendency by doing a few things. One, the check instructors were reminded to continue to “look at the big picture.” Although a particular pilot’s performance may have lacked the context of an entire EOC test, a few performances were clearly unsatisfactory. Prior to the calibration, the check instructors tended to grade such performances based on the scoring matrix and derived a score that didn’t agree with their so-called gut feeling or what they knew to be correct.

The proper approach that was discussed during calibration was instead to use the Basic Competencies and Behavioral Indicators to justify their decision, tie the behavior back to the appropriate ACS proficiency element, and then use the scoring matrix to fine-tune the score. The statistics revealed how the discussion affected the scores.

For example, the task Non-Precision Instrument Approach (VOR with Procedure Turn) on Video 9 had a gold standard score of 2, which was unsatisfactory. Prior to calibration, scores awarded by the check instructors for that task, including Rater 6, ranged from 1 to 4 with a mean of 2.8 and a standard deviation of .919. After calibration,

scores awarded by the check instructors for that task, including Rater 6, ranged from 2 to 4. The mean was still 2.8, but with no 1s awarded and one less 4 awarded, the standard deviation decreased to .632. Considering that task with the others in the video, the average SMAD for Video 9 across all check instructors before calibration was .894 but improved to .938 after calibration. As Video 9 was one of the three with least accuracy and therefore targeted during the calibration guided discussions, the improvement showed a positive effect of the calibration.

Pratt's Four-Quadrant Model. As mentioned, Pratt's Four-Quadrant Model was used to give focus to the concepts of human resource development. Specifically, during the reactions phase of the calibration guided discussions and post-calibration debriefing, the model was used to both help stimulate discussion and organize the feedback about calibration into two categories—individualized learning and organizational training. The model structured the reactional feedback and helped provide context and stimulate further discussion during the understanding phase of the calibration guided discussions and post-calibration debriefing when specific statistical analyses were presented and discussed.

Generally, the feedback was mixed and showed balance with respect to the need for direction, or organizational training, but showed a greater desire for support. As a group, the check instructors fell in Quadrant 1 (see Figure 4). The check instructors expressed the following points:

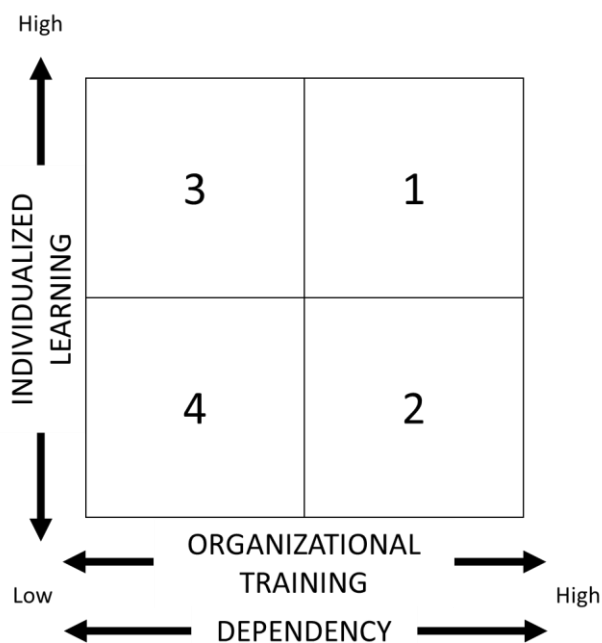
- More training was required to fully realize the benefits of evaluator calibration, but the training already received provided a much more objective and precise understanding of how to evaluate pilot performance compared to

previous training, even while using traditional tools such as plans of action and the FAA ACS; this supports the need for greater organizational training.

- More self-study on and time to review the Basic Competencies and Behavioral Indicators in Appendix B was needed to be able to more efficiently identify behaviors, which validated the discussions that took place about specific videos; this supports the desire for greater individualized learning.
- Practice videos with practice grading as a group would have been very beneficial at helping to apply the Basic Competencies and Behavioral Indicators in Appendix B; this supports the need for both increased individualized learning through collaboration and encouragement from peers and greater organizational training through use of guided discussions to ensure the practice grading is within reasonable accuracy to the gold standard.

Figure 4

Pratt's Four-Quadrant Model Applied to Check Instructor Calibration



Note. Adapted from (Pratt, 1988).

Conclusions

Can pilot school check instructors be calibrated against a gold standard to provide reliable, accurate, valid, and consistent evaluations? The answer is yes. While the statistical results were mixed overall, many of the individual and specific results showed positive changes as a result at the calibration attempt. In particular, improvements in RRR, sensitivity correlations, and SMAD accuracy measurements were shown. These positive changes, considered together and with the feedback collected from the check instructors showed support for calibration training for pilot school check instructors.

To conclude the debriefing session following calibration, the group of 10 check instructors were encouraged to come to a consensus on three items that went well with

the calibration and three items that should be targeted for improvement should further calibration efforts take place. The conclusion discussion followed the plus-delta tool described by Gardner (2013).

The three items that went well with the calibration were:

- The check instructors were able to focus on more proficiency elements, were able to be more objective with their judgements, and were able to more easily justify the grades awarded.
- The training received about grading and behavior was effective and beneficial.
- Knowledge of the ACS and experience conducting EOC tests allowed easy adaptation to the new grading methods.

The three items that should be targeted for improvement were:

- More tools and guidance for each of the videos used during calibration should be considered to counteract the passive involvement of the check instructors.
- More practice videos and group grading should be used to improve accuracy, especially whether unsatisfactory performance should be scored a 2 or a 3 or whether satisfactory performance should be scored a 4 or a 5.
- More guidance and support during practice grading sessions to help identify behavioral evidence instead of performance outcomes.

These plus-delta debriefing items should not only be interpreted as reflective critique, but also as insight into how further research about pilot school check instructor calibration should take place.

Theoretical Contributions

It was shown that while rater calibration training is a standard method of training instructors and evaluators in air carrier settings, the literature about pilot school check instructor calibration is lacking. The results and findings of this study contribute to the body of knowledge because they focus specifically on pilot school check instructor calibration and show that there are merits in continuing research in this specific area. Additional contributions can be made regarding the differentiation between behavioral evidence and proficiency outcomes in a primary or general aviation flight training setting. By providing a starting point for developing a competency system that complements the FAA ACS, the intention is that primary flight training and practical testing can target and improve the fundamental behaviors that generate certain outcomes of performance.

Practical Contributions

The results and findings of this study provide needed data and guidance toward the ultimate goal of improving flight safety—a goal shared by regulators, organizations, and individual pilots and flight instructors. The statistical analyses and qualitative feedback show that it is possible to reorient and calibrate pilot school check instructors toward evaluating fundamental behavior. Because it was already shown to be possible for air carrier evaluators and because this study finds it possible for pilot school check instructors, it is fair to say that the process and methods can be applied to any type of pilot evaluator, such as an FAA-designated pilot examiner or any flight instructor.

This paper also describes how teaching and learning processes and methods were entangled with the calibration activities to supplement the lack of experience with calibration and the lack of already existing data regarding calibration at pilot schools.

Similarly, other pilot schools or smaller organizations or individuals may lack the experience, data, or even resources to implement their own calibration programs. By using carefully selected teaching and learning methods, collaborating with organizations that do possess the resources, and implementing technological tools, the lack of experience and resources can be mitigated, producing at least nominal improvements in general aviation flight safety.

A final contribution that the results and findings this study provides is an overview and explanation of the types of data that can be generated from calibration efforts and the insight the data provide in driving training and organizational development or personal performance improvement. There is so much data about a pilot's performance on a check ride that is currently not collected or even known. The advent of the FAA ACS is a step in the right direction because they provide a robust standard and a coding system that pinpoints proficiency weaknesses. However, creating grading scales and training programs specifically geared toward collecting and analyzing the behavioral evidence related to those proficiency elements will be much more beneficial than the traditional pass or fail grading method.

Limitations of the Findings

Limitations do exist with the findings in this study, so future research must take them into account but also develop and suggest methods to overcome them. Some of the key limitations and suggestions follow.

First, limited improvement in the statistical measures themselves suggests the possibility that repeated or future calibration attempts or studies may not be successful. The feedback from the check instructors, however, suggests it was beneficial and that

there is a desire for more time and practice with grading and behavioral evaluation. The time allotted and approved for this study was not sufficient. The calibration guided discussions were only 1 hour in duration. A greater amount should be allotted to see greater results, but the limited statistical results from this study make the justification for doing so difficult.

Second, while the check instructors stated that the video recordings and pilot performances were authentic, the pilots who volunteered to help produce the videos came from a very restrictive subset of the ERAU pilot population, which is itself a limited subset of the entire pilot population. At the time, the pilots were within 1 month of completing their instrument airplane flight training, either having already completed it or about to. The intention was to ensure realistic and authentic performances of what is normally observed on a real EOC test, but in doing so, the range of performances may have been too limited, limiting the range of gold standard scores, and thereby limiting the check instructors' ability or opportunity to evaluate a wider range of behaviors and competencies. The reason it was necessary to have volunteer pilots assist with producing the videos was because no videos existed. Moving forward, it would be important for organizations to begin recording actual check ride performances from the broader pilot population in advance of any calibration efforts. Experts should review the videos and retain the ones that present a wide range of behaviors and competencies for future calibration training.

Third, it is important to understand that the calibration that took place in this study had no impact on actual flight training or flight evaluation activities. No data previously existed or is currently being collected for which analysis can be done or

against which changes can be measured. A true calibration program would measure impacts to actual evaluation data, such as a particular check instructor or evaluator improving grading accuracy over time. Just like the recording and selection of videos must begin well in advance of any calibration efforts, so too must appropriate grading systems be implemented and data collection begin in advance.

Recommendations

Recommendations for Pilot Schools

Pilot schools that have self-examining authority should desire to improve the quality of their workforce and the quality of their graduates. In doing so, pilot school administrators should evaluate how the implementation of different grading systems can generate data collection streams and can be used for calibration efforts. Implementation of effective human resource development methods blended with calibration may significantly improve check instructor evaluator accuracy and reliability. It is recommended that pilot schools focus on these areas:

- Move away from traditional pass or fail grading systems and toward grading scales that objectively and precisely describe behavior and performance. Grading should continue to be learner-centered regardless of the scale used.
- Partner with other pilot schools and large flight training organization to develop and standardize a broader set of core competencies that can apply to all of general aviation primary flight training and prepare pilot school graduates for similar competency-based evaluation as they advance in their aviation careers. The Basic Competencies and Behavioral Indicators presented in Appendix B serve as substantial starting point. However, the list of

competencies might be expanded to include instructional competency and professional competency of pilot school staff, thereby linking organizational success with human resource development and graduate success.

- Regardless of the grading system or competency system used, it is recommended to begin collecting data related to evaluator performance during EOC testing or practical testing. Such data, in the form of video recordings and grading data can be used to support practice evaluation, training, and calibration efforts of pilot school check instructors.

Recommendations for Future Research

While this study focused on check instructor calibration itself, prerequisite materials needed to be created prior to conducting the calibration. Each of those warrants research into their appropriateness and applicability. The following is recommended:

- Further research should focus on the development and analysis of general aviation competencies and behaviors as they apply to specific proficiency elements detailed in the FAA ACS and other standards of performance. As recommended earlier, development of such competencies should be a partnership between flight training organizations toward an industry-wide standard. The FAA should also be involved in any research and development of such competency systems.
- Additional research should be done on grading scales and grading systems appropriate to general aviation primary flight training and that properly link behavior, proficiency, competency, and certification standards and that

simultaneously detail learning progress while undergoing training and assessment results while undergoing testing.

References

- Air Carrier Operations Branch. (2017). *Advanced qualification program* (AC 120-54A, Change 1). U.S. Department of Transportation, Federal Aviation Administration. http://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_120-54A_CHG_1.pdf
- Airman Testing Standards Branch. (2019). *Instrument rating - airplane airman certification standards* (FAA-S-ACS-8B, Change 1). U.S. Department of Transportation, Federal Aviation Administration. https://www.faa.gov/training_testing/testing/acs/media/instrument_rating_acs_change_1.pdf
- Airman Testing Standards Branch. (2020). *Aviation instructor's handbook* (FAA-H-8083-9B). U.S. Department of Transportation, Federal Aviation Administration. https://www.faa.gov/regulations_policies/handbooks_manuals/aviation/aviation_instructors_handbook/media/aviation_instructors_handbook.pdf
- Al-Romaithi, S. A. (2006). *An airline ab initio flight training program in the United Arab Emirates (UAE): An analytical approach toward emiratization* [Master's thesis, Embry-Riddle Aeronautical University - Daytona Beach]. Available from ProQuest Dissertations and Theses database. (UMI No. EP32105)
- Baker, D. P. (2002). *CRM assessment: Determining the generalization of rater calibration training*. Washington: American Institutes for Research. <http://hdl.handle.net/2060/20020072221>
- Baker, D. P., & Dismukes, R. K. (2002). A framework for understanding crew performance assessment issues. *The International Journal of Aviation Psychology*, 12(3), 205-222. https://doi.org/10.1207/S15327108IJAP1203_2
- Beaudin-Seiler, B. M., & Seiler, R. (2015). A study of how flight instructors assess flight maneuvers and give grades: Inter-rater reliability of instructor assessments. *Journal of Aviation/Aerospace Education & Research*, 25(1), 73-102. <https://doi.org/10.15394/jaaer.2015.1652>
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Handbook I: cognitive domain. *Taxonomy of educational objectives: The classification of educational goals*. [https://www.uky.edu/~rsand1/china2018/texts/Bloom et al -Taxonomy of Educational Objectives.pdf](https://www.uky.edu/~rsand1/china2018/texts/Bloom%20et%20al%20-%20Taxonomy%20of%20Educational%20Objectives.pdf)
- Bujang, M. A., & Baharum, N. (2017). Guidelines of the minimum sample size requirements for Cohen's kappa. *Epidemiology Biostatistics and Public Health*, 14(2), e12267-1 to 10. doi:10.2427/12267
- Carlson, R. (1989). Malcolm Knowles: Apostle of andragogy. *Vitae Scholasticae*, 8(1). <http://www.nl.edu/ace/Resources/Knowles.html>

- Department of the Air Force. (2003). *Guidebook for air force instructors* (AFMAN 36-2236). U.S. Department of Defense.
<https://www.angtec.af.mil/Portals/10/Courses%20resources/afman36-2236.pdf?ver=2018-10-02-084122-173>
- Embry-Riddle Aeronautical University. (2021, January 15). *New grading system: For FA 323 only: Spring 2021* [PowerPoint slides]. Flight Training Department.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), pp. 378-382. <https://doi.org/10.1037/h0031619>
- Flin, R., & Martin, L. (2001). Behavioral markers for crew resource management: A review of current practice. *The International Journal of Aviation Psychology*, 11(1), 95-118. https://doi.org/10.1207/S15327108IJAP1101_6
- Gardner, R. (2013). Introduction to debriefing. *Seminars in Perinatology*, 37(3), 166-174. <https://doi.org/10.1053/j.semperi.2013.02.008>
- Goldsmith, T. E., & Johnson, P. J. (2002). Assessing and improving evaluation of aircrew performance. *The International Journal of Aviation Psychology*, 12(3), 223-240. https://doi.org/10.1207/S15327108IJAP1203_3
- Gontar, P., & Hoermann, H.-J. (2015). Interrater reliability at the top end: Measures of pilots' nontechnical performance. *The International Journal of Aviation Psychology*, 25(3-4), 171-190. <http://doi.org/10.1080/10508414.2015.1162636>
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutor Quant Methods Psychol.*, 8(2), pp. 23-34. <http://doi.org/10.20982/tqmp.08.1.p023>
- Holt, R. W., Hansberger, J. T., & Boehm-Davis, D. A. (2002). Improving rater calibration in aviation: A case study. *The International Journal of Aviation Psychology*, 12(3), 305-330. https://doi.org/10.1207/S15327108IJAP1203_7
- Holt, R. W., Johnson, P. J., & Goldsmith, T. E. (1997). Application of psychometrics to the calibration of air carrier evaluators. *Proceeding of the Human Factors and Ergonomics Society Annual Meeting*, 41(2), 916-920. <http://doi.org/10.1177/107118139704100244>
- International Air Transport Association. (2013). *Evidence-Based Training Implementation Guide*. <https://www.iata.org/contentassets/c0f61fc821dc4f62bb6441d7abedb076/ebt-implementation-guide.pdf>
- International Civil Aviation Organization. (2013). *Manual of evidence-based training* (Doc 9995). United Nations.

- Kim, A. H., Chutinan, S., & Park, S. E. (2015). Assessment skills of dental students as peer evaluators. *Journal of Dental Education*, 79(6), 653-657.
<https://doi.org/10.1002/j.0022-0337.2015.79.6.tb05937.x>
- Knowles, M. S., Holton III, E. F., & Swanson, R. A. (2012). *The adult learner: The definitive classic in adult education and human resource development*. London: Routledge.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), pp. 159-174. <http://doi.org/10.2307/2529310>
- Mulqueen, C., Baker, D. P., & Dismukes, R. K. (2002). Pilot instructor rater training: The utility of the multifacet item response theory model. *The International Journal of Aviation Psychology*, 12(3), 287-303.
https://doi.org/10.1207/S15327108IJAP1203_6
- O'Connor, P., & Long, W. M. (2011). The development of a prototype behavioral marker system for US Navy officers of the deck. *Safety Science*, 49(10), 1381-1387.
<https://doi.org/10.1016/j.ssci.2011.05.009>
- Pratt, D. D. (1988). Andragogy as a relational construct. *Adult Education Quarterly*, 38(3), pp. 160-181.
- Privitera, G. J. (2017). *Research Methods for the Behavioral Sciences: Second Edition*. Los Angeles: Sage Publications.

Appendix A

Permission to Conduct Research

Embry-Riddle Aeronautical University
Application for IRB Approval
EXEMPT Determination Form

Principal Investigator: Paul M. Cairns

Other Investigators: Andrew R. Dattel


Role: Student Campus: Daytona Beach College: Aviation/Aeronautics

Project Title: Gold Standards Training and Evaluator Calibration of Pilot School Check Instructors

Review Board Use Only

Initial Reviewer: Teri Gabriel Date: 01/26/2021 Approval #: 21-070

Determination: Exempt

Dr. Beth Blickensderfer  Elizabeth L.
 IRB Chair Signature: Blickensderfer, Ph.D. Digitally signed by Elizabeth L. Blickensderfer, Ph.D.
Date: 2021.01.26 16:43:28 -0500

Brief Description:

The purpose of this study is to determine applicability and effectiveness of airline-style pilot evaluator training as applied to pilot school evaluation activities. The check instructor calibration this study details is built around the evaluation activities for the instrument airplane rating conducted in qualified flight training devices. To enhance the effectiveness of check instructor calibration, appropriate application of learning and teaching theories is necessary. This study will also detail learning facilitation principles, adult teaching methods, and human resource development (HRD) methods that may yield such enhancement. Participants will be asked to complete a demographics questionnaire; then receive classroom instruction about general evaluator calibration and associated grading scales, view pre-recorded simulated flight check ride segments, grade them, receive additional classroom instruction, view additional pre-recorded simulated flight check ride segments, and grade those. They will also complete pre-test and post-test analysis of inter-rater and referent-rater reliability.

This research falls under the EXEMPT category as per 45 CFR 46.104:

- (1) Research, conducted in established or commonly accepted educational settings, that specifically involves normal educational practices that are not likely to adversely impact students' opportunity to learn required educational content or the assessment of educators who provide instruction. This includes most research on regular and special education instructional strategies, and research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods. (Applies to Subpart B [Pregnant Women, Human Fetuses and Neonates], does not apply for Subpart C [Prisoners] except for research aimed at involving a broader subject population that only incidentally includes prisoners, and Subpart D [Children] involved in research.)

Human Subject Protocol Application

Campus: Daytona Beach College: COA
 Applicant: Paul Calms Degree Level: Master
 ERAU ID: ERAU Affiliation: Student

Project Title: Gold Standards Training and Evaluator Calibration of Pilot School Check Instructors

Principal Investigator: Paul M. Calms
 Other Investigators: Dr. Andrew R. Dattel

Submission Date: 12/10/2020
 Beginning Date: 02/01/2021
 Type of Project: Experiment
 Type of Funding Support (if any):

Questions:

1. Background and Purpose: Briefly describe the background and purpose of the research.

For decades, many large, U.S.-based air carriers have used a voluntary, alternative training program called advanced qualification program (AQP) to train their pilots, instructors, and evaluators. One of the key components of AQP is the calibration of evaluators in observing and grading crew resource management behaviors that are required to be demonstrated by pilots to evaluators during evaluation activities. AQP requires that rater reliability training be provided to evaluators to ensure the AQP data remains reliable and valid. The purpose of this study is to determine applicability and effectiveness of airline-style pilot evaluator training as applied to pilot school evaluation activities. The check instructor calibration this study details is built around the evaluation activities for the instrument airplane rating conducted in qualified flight training devices. To enhance the effectiveness of check instructor calibration, appropriate application of learning and teaching theories is necessary. This study will also detail learning facilitation principles, adult teaching methods, and human resource development (HRD) methods that may yield such enhancement.

2. Time: Approximately how much time will be required of each participant?

A total of approximately 8 hours spread over 6 days will be required for each participant to complete the study.

On day 1, all participants will receive an introductory briefing lasting about 10 minutes. They will then complete a participant questionnaire to collect demographic data. This will take approximately 1 minute. Participants will then receive a briefing about general evaluator calibration and associated grading scales. This will take approximately 1 hour. The participants will then be sub-divided into two smaller groups, each of which will participate on either days 2 and 3 or days 4 and 5.

On days 2 and 4 each smaller group of participants will spend approximately 2 hours viewing 18 videos while simultaneously completing maneuver evaluation grade sheets. A 10-minute break will be added after the first 9 videos. On days 3 and 5 each smaller group of participants will participate in a guided discussion about the previous day's videos and the pretest scores as analyzed from their completed maneuver evaluation grade sheets. This will take approximately 1 hour followed by a 10-minute break. They will then spend approximately 2 hours watching the other 18 videos and simultaneously completing maneuver evaluation grade sheets. A 10-minute break will be added after the first 9 videos.

On day 6, the entire group of participants will participate in a guided discussion about the previous day's videos and the post-test scores as analyzed from their completed maneuver evaluation grade sheets. This will take approximately 1 hour. A conclusion briefing will end the participants' involvement and will last approximately 20 minutes.

3. Design, Procedures and Methods: Describe the details of the procedure(s) to be used and the type of data that will be collected.

The study will be a within-subjects design with one group of participants. A pretest and post-test analysis of data collected on the maneuver evaluation grade sheets will take place.

Participants will receive classroom instruction about general evaluator calibration and associated grading scales. The participants will then be split into two smaller groups for counter-balancing.

Each smaller group will view 18 videos of pre-recorded simulated flight check ride segments, fill out the maneuver evaluation grade sheet associated with each video (pretest), participate in a guided discussion lesson about the grades, view an additional 18 videos of pre-recorded simulated flight check ride segments, fill out the maneuver evaluation grade sheet associated with each video (post-test), and receive feedback about the second set of grades. For each small group, videos will be presented in a different order for counter-balancing.

The pretest and post-test analysis of inter-rater and referent-rater reliability will determine the effectiveness of the guided discussion lesson and the applicability of airline-style pilot evaluator training to pilot school pilot evaluation. The pretest will be the data collected on the maneuver evaluation grade sheets associated with the first set of 18 videos prior to the guided discussion that follows. The post-test will be the data collected on the maneuver evaluation grade sheets associated with the second set of 18 videos after the guided discussion about the scores from the first set. The guided discussion between the two sets of videos is the stimulus. The attached video screenshots and lesson plan show the details. For example, one of the small groups of participants will view and grade Pool A Video 1 (among a total of 18 videos in Pools A and B), receive feedback about their grading and level of inter-rater and referent-rater reliability during the guided discussion, then view and grade Pool C Video 1 (among a total of 18 videos in Pools C and D), and receive feedback about the change in the level of inter-rater and referent-rater reliability.

There are 36 unique videos (either the pilot is different or the maneuver segment is different). However, every video looks the same and the perspective never changes. The screenshots in the attached document are representative of all the videos.

Training will take place at the ERAU Daytona Beach Flight Training Department. Training will be provided by me, Paul Cairns, Assistant Chief Flight Instructor and principal investigator for the study. Approval for the participants to take part in the study during normal working hours was granted by email and a copy of the approval is attached to the application. The maneuver evaluation grade sheets are attached to this application.

4. Measures and Observations: What measures or observations will be taken in the study?

Participant demographics to include age, flight hours, years of flying experience, and years of check instructor experience will be collected using a participant questionnaire at the beginning of the classroom instruction. Maneuver evaluation grade sheets will be used by each participant to grade several pre-recorded flight check ride segments on a scale from 1 to 5 representing varying levels of pilot performance and learning achievement. The data analysis of the maneuver evaluation grade sheets will be done in SPSS and determine the inter-rater and referent-rater reliability based on the scores from the maneuver evaluation grade sheets.

5. Participant Population and Recruitment Procedures: Who will be recruited to be participants and how will they be recruited. Any recruitment email, flyer or document(s) must be reviewed by the IRB. Note that except for anonymous surveys, participants must be at least 18 years of age to participate.

Ten check instructor participants will be selected from among ERAU Daytona Beach Flight Training Department staff; the participants will be asked verbally to participate in the study and advised that participation is voluntary. The following verbiage will be used: "I am asking for your help with a research project. The project is designed to measure the effectiveness of check instructor calibration training. Calibration means that each check instructor improves their ability identify and evaluate piloting behavior in the most standardized fashion. The research will involve your participation in classroom training, watching pre-recorded simulated flight check ride segments, and grading the pilot performance. Please let me know if you are interested in participating. Further details will be provided and your written consent will be required before the study begins."

6. Risks or Discomforts: Describe any potential risks to the dignity, rights, health or welfare of the human subjects. All other possible options should be examined to minimize any risks to the participants.

Participants will not be exposed to any harm or adverse conditions. The classroom environment used will be the same as that experienced by the participants in their normal day-to-day work and educational activities at ERAU. In these uncertain times, there is a risk of contracting COVID-19. In addition to following the established ERAU university policies, the following cleaning procedures will be conducted prior to and during the research study to mitigate these risks: The researcher and participants will be required to wash their hands before beginning the study and will touch nothing between the bathroom and research area. The researcher will use a disinfectant wipe to wipe all surfaces that are touched by the participants or researcher prior to and after the study. Participants and the researcher will remain socially distanced throughout the study.

7. Benefits: Assess the potential benefits to be gained by the subjects as well as to society in general as a result of this project.

Participants of this study may benefit by receiving on-the-job training that may assist with their normal work duties at ERAU. The results of this study may be important in advancing flight training and evaluation methods at pilot schools, especially those that hold examining authority. Large flight training organizations with teams of flight instructors and check instructors may benefit from the study through increased standardization, improved human resource development, and implementation of data-collections streams. In general, any flight training organization may find the results of this study useful in making nominal improvements to their processes and procedures.

8. Informed Consent: Describe the procedure you will use to obtain informed consent of the subjects. How and where will you obtain consent? See Informed Consent Guidelines for more information on Informed Consent requirements.

The informed consent form will be presented to each participant and the study will be explained and any and all questions answered. Participants who wish to participate in the study will be asked to sign the consent form and those who do not wish to participate do not have to sign the form and may leave the classroom while the study is conducted.

9. Confidentiality of Records: Will participant information be anonymous (not even the researcher can match data with names), confidential (Names or any other identifying demographics can be matched, but only members of the research team will have access to that information. Publication of the data will not include any identifying information.), or public (Names and data will be matched and individuals outside of the research team will have either direct or indirect access. Publication of the data will allow either directly or indirectly, identification of the participants.)?

Confidential

9b. Justify the classification and describe how privacy will be ensured/protected.

Names or other identifying information will not be asked for. Demographic data collected on the participant questionnaire will not be associated with the individual completing it. A randomly assigned identifier for each participant will be used to link demographic data and data sets, but its association with individual participants will not be recorded.

10. Privacy: Describe the safeguards (including confidentiality safeguards) you will use to minimize risks. Indicate what will happen to data collected from participants that choose to "opt out" during the research process. If video/audio recordings are part of the research, describe how long that data will be stored and when it will be destroyed.

Data will be collected confidentially. The association of the randomly assigned identifiers with participants will not be recorded. If a participant opts-out after beginning the study, data collected to that point will be maintained but not included in the results. All data collected and all pre-recorded simulated flight check ride segment videos will be destroyed three months after completion of the study.

11. Economic Considerations: Are participants going to be paid for their participation?

No

11b. What will the compensation be?

Describe your policy for dealing with participants who 1) Show up for research, but refuse informed consent; 2) Start but fail to complete research.

Check instructor participants will be participating as part of their normal work duties at ERAU and will be paid their regular hourly wage for the on-duty work performed and training received. Although the participants are participating in the study during on-duty work periods and being paid accordingly, participation is strictly voluntary. Participation will not provide extra work-related benefits such as overtime pay or bonuses. If participants fail to complete the study or refuse informed the informed consent, they will keep all wages already earned.

By submitting this application, you are signing that the Principal Investigator and any other investigators certify the following:

1. The information in this application is accurate and complete
2. All procedures performed during this project will be conducted by individuals legally and responsibly entitled to do so
3. We will comply with all federal, state, and institutional policies and procedures to protect human subjects in research
4. We will assure that the consent process and research procedures as described herein are followed with every participant in the research
5. That any significant systematic deviation from the submitted protocol (for example, a change in the principal investigator, sponsorship, research purposes, participant recruitment procedures, research methodology, risks and benefits, or consent procedures) will be submitted to the IRB for approval prior to its implementation
6. We will promptly report any adverse events to the IRB

Electronic Signature:

Paul M. Calms

Appendix B

Basic Competencies and Behavioral Indicators

Competency	Description	Behavioral Indicators/Levels
Use of knowledge	Demonstrates the knowledge level required of each task/line item in accordance with the knowledge elements found in the applicable FAA Airman Certification Standard.	<p><u>Rote</u> = The learner can remember information. The learner can define, identify, label, state, list, match, or select.</p> <p><u>Understanding</u> = The learner comprehends and grasps the nature and meaning of the knowledge as it relates to flight operations. The learner can describe, generalize, paraphrase, summarize, estimate, and discuss. The knowledge is used as the basis of explaining risk management and aeronautical decision making.</p> <p><u>Application</u> = The learner uses the knowledge in actual flight operational settings. The learner can determine, chart, implement, prepare, solve, use, develop, explain, apply, relate, instruct, show, or teach. The knowledge is used as the basis of practicing risk management and aeronautical decision making.</p> <p><u>Correlation</u> = The learner associates the knowledge with previous or subsequent learning. The learner can analyze, synthesize, and evaluate. The knowledge is used as the basis of independently managing risk and making aeronautical decisions.</p>
Risk management and aeronautical decision-making	Accurately identifies risks and resolves problems. Uses the appropriate decision-making processes. Completes each task/line item while considering the risk management elements found in the applicable FAA Airman Certification Standard.	<p><u>Describe</u> = The learner can recite or repeat the hazards or risks associated with the activity, but lacks understanding about their meaning, application, and management.</p> <p><u>Explain</u> = The learner can verbally identify, describe, and understand the risks inherent in the flight scenario, but needs to be prompted to identify risks and make decisions.</p> <p><u>Practice</u> = The learner can identify, understand, and apply SRM principles to the actual flight situation. Coaching, instruction, and/or assistance quickly corrects minor deviations and errors identified by the instructor. The learner is an active decision maker.</p> <p><u>Manage-Decide</u> = The learner can correctly gather the most important data available both inside and outside the flight deck, identify possible courses of action, evaluate the risk inherent in each course of action, and make the appropriate decision. Instructor intervention is not required for the safe completion of the flight.</p>
Adherence to standard operating procedures	Identifies and applies procedures in accordance with ERAU-published operating instructions, FAA guidance material, and applicable regulations. Performs each checklist using a read/do or do/verify method as required by the ERAU SOPM.	<p><u>Describe</u> = The learner can recite or repeat the elements of the procedure, but lacks understanding about their meaning, application, and implementation.</p> <p><u>Explain</u> = The learner can verbally identify, describe, and understand the procedure's underlying concepts and principles. Errors or omissions are acceptable.</p> <p><u>Practice</u> = The learner can apply the procedure to the actual flight operational scenarios with coaching and assistance. Errors or omissions are corrected in a timely manner.</p> <p><u>Perform</u> = The learner can independently apply the procedure to the actual flight operational scenarios without errors or omissions.</p>
Aircraft flight path management	Manages the aircraft flight path through manual and automated flight controls, including appropriate use of flight management system(s) and guidance. Performs each task/line item in accordance with the skill elements found in the applicable FAA Airman Certification Standard.	<p><u>Describe</u> = The learner can recite or repeat the physical characteristics/cognitive elements of the maneuver.</p> <p><u>Explain</u> = The learner can verbally identify, describe, and understand the maneuver's underlying concepts, principles, and procedures. Uncorrected deviations from the ACS tolerances is acceptable.</p> <p><u>Practice</u> = The learner can plan and execute the maneuver with coaching and assistance to correct deviations from the ACS tolerances in a timely manner.</p> <p><u>Perform</u> = The learner can plan and execute the maneuver independently within ACS tolerances.</p>

Appendix C

Maneuver Evaluation Grade Sheets

MANEUVER EVALUATION GRADE SHEET		Randomly Assigned Identifier							
<u>End-of-Course Test</u> Instrument Airplane	Video Number 1								
<u>Objective:</u> Using the Basic Competencies and Behavioral Indicators, the check instructor will evaluate the applicant's use of knowledge, risk management and aeronautical decision making, adherence to standard operating procedures, and aircraft flight path management. The applicants performance of each competency will correlate to the elements and standards established the Instrument Airplane Rating Airman Certification Standards.									
<u>Completion Standard:</u> A minimum grade of 4 is required for each task to achieve satisfactory performance. A grade of correlates to the minimum completion standards for each task as outlined in the current FAA Instrument Airplane Rating Airman Certification Standards.									
<u>Segment 1 Tasks</u>	<u>Area of Operation</u>	<u>Grade Awarded:</u> <u>Minimum:</u>							
Instrument Flight Deck Check	Preflight Preparation	<table border="1" style="display: inline-table; border-collapse: collapse;"> <tr> <td style="width: 10%;">5</td> <td style="width: 10%;">4</td> <td style="width: 10%;">3</td> <td style="width: 10%;">2</td> <td style="width: 10%;">1</td> <td style="width: 10%;">Inc</td> <td style="width: 10%;"></td> </tr> </table>	5	4	3	2	1	Inc	
5	4	3	2	1	Inc				

MANEUVER EVALUATION GRADE SHEET		Randomly Assigned Identifier							
<u>End-of-Course Test</u> Instrument Airplane	Video Number 2								
<u>Objective:</u> Using the Basic Competencies and Behavioral Indicators, the check instructor will evaluate the applicant's use of knowledge, risk management and aeronautical decision making, adherence to standard operating procedures, and aircraft flight path management. The applicants performance of each competency will correlate to the elements and standards established the Instrument Airplane Rating Airman Certification Standards.									
<u>Completion Standard:</u> A minimum grade of 4 is required for each task to achieve satisfactory performance. A grade of correlates to the minimum completion standards for each task as outlined in the current FAA Instrument Airplane Rating Airman Certification Standards.									
<u>Segment 2 Tasks</u>	<u>Area of Operation</u>	<u>Grade Awarded:</u> <u>Minimum:</u>							
Compliance with Air Traffic Control Clearances	ATC Clearances and Procedures	<table border="1" style="display: inline-table; border-collapse: collapse;"> <tr> <td style="width: 10%;">5</td> <td style="width: 10%;">4</td> <td style="width: 10%;">3</td> <td style="width: 10%;">2</td> <td style="width: 10%;">1</td> <td style="width: 10%;">Inc</td> <td style="width: 10%;">4</td> </tr> </table>	5	4	3	2	1	Inc	4
5	4	3	2	1	Inc	4			
Instrument Flight	Flight by Reference to Instruments	<table border="1" style="display: inline-table; border-collapse: collapse;"> <tr> <td style="width: 10%;">5</td> <td style="width: 10%;">4</td> <td style="width: 10%;">3</td> <td style="width: 10%;">2</td> <td style="width: 10%;">1</td> <td style="width: 10%;">Inc</td> <td style="width: 10%;">4</td> </tr> </table>	5	4	3	2	1	Inc	4
5	4	3	2	1	Inc	4			
Departure, En Route, and Arrival Operations	Navigation Systems	<table border="1" style="display: inline-table; border-collapse: collapse;"> <tr> <td style="width: 10%;">5</td> <td style="width: 10%;">4</td> <td style="width: 10%;">3</td> <td style="width: 10%;">2</td> <td style="width: 10%;">1</td> <td style="width: 10%;">Inc</td> <td style="width: 10%;">4</td> </tr> </table>	5	4	3	2	1	Inc	4
5	4	3	2	1	Inc	4			

MANEUVER EVALUATION GRADE SHEET		Randomly Assigned Identifier							
<u>End-of-Course Test</u> Instrument Airplane	Video Number 3								
<u>Objective:</u> Using the Basic Competencies and Behavioral Indicators, the check instructor will evaluate the applicant's use of knowledge, risk management and aeronautical decision making, adherence to standard operating procedures, and aircraft flight path management. The applicants performance of each competency will correlate to the elements and standards established the Instrument Airplane Rating Airman Certification Standards.									
<u>Completion Standard:</u> A minimum grade of 4 is required for each task to achieve satisfactory performance. A grade of correlates to the minimum completion standards for each task as outlined in the current FAA Instrument Airplane Rating Airman Certification Standards.									
<u>Segment 3 Tasks</u>	<u>Area of Operation</u>	<u>Grade Awarded:</u> <u>Minimum:</u>							
Instrument Flight	Flight by Reference to Instruments	<table border="1" style="display: inline-table; border-collapse: collapse;"> <tr> <td style="width: 10%;">5</td> <td style="width: 10%;">4</td> <td style="width: 10%;">3</td> <td style="width: 10%;">2</td> <td style="width: 10%;">1</td> <td style="width: 10%;">Inc</td> <td style="width: 10%;">4</td> </tr> </table>	5	4	3	2	1	Inc	4
5	4	3	2	1	Inc	4			
Recovery from Unusual Flight Attitudes	Flight by Reference to Instruments	<table border="1" style="display: inline-table; border-collapse: collapse;"> <tr> <td style="width: 10%;">5</td> <td style="width: 10%;">4</td> <td style="width: 10%;">3</td> <td style="width: 10%;">2</td> <td style="width: 10%;">1</td> <td style="width: 10%;">Inc</td> <td style="width: 10%;">4</td> </tr> </table>	5	4	3	2	1	Inc	4
5	4	3	2	1	Inc	4			

MANEUVER EVALUATION GRADE SHEET		Randomly Assigned Identifier						
<u>End-of-Course Test</u> Instrument Airplane	Video Number 4							
<u>Objective:</u> Using the Basic Competencies and Behavioral Indicators, the check instructor will evaluate the applicant's use of knowledge, risk management and aeronautical decision making, adherence to standard operating procedures, and aircraft flight path management. The applicants performance of each competency will correlate to the elements and standards established the Instrument Airplane Rating Airman Certification Standards.								
<u>Completion Standard:</u> A minimum grade of 4 is required for each task to achieve satisfactory performance. A grade of correlates to the minimum completion standards for each task as outlined in the current FAA Instrument Airplane Rating Airman Certification Standards.								
<u>Segment 4 Tasks</u>	<u>Area of Operation</u>	<u>Grade Awarded:</u>						<u>Minimum:</u>
Compliance with Air Traffic Control Clearances	ATC Clearances and Procedures	5	4	3	2	1	Inc	4
Intercepting and Tracking Navigational Systems	Navigation Systems	5	4	3	2	1	Inc	4
Departure, En Route, and Arrival Operations	Navigation Systems	5	4	3	2	1	Inc	4

MANEUVER EVALUATION GRADE SHEET		Randomly Assigned Identifier						
<u>End-of-Course Test</u> Instrument Airplane	Video Number 5							
<u>Objective:</u> Using the Basic Competencies and Behavioral Indicators, the check instructor will evaluate the applicant's use of knowledge, risk management and aeronautical decision making, adherence to standard operating procedures, and aircraft flight path management. The applicants performance of each competency will correlate to the elements and standards established the Instrument Airplane Rating Airman Certification Standards.								
<u>Completion Standard:</u> A minimum grade of 4 is required for each task to achieve satisfactory performance. A grade of correlates to the minimum completion standards for each task as outlined in the current FAA Instrument Airplane Rating Airman Certification Standards.								
<u>Segment 5 Tasks</u>	<u>Area of Operation</u>	<u>Grade Awarded:</u>						<u>Minimum:</u>
Compliance with Air Traffic Control Clearances	ATC Clearances and Procedures	5	4	3	2	1	Inc	4
Holding Procedures	ATC Clearances and Procedures	5	4	3	2	1	Inc	4
Departure, En Route, and Arrival Operations	Navigation Systems	5	4	3	2	1	Inc	4

MANEUVER EVALUATION GRADE SHEET		Randomly Assigned Identifier						
<u>End-of-Course Test</u> Instrument Airplane	Video Number 6							
<u>Objective:</u> Using the Basic Competencies and Behavioral Indicators, the check instructor will evaluate the applicant's use of knowledge, risk management and aeronautical decision making, adherence to standard operating procedures, and aircraft flight path management. The applicants performance of each competency will correlate to the elements and standards established the Instrument Airplane Rating Airman Certification Standards.								
<u>Completion Standard:</u> A minimum grade of 4 is required for each task to achieve satisfactory performance. A grade of correlates to the minimum completion standards for each task as outlined in the current FAA Instrument Airplane Rating Airman Certification Standards.								
<u>Segment 6 Tasks</u>	<u>Area of Operation</u>	<u>Grade Awarded:</u>						<u>Minimum:</u>
Compliance with Air Traffic Control Clearances	ATC Clearances and Procedures	5	4	3	2	1	Inc	4
Intercepting and Tracking DME Arcs	Navigation Systems	5	4	3	2	1	Inc	4
Departure, En Route, and Arrival Operations	Navigation Systems	5	4	3	2	1	Inc	4

MANEUVER EVALUATION GRADE SHEET		Randomly Assigned Identifier						
<u>End-of-Course Test</u> Instrument Airplane	Video Number 7							
<u>Objective:</u> Using the Basic Competencies and Behavioral Indicators, the check instructor will evaluate the applicant's use of knowledge, risk management and aeronautical decision making, adherence to standard operating procedures, and aircraft flight path management. The applicants performance of each competency will correlate to the elements and standards established the Instrument Airplane Rating Airman Certification Standards.								
<u>Completion Standard:</u> A minimum grade of 4 is required for each task to achieve satisfactory performance. A grade of correlates to the minimum completion standards for each task as outlined in the current FAA Instrument Airplane Rating Airman Certification Standards.								
<u>Segment 7 Tasks</u>	<u>Area of Operation</u>	<u>Grade Awarded:</u>					<u>Minimum:</u>	
Compliance with Air Traffic Control Clearances	ATC Clearances and Procedures	5	4	3	2	1	Inc	4
Departure, En Route, and Arrival Operations	Navigation Systems	5	4	3	2	1	Inc	4
Non-Precision Approach (VOR with Radar Vectors)	IAP	5	4	3	2	1	Inc	4
Missed Approach	IAP	5	4	3	2	1	Inc	4

MANEUVER EVALUATION GRADE SHEET		Randomly Assigned Identifier						
<u>End-of-Course Test</u> Instrument Airplane	Video Number 8							
<u>Objective:</u> Using the Basic Competencies and Behavioral Indicators, the check instructor will evaluate the applicant's use of knowledge, risk management and aeronautical decision making, adherence to standard operating procedures, and aircraft flight path management. The applicants performance of each competency will correlate to the elements and standards established the Instrument Airplane Rating Airman Certification Standards.								
<u>Completion Standard:</u> A minimum grade of 4 is required for each task to achieve satisfactory performance. A grade of correlates to the minimum completion standards for each task as outlined in the current FAA Instrument Airplane Rating Airman Certification Standards.								
<u>Segment 8 Tasks</u>	<u>Area of Operation</u>	<u>Grade Awarded:</u>					<u>Minimum:</u>	
Compliance with Air Traffic Control Clearances	ATC Clearances and Procedures	5	4	3	2	1	Inc	4
Holding Procedures	ATC Clearances and Procedures	5	4	3	2	1	Inc	4
Departure, En Route, and Arrival Operations	Navigation Systems	5	4	3	2	1	Inc	4
Non-Precision Approach (VOR with Course Reversal)	IAP	5	4	3	2	1	Inc	4

MANEUVER EVALUATION GRADE SHEET		Randomly Assigned Identifier						
<u>End-of-Course Test</u> Instrument Airplane	Video Number 9							
<u>Objective:</u> Using the Basic Competencies and Behavioral Indicators, the check instructor will evaluate the applicant's use of knowledge, risk management and aeronautical decision making, adherence to standard operating procedures, and aircraft flight path management. The applicants performance of each competency will correlate to the elements and standards established the Instrument Airplane Rating Airman Certification Standards.								
<u>Completion Standard:</u> A minimum grade of 4 is required for each task to achieve satisfactory performance. A grade of correlates to the minimum completion standards for each task as outlined in the current FAA Instrument Airplane Rating Airman Certification Standards.								
<u>Segment 9 Tasks</u>	<u>Area of Operation</u>	<u>Grade Awarded:</u>					<u>Minimum:</u>	
Compliance with Air Traffic Control Clearances	ATC Clearances and Procedures	5	4	3	2	1	Inc	4
Departure, En Route, and Arrival Operations	Navigation Systems	5	4	3	2	1	Inc	4
Non-Precision Approach (VOR with Procedure Turn)	IAP	5	4	3	2	1	Inc	4
Missed Approach	IAP	5	4	3	2	1	Inc	4

MANEUVER EVALUATION GRADE SHEET		Randomly Assigned Identifier						
<u>End-of-Course Test</u> Instrument Airplane	Video Number 10							
<u>Objective:</u> Using the Basic Competencies and Behavioral Indicators, the check instructor will evaluate the applicant's use of knowledge, risk management and aeronautical decision making, adherence to standard operating procedures, and aircraft flight path management. The applicants performance of each competency will correlate to the elements and standards established the Instrument Airplane Rating Airman Certification Standards.								
<u>Completion Standard:</u> A minimum grade of 4 is required for each task to achieve satisfactory performance. A grade of correlates to the minimum completion standards for each task as outlined in the current FAA Instrument Airplane Rating Airman Certification Standards.								
<u>Segment 10 Tasks</u>	<u>Area of Operation</u>	<u>Grade Awarded:</u>					<u>Minimum:</u>	
Compliance with Air Traffic Control Clearances	ATC Clearances and Procedures	5	4	3	2	1	Inc	4
Departure, En Route, and Arrival Operations	Navigation Systems	5	4	3	2	1	Inc	4
Non-Precision Approach (VOR with Procedure Turn)	IAP	5	4	3	2	1	Inc	4
Approach with Loss of Primary Flight Instrument Indicators	Emergency Operations	5	4	3	2	1	Inc	4

MANEUVER EVALUATION GRADE SHEET		Randomly Assigned Identifier						
<u>End-of-Course Test</u> Instrument Airplane	Video Number 11							
<u>Objective:</u> Using the Basic Competencies and Behavioral Indicators, the check instructor will evaluate the applicant's use of knowledge, risk management and aeronautical decision making, adherence to standard operating procedures, and aircraft flight path management. The applicants performance of each competency will correlate to the elements and standards established the Instrument Airplane Rating Airman Certification Standards.								
<u>Completion Standard:</u> A minimum grade of 4 is required for each task to achieve satisfactory performance. A grade of correlates to the minimum completion standards for each task as outlined in the current FAA Instrument Airplane Rating Airman Certification Standards.								
<u>Segment 11 Tasks</u>	<u>Area of Operation</u>	<u>Grade Awarded:</u>					<u>Minimum:</u>	
Compliance with Air Traffic Control Clearances	ATC Clearances and Procedures	5	4	3	2	1	Inc	4
Departure, En Route, and Arrival Operations	Navigation Systems	5	4	3	2	1	Inc	4
Non-Precision Approach (GPS with Radar Vectors)	IAP	5	4	3	2	1	Inc	4
Missed Approach	IAP	5	4	3	2	1	Inc	4

MANEUVER EVALUATION GRADE SHEET		Randomly Assigned Identifier						
<u>End-of-Course Test</u> Instrument Airplane	Video Number 12							
<u>Objective:</u> Using the Basic Competencies and Behavioral Indicators, the check instructor will evaluate the applicant's use of knowledge, risk management and aeronautical decision making, adherence to standard operating procedures, and aircraft flight path management. The applicants performance of each competency will correlate to the elements and standards established the Instrument Airplane Rating Airman Certification Standards.								
<u>Completion Standard:</u> A minimum grade of 4 is required for each task to achieve satisfactory performance. A grade of correlates to the minimum completion standards for each task as outlined in the current FAA Instrument Airplane Rating Airman Certification Standards.								
<u>Segment 12 Tasks</u>	<u>Area of Operation</u>	<u>Grade Awarded:</u>					<u>Minimum:</u>	
Compliance with Air Traffic Control Clearances	ATC Clearances and Procedures	5	4	3	2	1	Inc	4
Holding Procedures	ATC Clearances and Procedures	5	4	3	2	1	Inc	4
Departure, En Route, and Arrival Operations	Navigation Systems	5	4	3	2	1	Inc	4
Non-Precision Approach (GPS with Course Reversal)	IAP	5	4	3	2	1	Inc	4

MANEUVER EVALUATION GRADE SHEET		Randomly Assigned Identifier						
<u>End-of-Course Test</u> Instrument Airplane	Video Number 13							
<u>Objective:</u> Using the Basic Competencies and Behavioral Indicators, the check instructor will evaluate the applicant's use of knowledge, risk management and aeronautical decision making, adherence to standard operating procedures, and aircraft flight path management. The applicants performance of each competency will correlate to the elements and standards established the Instrument Airplane Rating Airman Certification Standards.								
<u>Completion Standard:</u> A minimum grade of 4 is required for each task to achieve satisfactory performance. A grade of correlates to the minimum completion standards for each task as outlined in the current FAA Instrument Airplane Rating Airman Certification Standards.								
<u>Segment 13 Tasks</u>	<u>Area of Operation</u>	<u>Grade Awarded:</u>						<u>Minimum:</u>
Compliance with Air Traffic Control Clearances	ATC Clearances and Procedures	5	4	3	2	1	Inc	4
Departure, En Route, and Arrival Operations	Navigation Systems	5	4	3	2	1	Inc	4
Non-Precision Approach (GPS with TAA)	IAP	5	4	3	2	1	Inc	4

MANEUVER EVALUATION GRADE SHEET		Randomly Assigned Identifier						
<u>End-of-Course Test</u> Instrument Airplane	Video Number 14							
<u>Objective:</u> Using the Basic Competencies and Behavioral Indicators, the check instructor will evaluate the applicant's use of knowledge, risk management and aeronautical decision making, adherence to standard operating procedures, and aircraft flight path management. The applicants performance of each competency will correlate to the elements and standards established the Instrument Airplane Rating Airman Certification Standards.								
<u>Completion Standard:</u> A minimum grade of 4 is required for each task to achieve satisfactory performance. A grade of correlates to the minimum completion standards for each task as outlined in the current FAA Instrument Airplane Rating Airman Certification Standards.								
<u>Segment 14 Tasks</u>	<u>Area of Operation</u>	<u>Grade Awarded:</u>						<u>Minimum:</u>
Compliance with Air Traffic Control Clearances	ATC Clearances and Procedures	5	4	3	2	1	Inc	4
Departure, En Route, and Arrival Operations	Navigation Systems	5	4	3	2	1	Inc	4
Precision Approach (ILS with Radar Vectors)	IAP	5	4	3	2	1	Inc	4
Missed Approach	IAP	5	4	3	2	1	Inc	4

MANEUVER EVALUATION GRADE SHEET		Randomly Assigned Identifier						
<u>End-of-Course Test</u> Instrument Airplane	Video Number 15							
<u>Objective:</u> Using the Basic Competencies and Behavioral Indicators, the check instructor will evaluate the applicant's use of knowledge, risk management and aeronautical decision making, adherence to standard operating procedures, and aircraft flight path management. The applicants performance of each competency will correlate to the elements and standards established the Instrument Airplane Rating Airman Certification Standards.								
<u>Completion Standard:</u> A minimum grade of 4 is required for each task to achieve satisfactory performance. A grade of correlates to the minimum completion standards for each task as outlined in the current FAA Instrument Airplane Rating Airman Certification Standards.								
<u>Segment 15 Tasks</u>	<u>Area of Operation</u>	<u>Grade Awarded:</u>						<u>Minimum:</u>
Compliance with Air Traffic Control Clearances	ATC Clearances and Procedures	5	4	3	2	1	Inc	4
Intercepting and Tracking DME Arcs	Navigation Systems	5	4	3	2	1	Inc	4
Departure, En Route, and Arrival Operations	Navigation Systems	5	4	3	2	1	Inc	4
Precision Approach (ILS with Transition)	IAP	5	4	3	2	1	Inc	4

MANEUVER EVALUATION GRADE SHEET		Randomly Assigned Identifier						
<u>End-of-Course Test</u> Instrument Airplane	Video Number 16							
<u>Objective:</u> Using the Basic Competencies and Behavioral Indicators, the check instructor will evaluate the applicant's use of knowledge, risk management and aeronautical decision making, adherence to standard operating procedures, and aircraft flight path management. The applicants performance of each competency will correlate to the elements and standards established the Instrument Airplane Rating Airman Certification Standards.								
<u>Completion Standard:</u> A minimum grade of 4 is required for each task to achieve satisfactory performance. A grade of correlates to the minimum completion standards for each task as outlined in the current FAA Instrument Airplane Rating Airman Certification Standards.								
<u>Segment 16 Tasks</u>	<u>Area of Operation</u>	<u>Grade Awarded:</u>					<u>Minimum:</u>	
Compliance with Air Traffic Control Clearances	ATC Clearances and Procedures	5	4	3	2	1	Inc	4
Holding Procedures	ATC Clearances and Procedures	5	4	3	2	1	Inc	4
Departure, En Route, and Arrival Operations	Navigation Systems	5	4	3	2	1	Inc	4
Precision Approach (ILS with Course Reversal)	IAP	5	4	3	2	1	Inc	4

MANEUVER EVALUATION GRADE SHEET		Randomly Assigned Identifier						
<u>End-of-Course Test</u> Instrument Airplane	Video Number 17							
<u>Objective:</u> Using the Basic Competencies and Behavioral Indicators, the check instructor will evaluate the applicant's use of knowledge, risk management and aeronautical decision making, adherence to standard operating procedures, and aircraft flight path management. The applicants performance of each competency will correlate to the elements and standards established the Instrument Airplane Rating Airman Certification Standards.								
<u>Completion Standard:</u> A minimum grade of 4 is required for each task to achieve satisfactory performance. A grade of correlates to the minimum completion standards for each task as outlined in the current FAA Instrument Airplane Rating Airman Certification Standards.								
<u>Segment 17 Tasks</u>	<u>Area of Operation</u>	<u>Grade Awarded:</u>					<u>Minimum:</u>	
Missed Approach	IAP	5	4	3	2	1	Inc	4
Compliance with Air Traffic Control Clearances	ATC Clearances and Procedures	5	4	3	2	1	Inc	4
Holding Procedures	ATC Clearances and Procedures	5	4	3	2	1	Inc	4

MANEUVER EVALUATION GRADE SHEET		Randomly Assigned Identifier						
<u>End-of-Course Test</u> Instrument Airplane	Video Number 18							
<u>Objective:</u> Using the Basic Competencies and Behavioral Indicators, the check instructor will evaluate the applicant's use of knowledge, risk management and aeronautical decision making, adherence to standard operating procedures, and aircraft flight path management. The applicants performance of each competency will correlate to the elements and standards established the Instrument Airplane Rating Airman Certification Standards.								
<u>Completion Standard:</u> A minimum grade of 4 is required for each task to achieve satisfactory performance. A grade of correlates to the minimum completion standards for each task as outlined in the current FAA Instrument Airplane Rating Airman Certification Standards.								
<u>Segment 18 Tasks</u>	<u>Area of Operation</u>	<u>Grade Awarded:</u>					<u>Minimum:</u>	
Checking Instruments and Equipment	Postflight Procedures	5	4	3	2	1	Inc	4

Appendix D

Slide Show Used for Grading System Training

The following Microsoft PowerPoint slideshow presentation was used to support the grading system training provided to the check instructors in this study. The slideshow was adapted from a similar slideshow (ERAU, 2021) and reproduced here with permission of ERAU.

AQP and Evaluator Calibration

Introduction to
Gold Standards Training and Evaluator Calibration of
Pilot School Check Instructors

EMBRY-RIDDLE
Aeronautical University

Flight Standards

January 29, 2021

Background

Advanced Qualification Program (AQP)

- ✓ Alternative training methods and requirements
- ✓ Primarily used at air carriers
- ✓ Each program approved by the FAA
- ✓ Proficiency-based and focused on CRM
- ✓ Train using line-operational simulation (LOS)
- ✓ Line-operational evaluations (LOE) designed to solicit behavior
- ✓ Behaviors are specific, observable, and measurable
- ✓ Data driven – AQP data normally integrated with ASAP and FOQA data
- ✓ Calibration of instructors and evaluators improves rater reliability
- ✓ Rater reliability improves validity of AQP



Flight Standards – New Grading System

January 29, 2021

Pilot Schools

Pitfalls

- ✓ No formal standardization program approved by the FAA
- ✓ May not include data collection and validation processes specific to check instructors
- ✓ Check instructors rely on individual experience, bias, and judgement

What does ERAU have?

- ✓ Formal standardization and training of instructors and check instructors
- ✓ Pass rate and course completion data
- ✓ Feedback mechanisms and data tools like QMS, IPQC, ASAP, FDM
- ✓ Check instructors training and reinforcement about how to evaluate

What is ERAU missing?

- ✓ Calibration training so check instructors standardized at identifying and evaluating pilot behavior
- ✓ A grading system with enough fidelity to differentiate check instructor evaluating skill and provide data



Flight Standards – New Grading System

January 29, 2021

Research Purpose

Determine the applicability of AQP concepts to pilot school evaluation activities

Employ human resource development (HRD) to improve organizational performance

Demonstrate possibilities for improving primary flight training processes and procedures

Answer the research question:

- ✓ Can pilot school check instructors be calibrated against a gold standard to provide, reliable, accurate, valid, and consistent evaluations?



Flight Standards – New Grading System

January 29, 2021

What is a Gold Standard?

A gold standard is the referent that evaluators are calibrated against

Simply calibrating one evaluator against a group of evaluators runs the risk of miscalibration

- ✓ If the group is wrong, then the individual is wrong
- ✓ As the group changes, the referent changes
- ✓ Key term = inter-rater reliability (IRR)

Developing a “true score” or gold standard solves this problem

- ✓ Individuals and groups can be calibrated against the same standard
- ✓ The referent never changes even if the group does
- ✓ Key term = referent-rater reliability (RRR)



Flight Standards – New Grading System

January 29, 2021

Grading Problems and Needs

Current grading system vaguely defined	Clearly define each grade based on behavioral evidence
Completion standards define unit-level minimums	Completion standards define line item-level minimums
Current grading system improperly applied	Train IPs on use and as an instructional aid
Current grading system doesn't track proficiency	Learners should earn higher scores based on evidence of proficiency
Current grading system not properly implemented	Test proposed system before implementation

Flight Standards – New Grading System

January 29, 2021

Objectives of New Grading System

Anchor to Certification Standard

- ✓ Tie grading scale to appropriate certification standard in ACS or PTS

Quality of TCOs

- ✓ Specify minimum completion standard per line item based on position in training course

Progression of Learning

- ✓ Measure learner progress toward achieving certification standard based on behavioral evidence of proficiency

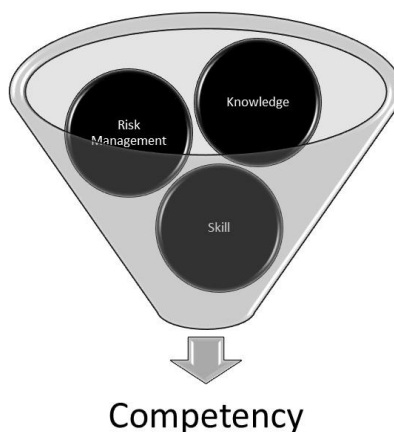


Flight Standards – New Grading System

January 15, 2021

Know, Consider, and Do

The ACS (and PTS) require applicants to demonstrate proficiency in knowledge, risk management, and skill in order to be competent at a complex task



Flight Standards – New Grading System

January 15, 2021

Levels of Learning

Aviation Instructor's Handbook – Chapter 3

Basic Levels of Learning

Flight Standards – New Grading System

January 15, 2021

Risk Management Assessment

Aviation Instructor's Handbook – Chapter 6

Explain

- The learner can verbally identify, describe, and understand the risks inherent in the flight scenario, but needs to be prompted to identify risks and make decisions.

Practice

- The learner is able to identify, understand, and apply SRM principles to the actual flight situation. Coaching, instruction, and/or assistance quickly corrects minor deviations and errors identified by the instructor. The learner is an active decision maker.

Manage-Decide

- The learner can correctly gather the most important data available both inside and outside the flight deck, identify possible courses of action, evaluate the risk inherent in each course of action, and make the appropriate decision. Instructor intervention is not required for the safe completion of the flight.

Flight Standards – New Grading System

January 15, 2021

Maneuver or Procedure Grades

Aviation Instructor's Handbook – Chapter 6

Rubric for Assessing Flight Training Maneuvers				
	Describe	Explain	Practice	Perform
Steep Turns	Pilot can describe physical characteristics/ cognitive elements of the maneuver.	Pilot can explain the maneuver's underlying concepts, principles, and procedures.	Pilot can plan and execute the maneuvers, with coaching and assistance to correct deviations and errors.	Pilot can plan and execute the maneuver to ACS standards without assistance or coaching. Pilot identifies and corrects errors and deviations.
Slow Flight				
Stalls				
Emergencies				

Flight Standards – New Grading System

January 15, 2021

Move Toward Evidence-Based Training

Competency

- ✓ A combination of knowledge, skills, and attitudes (KSAs) required to perform a complex task to a specified standard

Behavioral indicator

- ✓ An action or statement performed or made by a pilot that indicates how a job is being handled

Core competencies

- ✓ Groups of related behavioral indicators that describe how to proficiently perform a job

Evidence-based training (EBT)

- ✓ Assessment of competencies that lead to completion of a task rather than measurement of the task outcomes alone



Flight Standards – New Grading System

January 15, 2021

ERAU Basic Competencies

Competency	Description	Behavioral Indicators / Levels
Use of knowledge	Demonstrates the knowledge level required of each task/line item in accordance with the knowledge elements found in the applicable FAA Airman Certification Standard.	<p>Recite = The learner can remember information. The learner can define, identify, label, state, list, match, or select.</p> <p>Understanding = The learner comprehends and grasps the nature and meaning of the knowledge as it relates to flight operations. The learner can describe, generalize, paraphrase, summarize, estimate, and discuss. The knowledge is used as the basis of explaining risk management and aeronautical decision making.</p> <p>Application = The learner uses the knowledge in actual flight operational settings. The learner can determine, chart, implement, prepare, solve, use, develop, explain, apply, relate, instruct, show, or teach. The knowledge is used as the basis of practicing risk management and aeronautical decision making.</p> <p>Correlation = The learner associates the knowledge with previous or subsequent learning. The learner can analyze, synthesize, and evaluate. The knowledge is used as the basis of independently managing risk and making aeronautical decisions.</p>

Flight Standards – New Grading System

January 15, 2021

ERAU Basic Competencies

Competency	Description	Behavioral Indicators / Levels
Risk management and aeronautical decision-making	Accurately identifies risks and resolves problems. Uses the appropriate decision-making processes. Completes each task/line item while considering the risk management elements found in the applicable FAA Airman Certification Standard.	<p>Describe = The learner can recite or repeat the hazards or risks associated with the activity, but lacks understanding about their meaning, application, and management.</p> <p>Explain = The learner can verbally identify, describe, and understand the risks inherent in the flight scenario, but needs to be prompted to identify risks and make decisions.</p> <p>Practice = The learner can identify, understand, and apply SRM principles to the actual flight situation. Coaching, instruction, and/or assistance quickly corrects minor deviations and errors identified by the instructor. The learner is an active decision maker.</p> <p>Manage-Decide = The learner can correctly gather the most important data available both inside and outside the flight deck, identify possible courses of action, evaluate the risk inherent in each course of action, and make the appropriate decision. Instructor intervention is not required for the safe completion of the flight.</p>

Flight Standards – New Grading System

January 15, 2021

ERAU Basic Competencies

Competency	Description	Behavioral Indicators / Levels
Adherence to standard operating procedures	Identifies and applies procedures in accordance with ERAU-published operating instructions, FAA guidance material, and applicable regulations. Performs each checklist using a read/do or do/verify method as required by the ERAU SOPM.	<p><u>Describe</u> = The learner can recite or repeat the elements of the procedure, but lacks understanding about their meaning, application, and implementation.</p> <p><u>Explain</u> = The learner can verbally identify, describe, and understand the procedure's underlying concepts and principles. Errors or omissions are acceptable.</p> <p><u>Practice</u> = The learner can apply the procedure to the actual flight operational scenarios with coaching and assistance. Errors or omissions are corrected in a timely manner.</p> <p><u>Perform</u> = The learner can independently apply the procedure to the actual flight operational scenarios without errors or omissions.</p>

Flight Standards – New Grading System

January 15, 2021

ERAU Basic Competencies

Competency	Description	Behavioral Indicators / Levels
Aircraft flight path management	Manages the aircraft flight path through manual and automated flight controls, including appropriate use of flight management system(s) and guidance. Performs each task/line item in accordance with the skill elements found in the applicable FAA Airman Certification Standard.	<p><u>Describe</u> = The learner can recite or repeat the physical characteristics/cognitive elements of the maneuver.</p> <p><u>Explain</u> = The learner can verbally identify, describe, and understand the maneuver's underlying concepts, principles, and procedures. Uncorrected deviations from the ACS tolerances is acceptable.</p> <p><u>Practice</u> = The learner can plan and execute the maneuver with coaching and assistance to correct deviations from the ACS tolerances in a timely manner.</p> <p><u>Perform</u> = The learner can plan and execute the maneuver independently within ACS tolerances.</p>

Flight Standards – New Grading System

January 15, 2021

Grading Scale

Each line item will be graded by the instructor using a scale from 1 to 5 or as incomplete (I).

The grading scale will be anchored to the applicable Airman Certification Standards (ACS) or Practical Test Standards (PTS).

The grading scale uses specific metrics to evaluate:

- ✓ Knowledge
- ✓ Risk management
- ✓ Procedures
- ✓ Skills



Flight Standards – New Grading System

January 15, 2021

Grading Scale – 1

1. Requisite knowledge is demonstrated; deviations from the prescribed task standards occur that are not recognized or corrected.

- ✓ Knowledge elements of the task are demonstrated at a rote level of learning.
- ✓ Risk management elements of the task are not demonstrated or can only be described.
- ✓ Procedural elements of the task can only be described.
- ✓ Skill elements of the task can only be described.



Flight Standards – New Grading System

January 15, 2021

Grading Scale – 2

2. Deviations from the prescribed task standards can be explained but not corrected.

- ✓ Knowledge elements of the task are demonstrated at an understanding level of learning.
- ✓ Risk management elements of the task can be explained.
- ✓ Procedural elements of the task can be explained; errors or omissions are acceptable.
- ✓ Skill elements of the task can be explained; deviations are acceptable.



Grading Scale – 3

3. Deviations from the prescribed task standards occur that are recognized and corrected.

- ✓ Knowledge elements of the task are demonstrated at an application level of learning.
- ✓ Risk management elements of the task are being practiced.
- ✓ Procedural elements of the task are being practiced; timely correction to errors or omissions is made.
- ✓ Skill elements of the task are being practiced; timely correction to deviations is made.



Grading Scale – 4

4. Performance remains within the prescribed task standards.

- ✓ Knowledge elements of the task are demonstrated at an application or correlation level of learning.
- ✓ Risk management elements of the task are being practiced or can be managed independently.
- ✓ Procedural elements of the task are performed without error or omission.
- ✓ Skill elements of the task are performed without deviation.



Flight Standards – New Grading System

January 15, 2021

Grading Scale – 5

5. Performance remains within the prescribed task standards; cognitive abilities are exemplary.

- ✓ Knowledge elements of the task are demonstrated at a correlation level of learning.
- ✓ Risk management elements of the task are managed independently.
- ✓ Procedural elements of the task are performed without error or omission.
- ✓ Skill elements of the task are performed without deviation.



Flight Standards – New Grading System

January 15, 2021

Grading Scale – I

- I. (Incomplete) Line item not performed, attempted, demonstrated, or discussed.



Flight Standards – New Grading System

January 15, 2021

Course Progression

The minimum score for a particular line item is indicated on each lesson.

The minimum score will increase progressively through the course commensurate with the trainee's position in training and relative to overall preparation for the Part 141 end-of-course test and Part 61 certification practical test.



Flight Standards – New Grading System

January 15, 2021

Scoring Matrix

The grade awarded for each lesson element will be a function of the interaction among the trainee's use of knowledge, risk management and ADM, adherence to the SOPM, and aircraft flight path management using the following matrix:

		Knowledge			
		Rote	Understanding	Application	Correlation
Risk Management	Describe	1	2	2	2
	Explain	2	2	3	3
	Practice	2	3	3	4
	Manage-Decide	2	3	4	5
		Describe	Explain	Practice	Perform
		SOPM / Flight Path			

Think of the scoring matrix like this:

- ✓ Knowledge is the prerequisite
- ✓ Risk management increases safety
- ✓ Skill tempers the final score

Flight Standards – New Grading System

January 15, 2021

Whiz-Wheel Method

Procedure:

- ✓ The matrix is entered on the knowledge scale, moving to the right until reaching the demonstrated level of knowledge.
- ✓ From there, the demonstrated level of risk management is determined, identifying an initial score.
- ✓ Finally, the demonstrated level of adherence to the SOPM and proficiency in flight path management is determined, possibly decreasing or increasing the final score.

		Knowledge			
		Rote	Understanding	Application	Correlation
Risk Management	Describe	1	2	2	2
	Explain	2	2	3	3
	Practice	2	3	3	4
	Manage-Decide	2	3	4	5
		Describe	Explain	Practice	Perform
		SOPM / Flight Path			

Flight Standards – New Grading System

January 15, 2021

Example 1

Maneuvering During Slow Flight

✓ Minimum score for the maneuver is shown to be a 3

✓ The learner:

- Understands the effects of the flight controls at critically slow airspeeds
- Demonstrates proper collision avoidance procedures
- Performs the procedure in accordance with ERAU SOPM
- Maintains altitude within 150 ft.; heading within 10 degrees; airspeed within 10 KIAS; bank angle within 5 degrees; makes own timely corrections

		Knowledge			
		Rote	Understanding	Application	Correlation
Risk Management	Describe	1	2	2	2
	Explain	2	3	3	3
	Practice	3	4	4	4
	Manage-Decision	3	4	4	5
		Describe	Explain	Practice	Perform
SOPM / Flight Path					

Satisfactory

Flight Standards – New Grading System

January 15, 2021

Example 2

Emergency Approach to Landing

✓ Minimum score for the maneuver is shown to be a 4

✓ The learner:

- Applies knowledge of engine systems to troubleshoot the emergency
- Chooses an emergency landing site that contains many obstacles despite a better location being nearby but explains the error after the maneuver
- Correctly follows the emergency checklist
- Maintains best glide speed with 5 KIAS

		Knowledge			
		Rote	Understanding	Application	Correlation
Risk Management	Describe	1	2	2	2
	Explain	2	3	3	3
	Practice	3	4	4	4
	Manage-Decision	3	4	4	5
		Describe	Explain	Practice	Perform
SOPM / Flight Path					

Unsatisfactory

Flight Standards – New Grading System

January 15, 2021

Example 3

Holding

- ✓ Minimum score for the maneuver is shown to be a 3
- ✓ The learner:
 - Understands knowledge of holding patterns
 - Interprets and adjusts for wind in an attempt to remain within protected holding pattern airspace
 - Correctly explains the SOPM procedures
 - Turns the wrong direction after crossing the fix and does not realize error

		Knowledge			
		Rote	Understanding	Application	Correlation
Risk Management	Describe	1	2	2	2
	Explain	2	2	3	3
	Practice	2	3	3	4
	Manage-Decide	3	3	4	5
		Describe	Explain	Practice	Perform
SOPM / Flight Path					

Unsatisfactory

Flight Standards – New Grading System

January 15, 2021

Is There a Problem?

		Knowledge			
		Rote	Understanding	Application	Correlation
Risk Management	Describe	1	2	2	2
	Explain	2	2	3	3
	Practice	2	3	3	4
	Manage-Decide	2	3	4	5
		Describe	Explain	Practice	Perform
SOPM / Flight Path					

Flight Standards – New Grading System

January 29, 2021

The Fix

		Knowledge			
		Rote	Understanding	Application	Correlation
Risk Management	Describe	1	2	2	2
	Explain	2	2	3	3
	Practice	2	3	3	4
	Manage-Decide	2	3	4	5
		Describe	Explain	Practice	Perform
		SOPM / Flight Path			

Flight Standards – New Grading System

January 29, 2021

Grading as a Teaching Aid

Whether satisfactory or unsatisfactory, the scoring matrix can be used to assist with debriefing, critique, assigning homework or additional practice, etc.

From Example 1 on Slide 17

- ✓ While satisfactory proficiency was demonstrated, improvements in knowledge (self-study, oral lesson) or risk management (scenario-based training) will ensure ACS performance when required later in the course

From Example 2 on Slide 18

- ✓ Unsatisfactory proficiency can be corrected by improving risk management skills (scenario-based training); knowledge and skill already meet ACS performance

From Example 3 on Slide 19

- ✓ Unsatisfactory proficiency can be corrected through additional skills training (drill-and-practice); risk management made up for knowledge

Flight Standards – New Grading System

January 15, 2021

Overall Lesson Grade

An overall satisfactory (S), unsatisfactory (U), or incomplete (I) grade will be awarded.

A lesson may only be graded satisfactory if all line items are awarded the minimum score.

Each lesson completion standard may specify additional general or supplementary completion standards.



Flight Standards – New Grading System

January 15, 2021

ETA Demo



Flight Standards – New Grading System

January 15, 2021

Frequently Asked Questions

How do you grade a learner who deviates 150 feet compared to a learner who deviates 300 feet?

- ✓ The amount of deviation is not important. What matters is how the learner responds to the deviation.
- ✓ If the learner recognizes and corrects the deviation (skill level 3 – “practicing”) then the deviation should be small compared to the learner who doesn’t recognize the deviation (skill level 1 – “describing”) or can only explain but not correct the deviation (skill level 2 – “explaining”).



Flight Standards – New Grading System

January 15, 2021

Frequently Asked Questions

How do you grade line items that aren’t tasks in the ACS?

- ✓ This is the advantage of this grading system – your identifying behaviors rather than outcomes.
- ✓ Let’s use checklist usage as an example. Where are the knowledge and risk elements found about checklist usage? The SOPM contains that information.
- ✓ What are the skills for checklist usage? Completing the flows and verifying the checklist without error is the standard at ERAU (skill level 4 – “performing”). Compare that to the pre-solo learner pilot who is only required to complete the checklist as a read/do checklist and may omit some items as long as errors are corrected (skill level 3 – “practicing”).



Flight Standards – New Grading System

January 15, 2021

Frequently Asked Questions

Is there a concern that IPs will rush through or “pencil-whip” grades at the end of a lesson if short on time?

- ✓ Grading should take place during the lesson. After each task or line item, jot down the grade and short-hand a note justifying it.
- ✓ During the debrief, the pre-recorded grades and short-hand notes should be used to aid the IP in conducting the debrief and focusing on which components of each grade (knowledge, risk management, procedure, or skill) should be addressed.
- ✓ The grades should have already been determined and simply recorded permanently onto the final grade sheet with appropriate comments.



Flight Standards – New Grading System

January 15, 2021

Frequently Asked Questions

Can high-level grades be awarded during orals that occur early in training?

- ✓ In some cases yes. This would be appropriate for line items that are primarily tested during the oral portion of the EOC or practical test. Such line items normally have scenario-based skills that the learner works through. If the learner demonstrates the higher-level behaviors, then the appropriate higher-level grade can be awarded.
- ✓ However, in many cases, demonstration of skill requires a flight-related scenario to be used. In those cases, the oral lessons are designed to achieve only an understanding level of knowledge (knowledge level 2 – “understanding”) and the associated skills are later demonstrated during FTD and flight lessons.
- ✓ Remember, there is no formal separation in the ACS or PTS between the oral and flight evaluation.



Flight Standards – New Grading System

January 15, 2021

Frequently Asked Questions

The minimum score seems too high for the first time the line item is done in the airplane. Is this intentional?

- ✓ Yes, it is. Remember, the minimum score for the EOC must always be a 4, which also means the learner should achieve a score of 4 prior to the EOC test.
- ✓ If the course is short, the minimum score required for the first time the line item is practiced in flight may be a 3 instead of a 2 because there are limited opportunities for practice.
- ✓ This means that the module is more likely to be repeated should additional practice be required. However, excessive repeats on the last module before the EOC test may not be needed.



Flight Standards – New Grading System

January 15, 2021

Frequently Asked Questions

Why do minimum scores only range from 2 to 4 when the grading scale is from 1 to 5?

- ✓ Remember that a score of 1 represents that the requisite knowledge is demonstrated. This refers to the learner having completed any pre-lesson study, homework, PACE exercises, or ground school training.
- ✓ For introductory oral material, it is the instructor's job to build upon this requisite knowledge during the lesson so that the learner achieves an understanding level of knowledge or better. If the instructor is unable to help the learner do this, the grade awarded would be a 1 and the lesson would be unsatisfactory.
- ✓ For near-EOC material, the learner must demonstrate knowledge, risk management, and skill that meets the ACS, which is represented by a score of 4. However, it is possible for the learner to be better than the minimum ACS, which is a score of 5.



Flight Standards – New Grading System

January 15, 2021

Conclusion

Pay attention to whether the learner identifies and corrects his or her own errors.

Use the defined metrics and the scoring matrix to assist in determining line item grades. Focus on evidence instead of outcomes.

Record grades and short-hand notes during the lesson so they can be used during the debrief.

Use the components of each grade to assist with constructive critique and lesson planning.



Appendix E

Lesson Plan for Check Instructor Calibration

LESSON PLAN FOR CHECK INSTRUCTOR CALIBRATION

**EMBRY-RIDDLE AERONAUTICAL UNIVERSITY
Daytona Beach, Florida**

**PART I
COVER SHEET**

LESSON TITLE: Check Instructor Calibration

RESOURCE PERSON: Paul M. Cairns, Assistant Chief Flight Instructor

TEACHING METHOD: Dramatization and guided discussion

REFERENCES: FAA Instrument Rating Airplane Airman Certification Standards (ACS)

AIDS/HANDOUT/NOTETAKERS: Pre-recorded check ride maneuver segments; maneuver evaluation grade sheets

STUDENT PREPARATION/READING ASSIGNMENT: Review FAA Instrument Rating Airplane ACS

PRESENTATION TIME: 9 hours

PART IA

COGNITIVE OBJECTIVE: Apply knowledge of calibration techniques.

COGNITIVE SAMPLES OF BEHAVIOR:

1. Describe, summarize, and discuss calibration techniques and behavioral indicator grading.
2. Determine pilot performance based on standards in the FAA Instrument Airplane ACS.
3. Use behavioral indicators to accurately grade pilot performance.

AFFECTIVE OBJECTIVE: Value class discussion about the importance of check instructor calibration.

AFFECTIVE SAMPLES OF BEHAVIOR:

1. Voluntarily participates in discussion about calibration techniques.
2. Complies with use of maneuver evaluation grade sheets.
3. Accepts calibration and grading methods as appropriate for evaluation of pilot performance.

PSYCHOMOTOR OBJECTIVE: None

PSYCHOMOTOR SAMPLES OF BEHAVIOR: None

PART IB

ORGANIZATIONAL PATTERN: Cause-Effect

STRATEGY: The lesson should begin by explaining background information regarding advance qualification programs (AQP), behavioral indicators, and evaluator calibration. A simple background in statistical methods such as inter-rater reliability should be presented. After that, begin the calibration session by playing nine pre-recorded maneuver segments to the group. The segments chosen will each feature a different pilot performing maneuver segments. It may be useful during the first pool of video demonstrations to briefly discuss each scenario ahead of time to prepare the check instructors to evaluate the proper areas. Each of the check instructors will score the tasks for each maneuver segment using the provided maneuver evaluation grade sheet. A short break can take place at this point, but it must be emphasized to the check instructors not to discuss the recordings or the scores each of them recorded. Then, play another nine pre-recorded maneuver segments, which may or may not feature the same pilots but will feature different maneuver segments. It should not be necessary at this point to interject before each video. Each of the check instructors will again score the tasks for each maneuver segment using the provided maneuver evaluation grade sheet. A longer break will then take place. During the break, a statistical analysis will be performed to determine the initial levels of rater agreement, reliability, sensitivity, and accuracy. After the break, each check instructor will be provided with individual feedback about his or her scores and how they compare to the group and to the gold standard. Explain the gold standard for each maneuver segment in the videos that were used and how the group differed from that standard, focusing on the least reliable and accurate items first. Facilitate a group discussion with the specific purpose of fostering learning and emphasizing methods for more reliable and accurate observation of the required behavioral indicators. Following the discussion portion of the calibration session, play another 18 videos. The participants will score the tasks for each maneuver segment. It is important to note here that the same videos used in the beginning of the session will not be used during this portion. Instead, different videos showing the same maneuver segments will be shown. This will help to limit testing effects. Another break can take place here while a statistical analysis of each check instructor's scores is again conducted to determine a change in the level of rater agreement, reliability, sensitivity, and accuracy. Facilitate a second group discussion and complete the calibration session by drawing conclusions about the group's change in performance.

LESSON OUTLINE:

Module 1. AQP, behavioral indicators, calibration, and inter-rater reliability.

Module 2. Video pools A and B.

Module 3. Guided discussion about check instructor feedback from Module 2.

Module 4. Video pools C and D.

Module 5. Guided discussion about change in evaluator performance

PART II LESSON DEVELOPMENT

Introduction

Time Allotted: 10 minutes

ATTENTION: You are about to head off to your first airline job. You have heard about AQP before but do not really know what it is or how it applies to you. Imagine if you had a leg up on your fellow new hires and knew exactly what it was and were prepared for the style of training you are about to receive.

MOTIVATION: As a flight standards team, it is important that we be the most standardized of any group in our flight training department. It is important for each check instructor to be able to evaluate the same check ride performance in the same manner. While this may seem impossible, we can get close by understanding behavioral indicators, gold standards of performance, and calibration techniques.

OVERVIEW: Today we are going to learn what AQP, or advanced qualification programs, are and how they are used at the airlines to benefit their training departments and enhance safety. We will learn specifically about check instructor calibration, evaluator reliability from a statistical perspective (do not worry; there's no math or statistics involved on your end), and what gold standards are and how they are used in calibration. After we have covered this material and you demonstrate comprehension of the material, we will watch several videos that will allow you the opportunity to evaluate pilot performance on simulated check ride scenarios. This is necessary to achieve a baseline statistical analysis of the accuracy of each of your evaluations. After a break, we will share the results of each of your evaluations and discuss them as a group. At this point you will be shown what the gold standards are, and we will discuss where and why there are differences between your evaluations and the gold standards. Using what you learn during this discussion period, you will watch another set of videos, again evaluating pilot performance. We will then analyze your second set of evaluations. The goal is to see a statistical change in your performance and have everyone be in greater alignment with one another and the gold standard.

Body

Module 1. AQP, behavioral indicators, calibration, and inter-rater reliability.

Time Allotted: 1 hour

Instructor Actions:

1. Using supplied PowerPoint presentation, give overview of AQP, behavioral indicators, calibration, and inter-rater reliability.
2. Present and explain the maneuver evaluation grade sheets; include completed examples that show how scores are totaled; discuss examples.

3. Regularly ask questions of the participants to ensure comprehension.

Participant Actions:

1. Ask questions to further enhance comprehension of the presented material.

Transition:

1. Discuss Module 2 and the format and length of each pre-recorded video.
2. Allow for a 10-minute break to use the facilities or obtain refreshment.

Module 2. Video pools A and B.

Time Allotted: 2 hours

Instructor Actions:

1. Play each video, briefly describing the scenario beforehand.

Participant Actions:

1. Apply knowledge of behavioral indicator grading to evaluate dramatized piloting performance.
2. Complete one maneuver evaluation grade sheet for each video.
3. Ask the instructor questions for anything not understood relating to completion of the maneuver evaluation grade sheets.

Transition:

1. Introduce Module 3 and the main points of the coming discussion to get the participants thinking ahead about questions they might have.
2. Allow for a 1-hour break to eat lunch or perform whatever other duties may be required. The break may span across days for scheduling convenience.

Module 3. Guided discussion about check instructor feedback from Module 2.

Time Allotted: 1 hour

Instructor Actions:

1. Present each participant the statistical feedback of their performance from Module 2.
2. Brief the class on the gold standards from each video in Module 2 and ask the participants to openly discuss their differences from the gold standards.
3. Lead a guided discussion about the group's performance it relates to the gold standards.

Participant Actions:

1. Voluntarily participate in the guided discussion about gold standards, individual difference, and group differences.

Transition:

1. Discuss Module 4 and the format and length of each pre-recorded video.
2. Allow for a 10-minute break to use the facilities or obtain refreshment.

Module 4. Video pools C and D.**Time Allotted:** 2 hours**Instructor Actions:**

1. Play each video without any briefing beforehand.

Participant Actions:

1. Apply knowledge of behavioral indicator grading to evaluate dramatized piloting performance.
2. Complete one maneuver evaluation grade sheet for each video.
3. Comply with proper completion of the maneuver evaluation grade sheets.

Transition:

1. Introduce Module 5 and the main points of the coming discussion to get the participants thinking ahead about questions they might have.
2. Allow for a 10-minute break to use the facilities or obtain refreshment.

Module 5. Guided discussion about change in evaluator performance**Time Allotted:** 1 hour**Instructor Actions:**

1. Present each participant the statistical feedback of their performance from Module 4.
2. Brief the class on the gold standards from each video in Module 4 and ask the participants to openly discuss any differences from the gold standards.
3. Lead a guided discussion about the group's performance as it relates to gold standards.
4. Identify and emphasize ratings that improved with accuracy and attribute the change to items learned in Modules 1 and 3.

Participant Actions:

1. Voluntarily participate in the guided discussion about gold standards, individual difference, and group differences.
2. Demonstrate acceptance of calibration and grading methods as appropriate for evaluation of pilot performance.

Conclusion

Time Allotted: 20 minutes

FINAL SUMAMRY: Briefly review the results of the check instructors' performances and how they improved from the evaluations made in Module 2 to those made in Module 4. Explain how this change in performance can be attributed to the concepts discussed in Module 1, Module 3, and Module 5. Highlight and reiterate the concepts of behavioral indicators, calibration, and referent-rater reliability (RRR) during this discussion. Explain that while calibration, in the form of improved RRR, was successful, it was limited to the scenarios presented on the videos and of similar performances likely to be evaluated on actual instrument airplane check rides. Similar calibration session must take place for each scenario where increased RRR is desired. Calibration may not be transferable to different scenarios.

REMOTIVATION: Encourage flight standards check instructors to discuss calibration with other instructor pilots and explain how these calibration sessions can make the flight standards team even more standardized than it already is. Draw the relationship between check instructor calibration and improved feedback to individual instructors and the organization's quality management system. Explain how the process of calibration also allows the collection of data that can drive curriculum changes and standardization methods.

CLOSURE: Sit yourself in day one of an airline's new hire indoctrination class. Think about how prepared you will be to understand the training and evaluation methods about to be presented to you. You will be ready to hit the ground running.

Appendix F

Video Scenarios and Associated Aeronautical Information

Instrument Airplane Mock Check Ride Recordings
Absence of NOTAMs indicates none were present at time of recording

Pool A

- Video 01 – Instrument flight deck check (KFMY runway 5 at A1)
- Video 03 – Airborne near KOCF
- Video 05 – TRV hold (beginning east of TRV)
- Video 07 – KFPR VOR 14 with radar vectors and missed
- Video 09 – KFMY VOR 13 with procedure turn
- Video 11 – KSFB RNAV 9L with radar vectors and missed
- Video 13 – KDAB RNAV 25R direct PASIY then cleared for approach
- Video 15 – KDAB ILS 7L with DME arc from south
- Video 17 – KFMY ILS 5, missed, hold at alternate missed fix (CALOO) as published

Pool B

- Video 02 – Takeoff occurring from KFMY runway 5 at A1
- Video 04 – Direct NUCIS for intercepting and tracking with GPS
- Video 06 – KVRB VOR 12R join 7 DME arc from WUBUR
- Video 08 – KFPR VOR 14 with hold at TRV as published
- Video 10 – KFMY VOR 13 with procedure turn and partial panel
- Video 12 – KSFB RNAV 9R with course reversal
- Video 14 – KDAB ILS 25R with radar vectors
- Video 16 – KFMY ILS 5 with course reversal
- Video 18 – After landing in KFMY (after partial panel approach)

Pool C

- Video 01 – Instrument flight deck check (KFMY runway 5 at A1)
- Video 03 – Airborne northwest of KFMY
- Video 05 – Holding over OCF
- Video 07 – KVRB VOR 12R, break off from segment 6 for radar vectors, missed
- Video 09 – KMLB VOR 9R with procedure turn and missed
- Video 11 – KFMY RNAV 23 with radar vectors
- Video 13 – KSFB RNAV 27R from GACNO then cleared for approach
- Video 15 – KDAB ILS 25R with DME arc from north
- Video 17 – After KFMY ILS 5, missed and holding at CALOO as published alternate missed fix

Pool D

- Video 02 – Takeoff occurring from KFMY runway 5 at A1
- Video 04 – Intercept and track to CALOO (FM NDB) which is published on KFMY ILS 5
- Video 06 – KOCF ILS 36 joining DME arc as published
- Video 08 – KVRB VOR 30L, direct ZAGGA, hold as published, course reversal and shoot approach
- Video 10 – MLB VOR 9R with procedure turn and partial panel
- Video 12 – KFMY RNAV 13 with course reversal
- Video 14 – KSFB ILS 9R with radar vectors and missed
- Video 16 – KFMY ILS 5 with course reversal at CALOO
- Video 18 – After landing in KFMY (after partial panel approach)

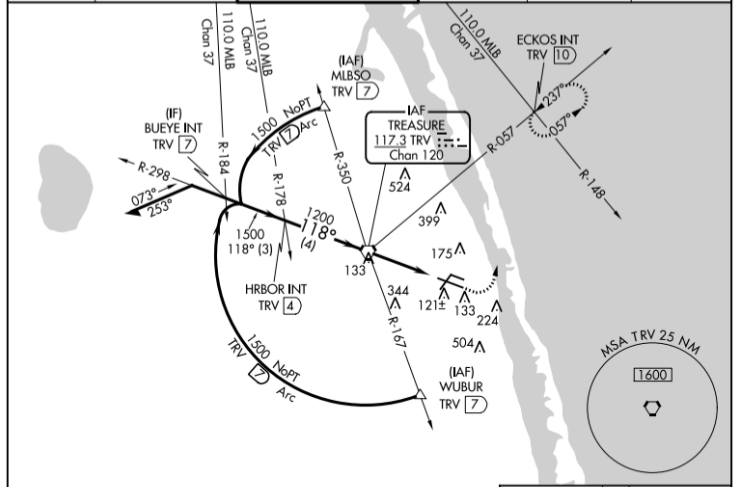
VERO BEACH, FLORIDA AL-437 (FAA) 20254
VOR RWY 12R
 VERO BEACH RGNL (VRB)

VORTAC TRV **117.3** APP CRS **118°** Rwy Idg **7314** TDZE **23**
 Chan **120** Apt Elev **24**

⚠ Circling to Rwy 30R NA at night. When local altimeter setting not received, use Fort Pierce altimeter setting and increase all MDAs 40 feet.

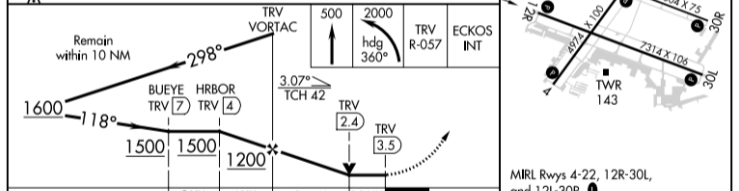
⚠ MISSED APPROACH: Climb to 500 then climbing left turn to 2000 heading 360° and on TRV VORTAC R-057 to ECKOS INT/TRV 10 DME and hold.

ATIS 120.575	PALM BEACH APP CON 123.625 225.4 (N)	VERO BEACH TOWER * 126.3 (CTAF)	GND CON 127.45	CINC DEL 134.975	UNICOM 122.95
------------------------	--	---	--------------------------	----------------------------	-------------------------



SE-3, 31 DEC 2020 to 28 JAN 2021

SE-3, 31 DEC 2020 to 28 JAN 2021



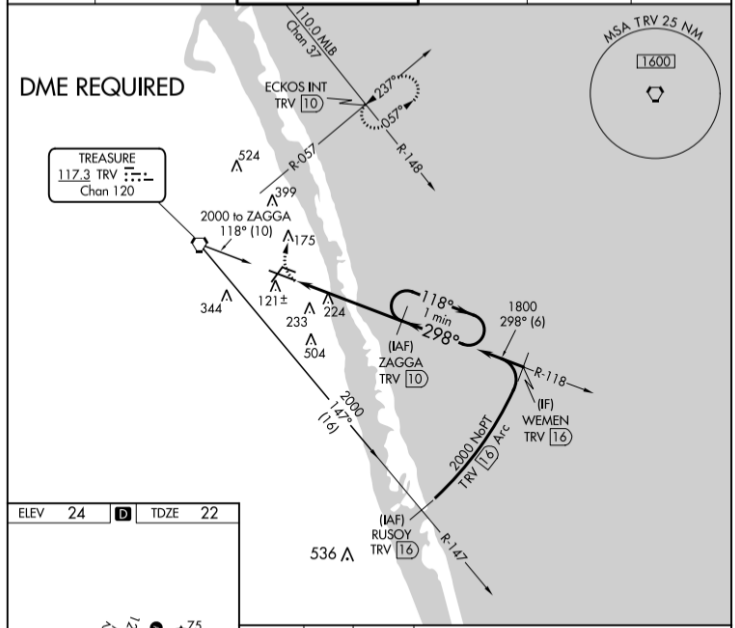
ELEV	24	TDZE	23
CATEGORY	A	B	C
S-12R	400-1 377 (400-1)		
CIRCLING	560-1 536 (600-1)	560-1½ 536 (600-1½)	660-2 636 (700-2)
	FAF to MAP 3.5 NM		
	Knots	60	90 120 150 180
	Min:Sec	3:30	2:20 1:45 1:24 1:10

VERO BEACH, FLORIDA Amdt 14D 01FEB18 27°39'N-80°25'W VERO BEACH RGNL (VRB) **VOR RWY 12R**

VERO BEACH, FLORIDA AL-437 (FAA) 20254
VOR RWY 30L
 VERO BEACH RGNL (V.R.B.)

VORTAC TRV 117.3 Chan 120 APP CRS 298° Rwy Idg 7276 TDZE 22 Apt Elev 24
 MISSED APPROACH: Climbing right turn to 2000 on heading 360° and TRV VORTAC R-057 to ECKOS INT/TRV 10 DME and hold.

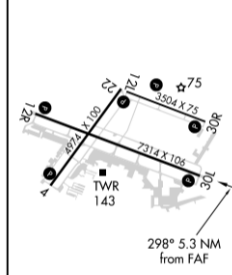
⚠ Circling to Rwy 30R NA at night. When local altimeter setting not received, use Fort Pierce altimeter setting and increase all MDAs 40 feet. VDP NA with Fort Pierce altimeter setting. DME required.
 ⚠ ATIS 120.575 PALM BEACH APP CON 123.625 225.4 (N) VERO BEACH TOWER * 126.3 (CTAF) GND CON 127.45 CLNC DEL 134.975 UNICOM 122.95



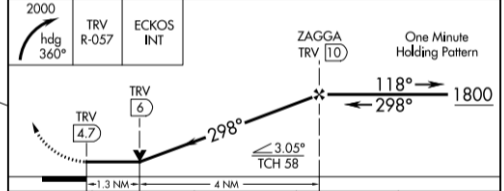
SE-3, 31 DEC 2020 to 28 JAN 2021

SE-3, 31 DEC 2020 to 28 JAN 2021

ELEV 24 TDZE 22



MRL Rwy 4-22, 12R-30L, and 12L-30R
 REIL Rwy 4, 12R, 22 and 30L



CATEGORY	A	B	C	D
S-30L	500-1 478 (500-1)		500-1 478 (500-1)	
CIRCLING	560-1 536 (600-1)		560-1 536 (600-1)	660-2 636 (700-2)

VERO BEACH, FLORIDA Amdt 4C 01FEB18 27°39'N-80°25'W VERO BEACH RGNL (V.R.B.) **VOR RWY 30L**

FORT MYERS, FLORIDA

AL-154 (FAA)

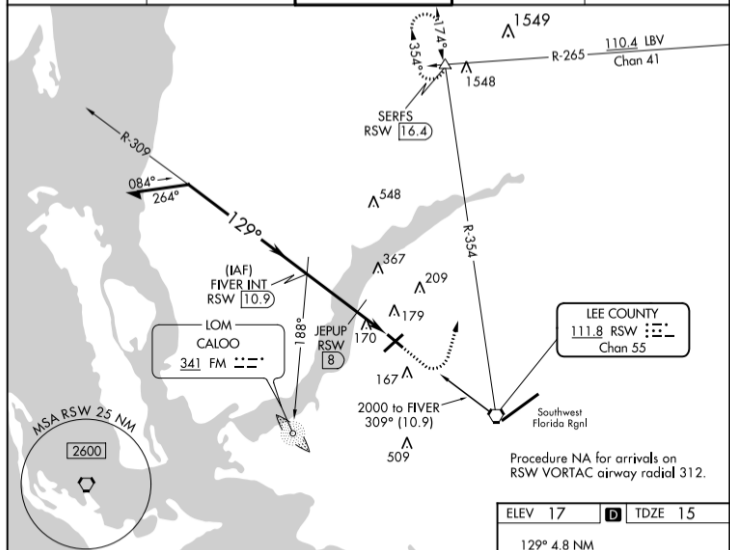
20198

VORTAC RSW 111.8 Chan 55	APP CRS 129°	Rwy Idg TDZE Apt Elev 4297 15 17
---------------------------------------	------------------------	--

VOR RWY 13
PAGE FIELD (FMY)

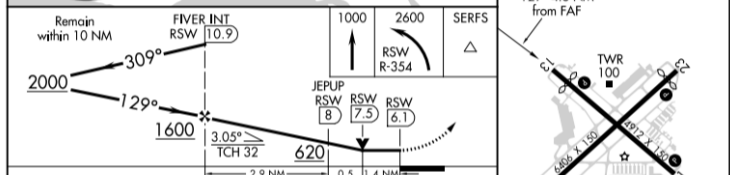
ADF or DME required.
 ▼ When Circling to Rwy 31 at night, operational VGSI required, remain on or above VGSI glidepath until threshold.
 MISSED APPROACH: Climb to 1000 then climbing left turn to 2600 on RSW VORTAC R-354 to SERFS INT/RSW 16.4 DME and hold.

ATIS 123.725	FORT MYERS APP CON* 126.8 306.2	PAGE TOWER* 119.0 (CTAF) 306.95	GND CON 121.7	CLNC DEL 121.7
------------------------	---	---	-------------------------	--------------------------

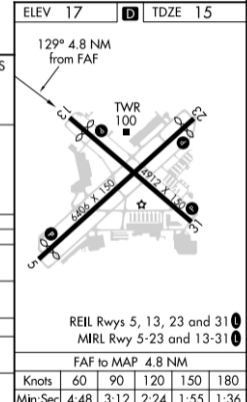


SE-3, 31 DEC 2020 to 28 JAN 2021

SE-3, 31 DEC 2020 to 28 JAN 2021



CATEGORY	A	B	C	D
S-13	620-1 605 (700-1)	620-1 605 (700-1)	620-1 605 (700-1)	620-1 605 (700-1)
CIRCLING	620-1 603 (700-1)	620-1 603 (700-1)	620-1 603 (700-1)	620-1 603 (700-1)
JEPUP FIX MINIMUMS				
S-13	480-1 465 (500-1)	480-1 465 (500-1)	480-1 465 (500-1)	480-1 465 (500-1)
CIRCLING	540-1 523 (600-1)	540-1 523 (600-1)	540-1 523 (600-1)	540-1 523 (600-1)



FORT MYERS, FLORIDA
Amdt 1 12SEP19
26°35'N-81°52'W

PAGE FIELD (FMY)
VOR RWY 13

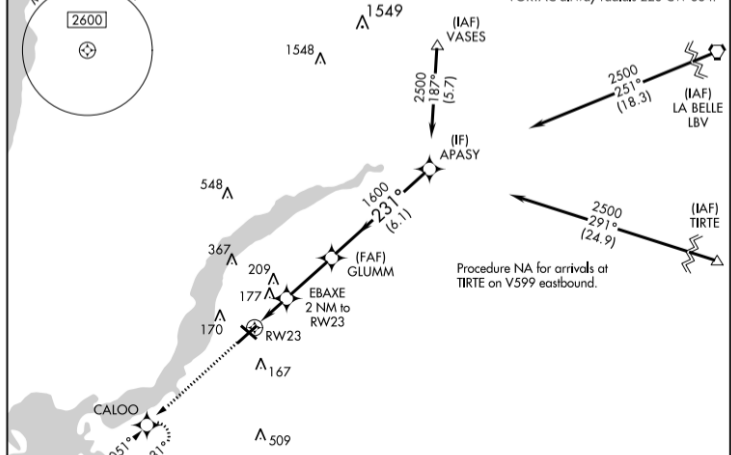
ELEV 17	TDZE 15
129° 4.8 NM from FAF	
TWR 100	
REIL Rwy 5, 13, 23 and 31	
MRL Rwy 5-23 and 13-31	
FAF to MAP 4.8 NM	
Knots	60 90 120 150 180
Min:Sec	4:48 3:12 2:24 1:55 1:36

FORT MYERS, FLORIDA AL-154 (FAA) 20058
RNAV (GPS) RWY 23
 PAGE FIELD (FMY)

WAAS CH **69424** APP CRS **231°** Rwy Idg **6007**
W23A TDZE **16** Apt Elev **17**
 RNP APCH

When Circling to Rwy 31 at night, operational VCSI required, remain on or above VCSI glidepath until threshold. Rwy 23 helicopter visibility reduction below ¼ SM NA. For uncompensated Baro-VNAV systems, LNAV/VNAV NA below 4°C or above 54°C.
 MISSED APPROACH: Climb to 2000 direct CALOO and hold.

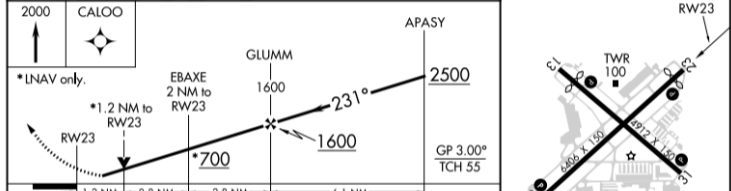
ATIS **123.725** FORT MYERS APP CON **126.8 306.2** PAGE TOWER **119.0 (CTAF) 306.95** GND CON **121.7** CLNC DEL **121.7**



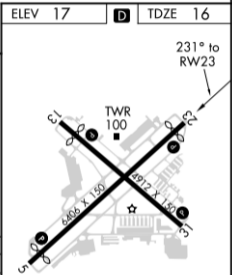
SE-3, 31 DEC 2020 to 28 JAN 2021

SE-3, 31 DEC 2020 to 28 JAN 2021

ELEV 17 TDZE 16



CATEGORY	A	B	C	D
LPV DA		274-¾	258 (300-¾)	
LNAV/VNAV DA		377-1	361 (400-1)	
LNAV MDA	440-1	424 (500-1)	440-1¼	424 (500-1¼)
CIRCLING	540-1	523 (600-1)	600-1½ 583 (600-1½)	680-2 663 (700-2)



FORT MYERS, FLORIDA PAGE FIELD (FMY)
 Amdt 1 12SEP19 26°35'N-81°52'W **RNAV (GPS) RWY 23**

FORT MYERS, FLORIDA AL-154 (FAA) 20058
RNAV (GPS) RWY 13
 PAGE FIELD (FMY)

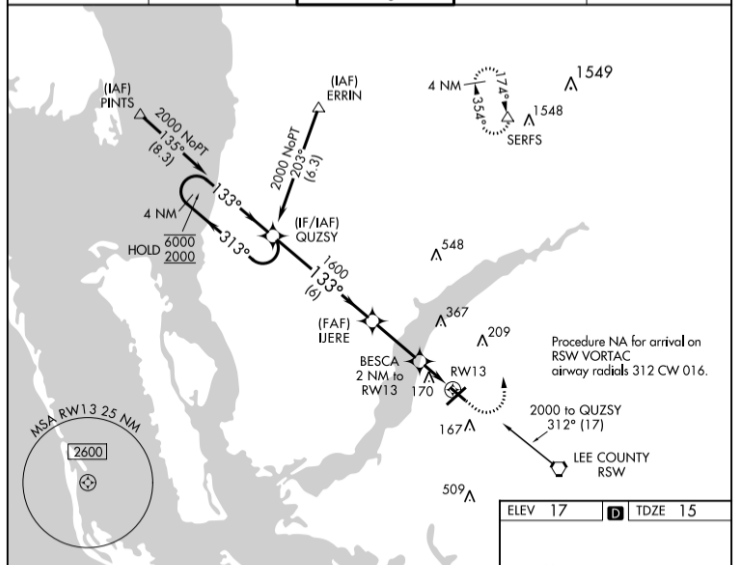
WAAS CH 73024 W13A	APP CRS 133°	Rwy Idg 4297	TDZE 15
		Apt Elev 17	

RNP APCH
 When Circling to Rwy 31 at night, operational VGSI required, remain on or above VGSI glidepath until threshold. For uncompensated Baro-VNAV systems, LNAV/VNAV NA below -15°C or above 54°C.
 MISSED APPROACH: Climb to 500 then climbing left turn to 2600 direct SERFS and hold.

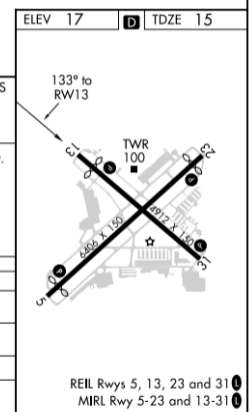
ATIS 123.725	FORT MYERS APP CON * 126.8 306.2	PAGE TOWER * 119.0 (CTAF) 306.95	GND CON 121.7	CINC DEL 121.7
-----------------	-------------------------------------	-------------------------------------	------------------	-------------------

SE-3, 31 DEC 2020 to 28 JAN 2021

SE-3, 31 DEC 2020 to 28 JAN 2021



4 NM Holding Pattern		QUZSY	500	2600	SERFS
6000	← 313°	QUZSY	1600	BESCA	2 NM to RWY 13
2000	→ 133°				
GP 3.00° TCH 30		← 133°	← 1600	* LNAV only.	
		← 6 NM	← 2.9 NM	← 0.6 NM	← 1.4 NM
CATEGORY	A	B	C	D	
LPV DA	265-1		250 (300-1)		
LNAV/VNAV DA	480-1½		465 (500-1½)		
LNAV MDA	480-1	465 (500-1)	480-1½	465 (500-1½)	
CIRCLING	540-1	523 (600-1)	600-1½ 583 (600-1½)	680-2 663 (700-2)	



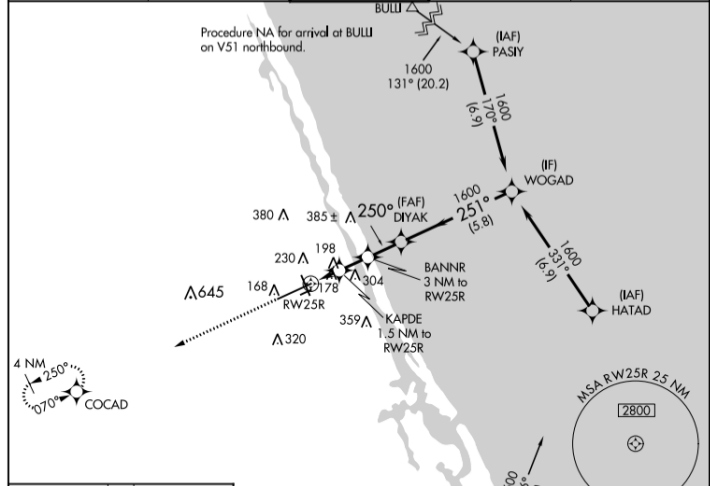
FORT MYERS, FLORIDA PAGE FIELD (FMY)
 Amdt 1F 12SEP19 26°35'N-81°52'W
RNAV (GPS) RWY 13

DAYTONA BEACH, FLORIDA AL-110 (FAA) 20254

WAAS CH 77533 W25A APP CRS 250° Rwy Idg 10293 TDZE 34 Aft Elev 34 **RNAV (GPS) RWY 25R** DAYTONA BEACH INTL (DAB)

DME/DME RNP-0.3 NA. MALSR MISSED APPROACH: Climb to 1700 direct COCAD and hold.

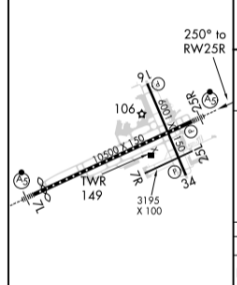
ATIS 132.875 DAYTONA APP CON 125.8 269.075 DAYTONA TOWER 120.7 257.8 GND CON 121.9 348.6 CLNC DEL 119.3



SE-3, 31 DEC 2020 to 28 JAN 2021

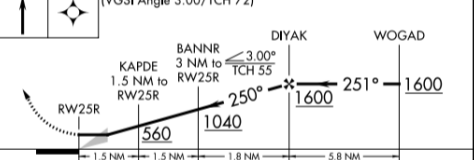
SE-3, 31 DEC 2020 to 28 JAN 2021

ELEV 34 TDZE 34



TDZ/CL Rwy 7L
HIRL Rwy 7L-25R
MIRL Rwy 7R-25L and 16-34
REIL Rwy 7R, 16, 25L and 34

1700 COCAD VGS and descent angles not coincident (VGS Angle 3.00/TCH 72)



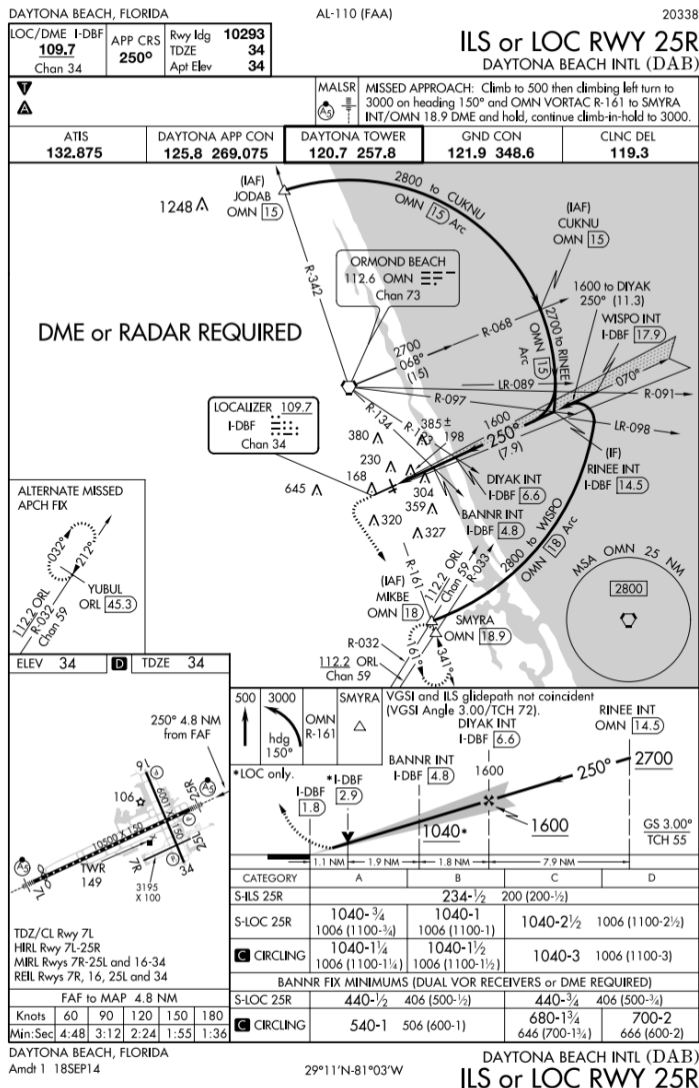
CATEGORY	A	B	C	D
LP MDA	440-1/2	407 (500-1/2)	440-3/4	407 (500-3/4)
INAV MDA	460-1/2	426 (500-1/2)	460-3/4	426 (500-3/4)
CIRCLING	540-1	506 (600-1)	680-1 3/4 646 (700-1 3/4)	700-2 666 (700-2)

DAYTONA BEACH, FLORIDA DAYTONA BEACH INTL (DAB) Amdt 4 18SEP14 29°11'N-81°03'W

RNAV (GPS) RWY 25R

DAB INSTRUMENT APPROACH PROCEDURE DAYTONA BEACH INTERNATIONAL, DAYTONA BEACH, FL. RNAV (GPS) RUNWAY 25R, AMENDMENT 4... LP MINIMUM DESCENT ALTITUDE 480/HEIGHT ABOVE TOUCHDOWN 426 ALL CATS. TEMPORARY CRANES UP TO 148 MEAN SEA LEVEL 2491FT E OF RUNWAY 25R (2020-ASO-245/249-OE). EXPIRATION ESTIMATED. FDC 0/4445

DAB AD AP WINDCONE FOR RUNWAY 25R OUT OF SERVICE. DAB 10/111



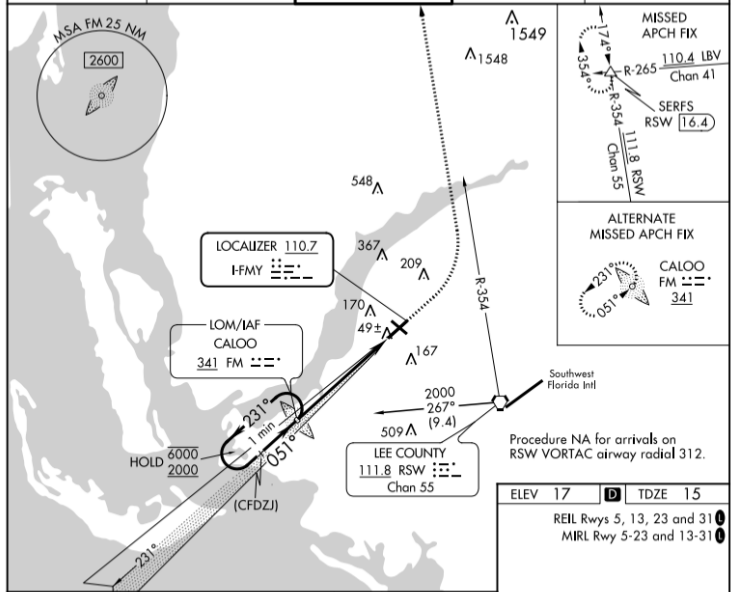
DAB INSTRUMENT APPROACH PROCEDURE DAYTONA BEACH INTERNATIONAL, DAYTONA BEACH, FL.
ILS OR LOC RUNWAY 25R, AMENDMENT 1...
 S-ILS 25R DA 393/HEIGHT ABOVE TOUCHDOWN 359 ALL CATS. BANNR FIX MINIMUMS (DUAL VOR RECEIVERS OR DISTANCE MEASURING EQUIPMENT REQUIRED) S-LOC 25R MINIMUM DESCENT ALTITUDE 460/HEIGHT ABOVE TOUCHDOWN 426 ALL CATS. VISUAL DESCENT POINT NA. TEMPORARY CRANES UP TO 148 MEAN SEA LEVEL 2491FT E OF RUNWAY 25R (2020-ASO-245/249-OE). EXPIRATION ESTIMATED. FDC 0/4449

DAB AD AP WINDCONE FOR RUNWAY 25R OUT OF SERVICE. DAB 10/111

FORT MYERS, FLORIDA AL-154 (FAA) 20086
ILS or LOC RWY 5
 PAGE FIELD (FMY)

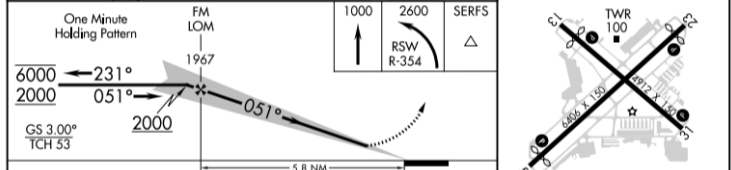
LOC I-FMY **110.7** APP CRS **051°** Rwy Idg **5947** TDZE **15** Apt Elev **17**
 ADF or RADAR required for procedure entry. ADF required for LOC only.
 When Circling to Rwy 31 at night, operational VGSI required, remain on or above VGSI glidepath until threshold. Rwy 5 helicopter visibility reduction below 3/4 SM NA.
 MISSED APPROACH: Climb to 1000 then climbing left turn to 2600 on RSW R-354 to SERFS INT/RSW 16.4 DME and hold.

ATIS **123.725** FORT MYERS APP CON * **126.8 306.2** PAGE TOWER * **119.0 (CTAF) 306.95** GND CON **121.7** CLNC DEL **121.7**



SE-3, 31 DEC 2020 to 28 JAN 2021

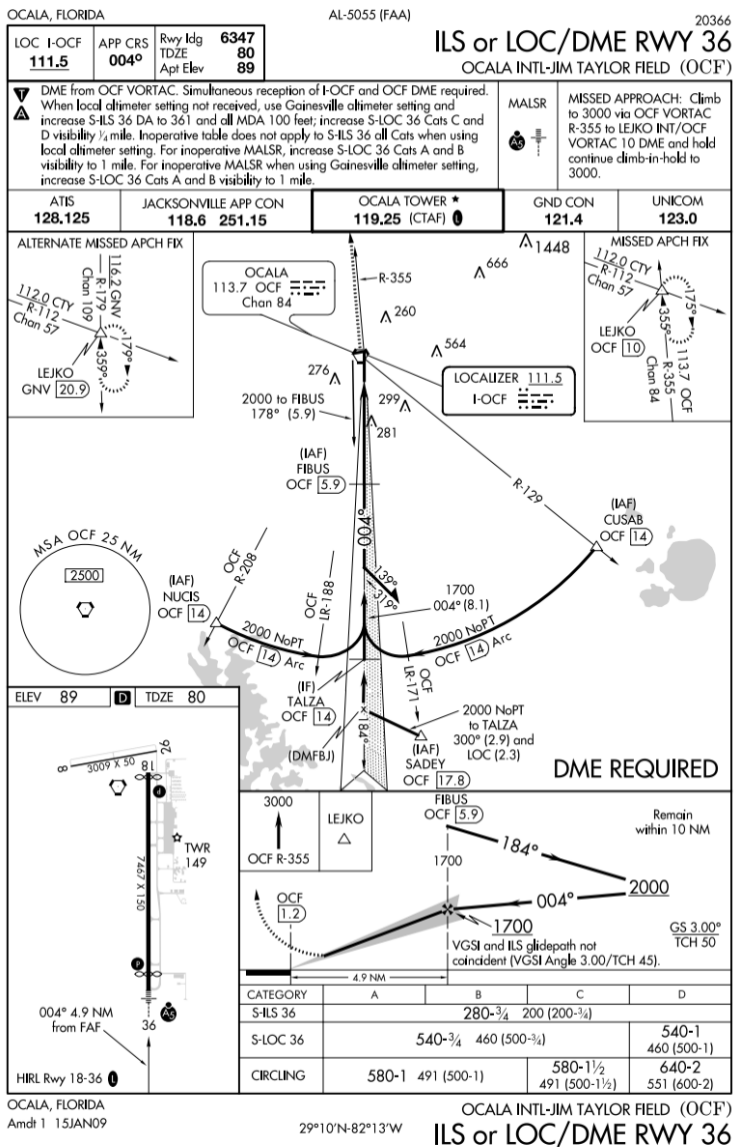
SE-3, 31 DEC 2020 to 28 JAN 2021



CATEGORY	A	B	C	D
S-ILS 5	265-1 250 (300-1)			
S-LOC 5	460-1	445 (500-1)	460-1 3/8	445 (500-1 3/8)
CIRCLING	540-1	523 (600-1)	600-1 1/2 583 (600-1 1/2)	680-2 663 (700-2)

FORT MYERS, FLORIDA Amdt 7D 12SEP19 26°35'N-81°52'W
 PAGE FIELD (FMY)
ILS or LOC RWY 5

ELEV 17	TDZE 15
REIL Rwys 5, 13, 23 and 31	
MRL Rwy 5-23 and 13-31	
TWR 100	
051° 5.8 NM from FAF	
FAF to MAP 5.8 NM	
Knots	60 90 120 150 180
Min:Sec	5:48 3:52 2:54 2:19 1:56



OCF NAVIGATION ILS RUNWAY 36 LOC/GP OUT OF SERVICE. OCF 01/021

OCF RUNWAY 36 TOUCHDOWN ZONE MARKINGS NOT STD. OCF 12/006

OCF INSTRUMENT APPROACH PROCEDURE OCALA INTERNATIONAL-JIM TAYLOR FIELD, OCALA, FL. ILS OR LOC/DISTANCE MEASURING EQUIPMENT RUNWAY 36, AMENDMENT 1...

CIRCLING CATEGORY A/B MINIMUM DESCENT ALTITUDE 600/HEIGHT ABOVE AIRPORT 510, CATEGORY C MINIMUM DESCENT ALTITUDE 760/HEIGHT ABOVE AIRPORT 670, CATEGORY D MINIMUM DESCENT ALTITUDE 920/HEIGHT ABOVE AIRPORT 830, VISIBILITY CATEGORY C 1 3/4, CATEGORY D 2 3/4.

CHANGE NOTE TO READ: WHEN LOCAL ALTIMETER SETTING NOT RECEIVED, USE GAINESVILLE ALTIMETER SETTING: INCREASE DA TO 361 FEET; INCREASE ALL MDAS 100 FEET AND VISIBILITY CATS C AND D 1/2 SM.

ALTERNATE MINIMUMS: ILS STANDARD; LOC STANDARD EXCEPT CATEGORY D 900-2 3/4; ILS AND LOC NA WHEN LOCAL WEATHER NOT AVAILABLE.

APT ELEVATION 90. EXPIRATION ESTIMATED. FDC 0/6524

SE-3, 31 DEC 2020 to 28 JAN 2021

SE-3, 31 DEC 2020 to 28 JAN 2021

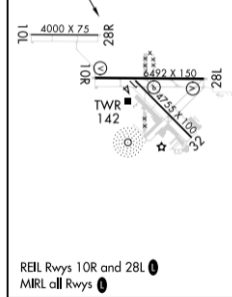
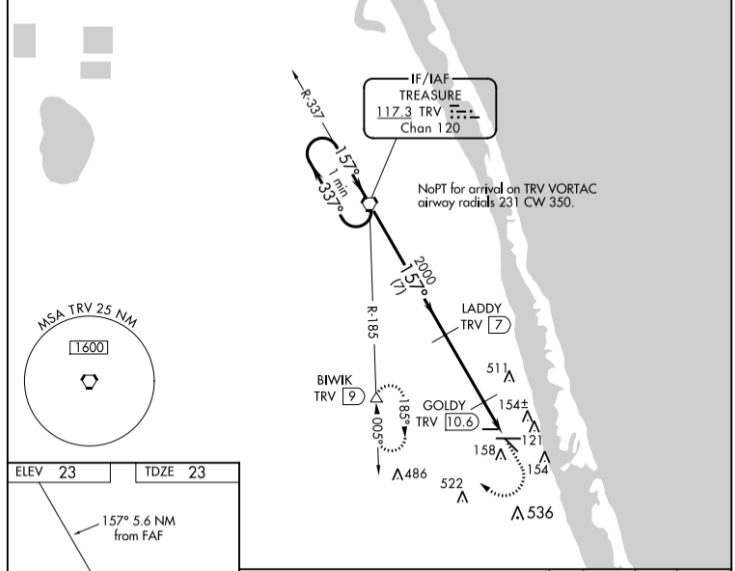
FORT PIERCE, FLORIDA AL-5343 (FAA) 20170
VOR/DME RWY 14
 TREASURE COAST INTL (FPR)

VORTAC TRV **117.3** APP CRS **157°** Rwy Idg **4755**
 Chan **120** TDZE **23** Apt Elev **23**

When local altimeter setting not received, use Vero Beach altimeter setting and increase all MDA 40 feet; increase S-14 visibility Cts C/D ¼ SM; increase Circling visibility C/D ¼ SM. Helicopter visibility reduction below 1 SM NA. Night Landing Rwy 14 NA.

MISSED APPROACH: Climb to 800 then descending right turn to 2600 on heading 270° and TRV VORTAC R-185 to BIWIK/TRV 9 DME and hold.

ATIS **134.825** PALM BEACH APP CON **132.8** FORT PIERCE TOWER* **128.2** (CTAF) **0** GND CON **119.55**



VGSI and descent angles not coincident (VGSI Angle 3.00/TCH 46).

One Minute Holding Pattern

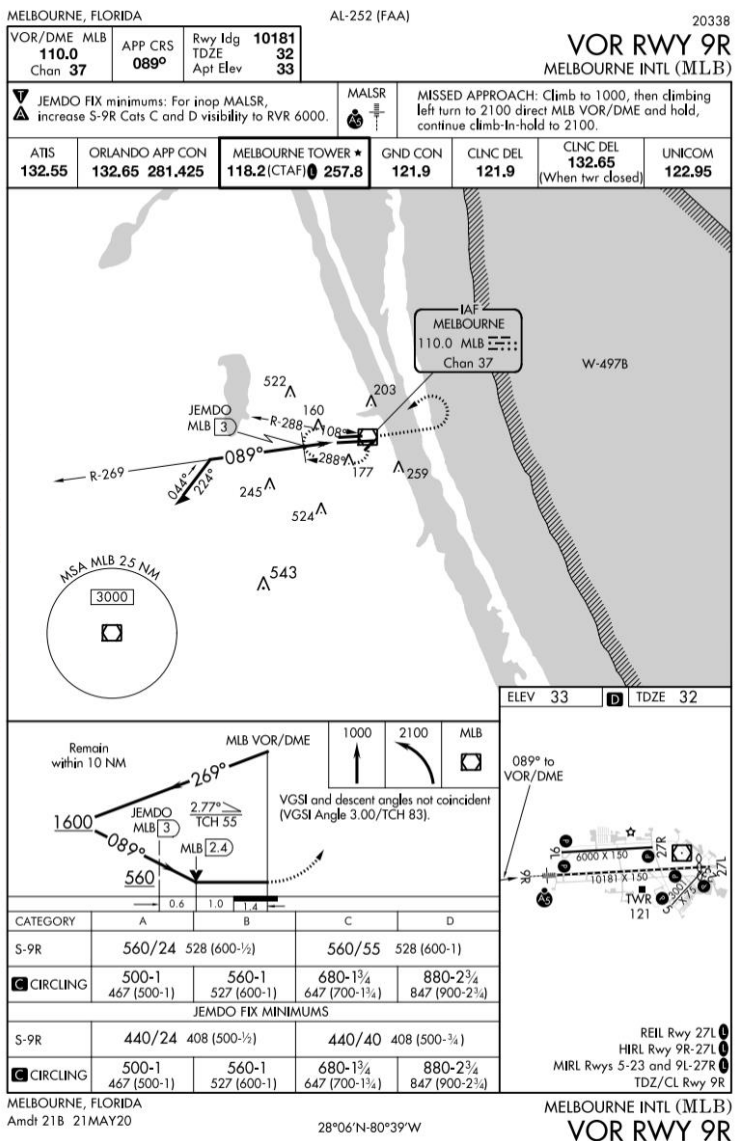
TRV VORTAC	LADY TRV (7)	GOLDY TRV (10.6)	TRV (12.6)	BIWIK
2000	2000	780	800	2600
337°	157°	3.40°	hdg 270°	TRV R-185
7 NM	3.6 NM	2 NM		

CATEGORY	A	B	C	D
S-14	420-1	397 (400-1)	420-1½	397 (400-1½)
CIRCLING	460-1 437 (500-1)	480-1 457 (500-1)	880-2½ 857 (900-2½)	880-2¾ 857 (900-2¾)

SE-3, 31 DEC 2020 to 28 JAN 2021

SE-3, 31 DEC 2020 to 28 JAN 2021

FORT PIERCE, FLORIDA TREASURE COAST INTL (FPR)
 Amdt 9D 18JUN20 27°30'N-80°22'W **VOR/DME RWY 14**

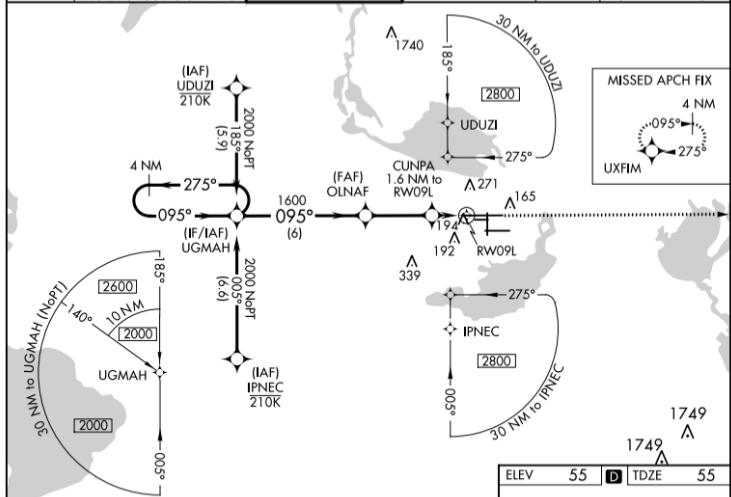


MLB INSTRUMENT APPROACH PROCEDURE
 MELBOURNE INTERNATIONAL, MELBOURNE, FL.
 VOR RUNWAY 9R, AMENDMENT 21B...
 S-9R MINIMUM DESCENT ALTITUDE 620/HEIGHT ABOVE TOUCHDOWN 588 ALL CATS, VISIBILITY CATS C/D RVR 6000. CIRCLING CATS
 A/B/C MINIMUM DESCENT ALTITUDE 620/HEIGHT ABOVE AIRPORT 587, VISIBILITY CATEGORY C 1 3/4. JEMDO FIX MINIMUMS: NA.
 TEMPORARY CRANES 252 MEAN SEA LEVEL 4884FT SE OF RUNWAY 9R
 (2019-ASO-3519/20/22/23/24-NRA). EXPIRATION ESTIMATED. FDC 0/6252

ORLANDO, FLORIDA AL-917 (FAA) 20254
RNAV (GPS) RWY 9L
 ORLANDO SANFORD INTL (SFB)

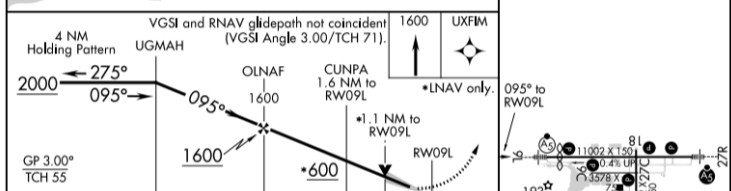
WAAS CH 78409 W09A	APP CRS 095°	Rwy Idg 10002 TDZE Apt Elev 55	MALSRS	MISSED APPROACH: Climb to 1600 direct UXFIM and hold.
---------------------------------	------------------------	--	--------	---

ATIS 125.975	ORLANDO APP CON 135.3 351.9 (NORTH) 119.775 351.9 (SOUTH)	SANFORD TOWER * 120.3 (CTAF) 0 254.35	GND CON 121.35 254.35	CLNC DEL 123.975	CLNC DEL 121.35 (when twr closed)
------------------------	---	--	---------------------------------	----------------------------	--



SE-3, 31 DEC 2020 to 28 JAN 2021

SE-3, 31 DEC 2020 to 28 JAN 2021

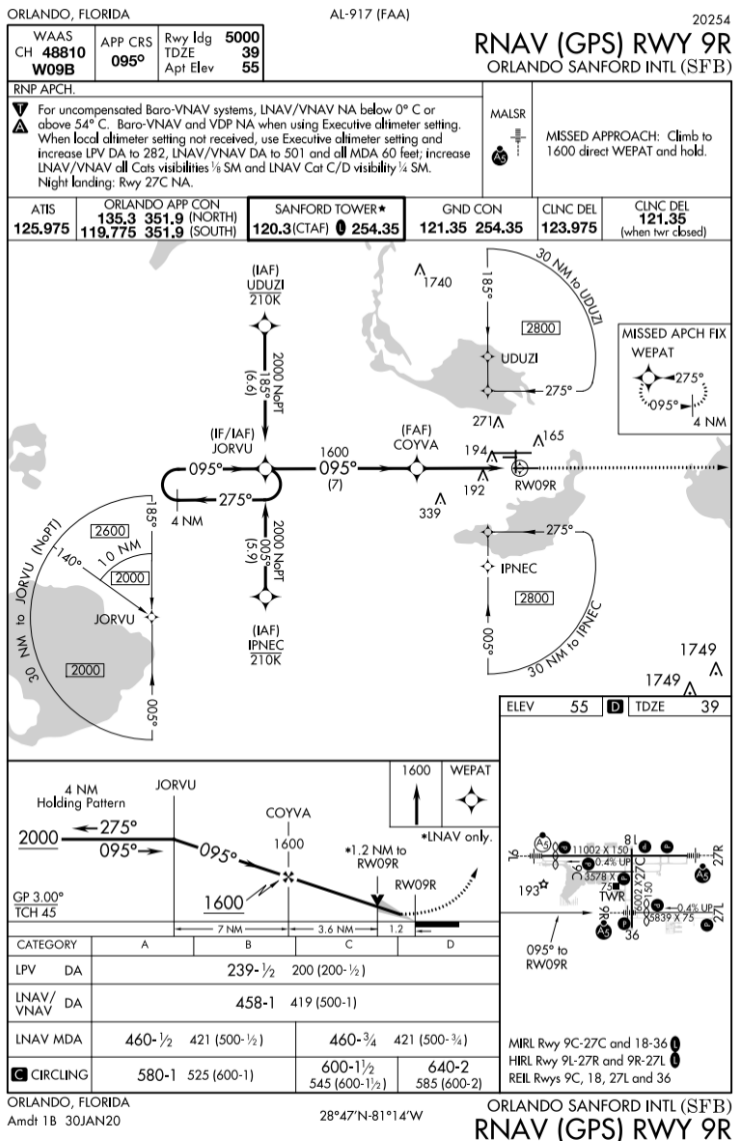


CATEGORY	A	B	C	D
LPV DA		255-1/2	200 (200-1/2)	
LNAV/VNAV DA		355-1/2	300 (300-1/2)	
LNAV MDA	460-1/2	405 (500-1/2)	460-3/4	405 (500-3/4)
CIRCLING	580-1	525 (600-1)	600-1 1/2	640-2
			545 (600-1 1/2)	585 (600-2)

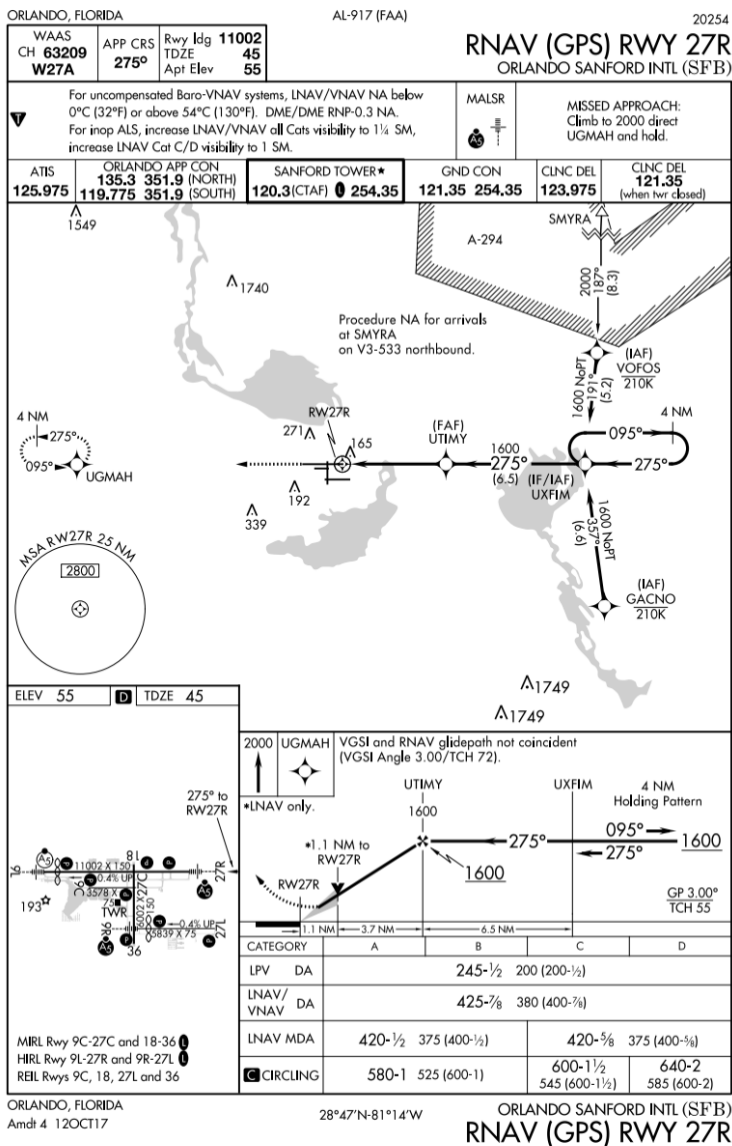
ORLANDO, FLORIDA Amdt 3B 30JAN20 28°47'N-81°14'W ORLANDO SANFORD INTL (SFB) RNAV (GPS) RWY 9L

SFB RUNWAY 18/36 CLOSED EXCEPT TAX BETWEEN TAXIWAY S AND APPROACH END RUNWAY 18. SFB 01/065

SFB RUNWAY 09L RAI LIGHT OUT OF SERVICE. EXPIRATION ESTIMATED. SFB 12/034



SFB RUNWAY 18/36 CLOSED EXCEPT TAX BETWEEN TAXIWAY S AND APPROACH END RUNWAY 18. SFB 01/065



rw SFB RUNWAY 18/36 **CLOSED** EXCEPT TAX BETWEEN TAXIWAY S AND APPROACH END RUNWAY 18. SFB 01/065

ORLANDO, FLORIDA AL-917 (FAA) 20254

LOC/DME I-OOS Chan 52 (Y)	APP CRS 095°	Rwy Idg TDZE Apt Elev	5000 39 55
-------------------------------------	------------------------	-----------------------------	---------------------------------------

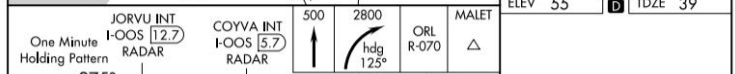
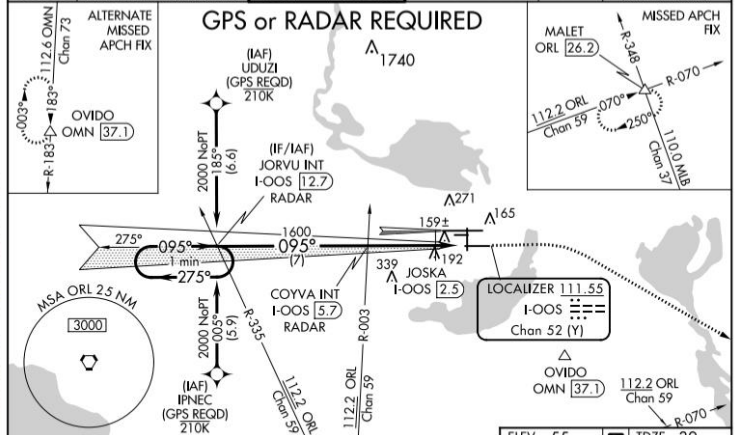
ILS or LOC RWY 9R
ORLANDO SANFORD INTL (SFB)

⚠ When local altimeter setting not received, use Executive altimeter setting and increase S-ILS DA to 282, and call MDA 60 feet; increase S-LOC 9R Cat C, D visibility $\frac{1}{4}$ mile, Circling Cat C visibility $\frac{1}{8}$ mile and JOSKA fix minimums S-LOC 9R Cats C, D visibility $\frac{1}{4}$ mile. Simultaneous approach authorized with Rwy 9L. Night landing: Rwy 27C NA.

MALSRL

MISSED APPROACH: Climb to 500 then climbing right turn to 2800 on heading 125° and ORL VORTAC R-070 to MALET INT/ ORL 26.2 DME and hold.

ATIS 125.975	ORLANDO APP CON 135.3 351.9 (NORTH) 119.775 351.9 (SOUTH)	SANFORD TOWER* 120.3 (CTAF) 0 254.35	GND CON 121.35 254.35	CLNC DEL 123.975	CLNC DEL 121.35 (when twr closed)
------------------------	---	---	---------------------------------	----------------------------	--



CATEGORY	ELEV 55 TDZE 39			
	A	B	C	D
S-ILS 9R	239- $\frac{1}{2}$ 200 (200- $\frac{1}{2}$)			
S-LOC 9R	540- $\frac{1}{2}$ 501 (500- $\frac{1}{2}$)	540-1 501 (500-1)		
CIRCLING	580-1 525 (600-1)	600-1 $\frac{1}{2}$ 545 (600-1 $\frac{1}{2}$)	640-2 585 (600-2)	
JOSKA FIX MINIMUMS (DME REQUIRED)				
S-LOC 9R	440- $\frac{1}{2}$ 401 (400- $\frac{1}{2}$)	440- $\frac{3}{4}$ 401 (400- $\frac{3}{4}$)		
CIRCLING	580-1 525 (600-1)	600-1 $\frac{1}{2}$ 545 (600-1 $\frac{1}{2}$)	640-2 585 (600-2)	

ORLANDO, FLORIDA 28°47'N-81°14'W ORLANDO SANFORD INTL (SFB) ILS or LOC RWY 9R
Amdt 1C 26MAR20

SE-3, 31 DEC 2020 to 28 JAN 2021

SE-3, 31 DEC 2020 to 28 JAN 2021

SFB RUNWAY 18/36 CLOSED EXCEPT TAX BETWEEN TAXIWAY S AND APPROACH END RUNWAY 18. SFB 01/065

DAYTONA BEACH, FLORIDA

AL-110 (FAA)

20338

LOC I-DAB	APP CRS	Rwy Idg	9810
109.7	070°	TDZE	30
		Apt Elev	34

ILS or LOC RWY 7L
DAYTONA BEACH INTL (DAB)

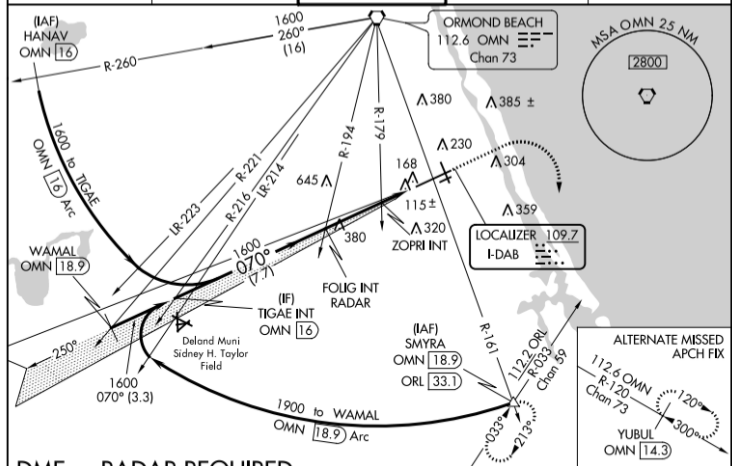
⚠ Inoperative table does not apply to S-ILS 07L. Helicopter visibility reduction below RVR 4000 NA. For inop MALSR, increase S-LOC 07L Cat A and B and ZOPRI fix minimums S-LOC 07L all Cats visibility to RVR 5500. Autopilot coupled approach NA below 535.

MALS MISSED APPROACH: Climb to 700 then climbing right turn to 3000 on heading 175° and ORL VORTAC R-033 to SMYRA/ORL 33.1 DME and hold.

ATIS	DAYTONA APP CON	DAYTONA TOWER	GND CON	CLNC DEL
132.875	125.8 269.075	120.7 257.8	121.9 348.6	119.3

SE-3, 31 DEC 2020 to 28 JAN 2021

SE-3, 31 DEC 2020 to 28 JAN 2021



DME or RADAR REQUIRED

WAMAL OMN 118.9		TIGAE INT OMN 116		FOLIG INT RADAR		ZOPRI INT		SMYRA INT	
1900		1600		1600		680		700	
GS 3.00° TCH 57		070°		175°		175°		175°	
3.3 NM		7.7 NM		2.9 NM		1.9 NM		070° 4.8 NM from FAF	
CATEGORY	A	B	C	D					
S-ILS 07L	230/40		200 (200-3/4)						
S-LOC 07L	680-40	650 (700-3/4)	680-1 3/8		650 (700-1 3/8)				
CIRCLING	680-1	646 (700-1)	680-1 7/8		700-2				
			646 (700-1 7/8)		666 (700-2)				
ZOPRI FIX MINIMUMS (DUAL VOR RECEIVERS REQUIRED)									
S-LOC 07L	380/40		350 (400-3/4)		FAF to MAP 4.8 NM				
CIRCLING	540-1	506 (600-1)	680-1 3/4		700-2				
			646 (700-1 3/4)		666 (700-2)				
DAYTONA BEACH, FLORIDA					DAYTONA BEACH INTL (DAB)				
Amdt 32A 28FEB19					29°11'N-81°03'W				

DAYTONA BEACH, FLORIDA

29°11'N-81°03'W

DAYTONA BEACH INTL (DAB)
ILS or LOC RWY 7L

Knots	60	90	120	150	180
Min:Sec	4:48	3:12	2:24	1:55	1:36