Evaluating the Suitability of the Evidence-Based Training Implementation in a Type Rating Training Course

Ing. Viktor Valenta

Thesis Supervisors:
  doc. Ing. Jakub Hospodka, Ph.D.
  doc. Ing. Bc. Vladimír Socha, Ph.D.

Type of programme: (D) doctoral
Study programme: (P3710) Technology in Transportation and Telecommunications
Branch of study / Specialization: (3708V017) Air Traffic Control and Management

Prague, 2024
Contents

1 Introduction

2 Literature Review
  2.1 Training Programs
    2.1.1 Training Path
    2.1.2 Regulatory Requirements
  2.2 Challenges in Pilot Training
  2.3 Addressing Challenges
  2.4 Competency-Based Training
  2.5 Evidence-Based Training
  2.6 Grading
    2.6.1 Purpose of grading
    2.6.2 Historical Perspectives on Grading
    2.6.3 Grading Practices and Their Variability
    2.6.4 The Predictive Value of Grades
    2.6.5 Standards-Based Grading
    2.6.6 Grading in Aviation Training
  2.7 Instructor Grading Reliability

3 Dissertation Aims
  3.1 Statement of the Problem
  3.2 Research Objectives
  3.3 Purpose Statement
  3.4 Implications

4 Methodology
  4.1 Data Sources and Analysis: Training Records
    4.1.1 Requested Data
    4.1.2 Data Types
    4.1.3 Processed Data
Chapter 0

4.2 Grading System ........................................... 37
  4.2.1 Free-Text Remarks ................................ 37
  4.2.2 Analyzing Grading ................................. 38
4.3 Instructor Grading Reliability .......................... 40

5 Results ....................................................... 42
  5.1 Trainee Progress Analysis .............................. 42
  5.2 Scenario Elements Evaluation .......................... 44
  5.3 Comparison of Outcomes from Type Ratings and Line Checks .......................... 50
  5.4 Instructor Grading Reliability .......................... 51

6 Discussion .................................................. 58
  6.1 Grading Methods ......................................... 58
  6.2 Scenario Elements Evaluation .......................... 59
  6.3 Comparison of Outcomes from Type Ratings and Line Checks .......................... 61
  6.4 Instructor Grading Reliability .......................... 62

7 Conclusions .................................................. 64

Research Work ................................................ 66

References ....................................................... 68

A Competency Tables ......................................... 82

B Comments Submitted to EASA ............................. 85

C ATO A Grading Data ........................................... 88
Abstract

Pilot training is fundamental to ensuring a high level of safety in air transport. Training programs designed primarily to meet regulatory requirements may not always provide optimum use of resources and maximum gain in safety. This thesis presents a comprehensive analysis of various aspects of a type rating training course, including syllabi, grading systems and instructor concordance. It analyzes training records from two Approved Training Organizations (ATOs). Data from ATO A show relatively high consistency, and a model reveals a statistically significant, positive progression in grades over sessions. This indicates an overall learning effect or improvement in performance, in line with the Competency-Based Training and Assessment (CBTA) philosophy. Conversely, ATO B, having recently transitioned from a traditional to a CBTA grading system, exhibited lower grading reliability, highlighting the necessity of this study. Finally, a method for ensuring instructor concordance is proposed. The findings suggest that shifting towards competency-based assessments could enhance the precision of training evaluations. Such a shift would enable instructors to objectively record trainee performance and provide more focused training. However, it is essential that tasks and technical skills are also adequately assessed in type rating courses.

Keywords: competency-based training, evidence-based training, inter-rater agreement, inter-rater reliability, instructor concordance, type rating training, pilot training
Declaration on honour

I declare that I have independently prepared this doctoral thesis entitled "Evaluating the Suitability of the Evidence-Based Training Implementation in a Type Rating Training Course" and that I have accurately cited all sources, which are listed in the attached bibliography of this thesis.

I have no significant objection to the use of this academic work under the terms of Section 60 of Act No. 121/2000 Coll., on Copyright, on Rights Related to Copyright, and on Amendment to Some Acts (Copyright Act).

In Prague, 28 March 2024

Signature

6
Chapter 1

Introduction

Air transport is a fast and reliable way of transporting both passengers and freight. Although air traffic has more than doubled in the last two decades (International Air Transport Association [IATA], 2015, 2023b), the number of major accidents did not rise (IATA, 2023a). Safety is the paramount aspect of aviation. All personnel receive extensive training in order to maintain an excellent safety record. The airline industry understands that training is an essential element of operations and not just an expense. However, there is still a strong pressure for effective use of resources. Airlines generally do not sponsor initial training for their future pilots. The cost of initial pilot training is covered by the trainees. They are not employed during the training and only after graduation from a flight school they can apply for a pilot job. If successful during assessment, the newly graduated pilots receive type rating training specific to the assigned aircraft type. The last step is an operator conversion course: newly hired pilots study company policies and participate in line flying under the supervision of a captain-instructor to get thoroughly familiar with company procedures and the area of operation. Initial training is expected to provide a sufficient basis for follow-up training.

A hiring slowdown after the Great Recession forced newly-graduated pilots to accept mediocre conditions for their first employment. Subsequent rapid airline growth led to a higher demand for pilots and shifted the focus on the so-called global pilot shortage (Caraway, 2020). The increased demand for pilots has brought new business opportunities for training organizations and has also raised concerns among airlines and regulators. Many training organizations focus only on producing more pilots and do not assess how their graduates adapt to the airline industry. Pre-COVID hiring needs had been covered due to improving terms and partially by the backlog of pilots accumulated during the slowdown. The pool of available pilots had been slowly diminishing, since the numbers of students enrolled in training
had not been rising as quickly as the demand. Initial professional pilot training usually takes between 18 to 36 months. It usually takes three to six years on-job first officer experience to upgrade to a captain position. The 2014 first officer shortage could have just become a captain shortage. A captain shortage is much more difficult to deal with. The question was for how long such a situation would be sustainable. Dramatic events caused by COVID-19 grounded the majority of world’s fleet (Kingsley-Jones, 2020) and brought pilot hiring to a halt.

Historically, the pilot job market has always been alternating between surplus and shortage. Although the US airline job market is quite unique, it may provide a valuable lesson for the global industry. In 2014, first officer starting salaries among US regional airlines were below the $25,000 (£1,500 per month) mark and many were struggling to meet the hiring needs. In 2015, Airline Pilot Association (ALPA) president Captain Tim Canoll stated, “the pilot shortage cited by US regional airlines is really a pilot pay shortage”. This statement was backed by Byrnes (2015) survey which showed: “a starting salary of the $40,000 to $45,000 range would virtually double the current regional pilot workforce”. Lutte (2017) follow-up research revealed an increase in starting salaries to an average of $51,400 (£2,700 per month) in 2017 and pointed out: “Regional airlines are generally meeting hiring needs but there is concern for sustainability of these hiring levels for the long term.”

European pilot job market is different as regional airlines do not usually supply pilots to major airlines; also, corporate and business aviation sectors are very small compared to the US. In 2015, an extensive survey of pilot employment in Europe (Jorens et al., 2015) showed that European pilots tend not to switch employers and all airlines regardless of business type can serve as pilots’ first employer. Low-cost carriers have the highest number of newly graduated pilots; however, this can be attributed to the rapid growth of this segment. With such volatile salaries, it is difficult to calculate the return on investment in training and attract talented candidates. The Byrnes (2015) study showed that only one third of accepted students at Embry-Riddle Aeronautical University (ERAU) enrolled in pilot training and approximately 65% pilot entrants completed the four-year training program. Lutte and Lovelace (2016) research revealed that 11% of the students changed their career goals and decided not to pursue airline pilot careers. The reasons given included concerns over the increased training cost, salary and lifestyle.

Pilot supply can hardly react to air travel demand as quickly as airlines would wish. The COVID-19 slowdown was an opportunity to optimize training resources so that a shortage similar to the one in the mid-2010s does not reoccur. OliverWyman (2020) survey shows that airlines are well aware of the possible challenges to ensure
the supply of qualified pilots. Boeing (2022) outlook forecasts a strong demand for more than 600,000 pilots between 2022 and 2041.

Competency-Based Training (CBT) and Evidence-Based Training (EBT) have become buzzwords in the early 2000s and gained similar popularity as Crew Resource Management (CRM) in the 1980s. The first major course built on the premises of CBT was training for the Multi-crew pilot license (MPL). The MPL was adopted by the International Civil Aviation Organization (ICAO) in 2006. It marked a shift towards a competency-based approach to training of professional pilots, focusing on preparing pilots for co-pilot duties in commercial air transportation on multi-crew airplanes from the first day of their training. This was intended to emphasize learning competencies over traditional hour-based training (Wikander & Dahlström, 2016a). European Union Aviation Safety Agency (EASA) adopted the MPL with the introduction of the "Aircrew Regulations" in 2011. The path for EBT was paved in 2013 when ICAO published Doc 9995 (International Civil Aviation Organization [ICAO], 2013). Mixed Evidence-Based Training (EBT) program as an addition to recurrent training and checking have been available for European airlines since 2015.

Innovation has not been overly affected by the COVID-19 slowdown. EASA finally approved and published rules for the implementation of a baseline EBT program in 2021. Several major European airlines immediately adopted the baseline EBT and abolished the conventional recurrent training and checking program. The question when and how to incorporate EBT data and CBT principles in the type rating training course, however, still remains open.

This work provides a detailed look into a type rating training course. Training records from two Approved Training Organizations (ATOs) and one operator are examined. Training syllabi and grading systems are assessed based on grades consistency and a method for assuring instructor concordance is proposed.
Chapter 2

Literature Review

2.1 Training Programs

Flight crew training is one of the fundamentals for ensuring a high level of safety in commercial air transport. It is difficult to define the relation between funds spent on promoting safety and trends in safety; similarly, it is problematic to quantify the effects of changes in flight crew training on flight safety. The market price of air transport has been shown to be steadily declining in the long term (Eriksson & Steenhuis, 2015). On the other hand, the cost of flight training follows the opposite direction and rises well above inflation levels (Valenta, 2018). This fact naturally leads to the build-up of strong commercial pressure on optimum use of resources during flight training (Browning, 2019). Every airline may encounter different pitfalls (e.g. areas of operation, crew experience), thus design of a training program that would perfectly match all needs of a specific airline is an extremely elaborate task. That is why airlines usually tend to build their programs in a way they simply follow all regulatory requirements.

Flight schools providing initial pilot training are put in an even more complex position; a candidate for a commercial pilot training course can pursue a career path as a second officer on ultra-longhaul flights or as a pilot in special operations such as agriculture work or aerial photography. Even though both jobs require the same pilot license, each assignment poses different challenges. The type of pilot’s first job depends on the current job market situation and it is often hard to predict at the beginning of training, which may take several years to complete. The ideal profiles of flight school graduates may be entirely different. Realistically, flight schools cannot design their training program to prepare their graduates for all available pilot jobs. Similarly to airlines, flight schools revert to building their training to meet regulatory requirements.


### 2.1.1 Training Path

![Diagram of Training Path]

A common career path to become an airline pilot in Europe is to enroll in an Airline transport pilot (ATP) integrated training course at an Approved Training Organization (ATO). An ATP integrated course lasts between 18 and 36 months and consists of at least 750 hours of theoretical knowledge instruction and 195 hours of flying training. This course is highly regulated and is designed to bring the pilot to the level of skill required for the issue of a Commercial pilot license (CPL) with a Multi-engine (ME) class rating and an Instrument Rating (IR). After a skill test, a graduate of this course is certified to operate a small multi-engine airplane (typically 4 or 6 seats) under Instrument meteorological conditions (IMC) for remuneration. The last phase of an ATP integrated course is a Multi-crew cooperation training course (MCC); it is meant to facilitate the transition from a small single-pilot airplane to a transport category multi-pilot airplane. There is very little regulatory guidance on MCC and no assessment is conducted upon its completion, thus competencies gained during this phase may vary significantly.

A pilot needs to undergo a type rating training course to be certified as a copilot on a multi-pilot multi-engine transport category airplane. A type rating training course can be provided by a different ATO than that of the ATP integrated course. A certain level of competency is expected by the ATO at the beginning of the course. Depending on the quality of the MCC training, a competency gap may

*Figure 2.1: Competency Gaps between Succeeding Training Courses*
exist prior to starting the course, as illustrated in fig. 2.1. Although this course aims to elevate a pilot to a level of competence required by an airline, its primary objective is to meet licensing requirements so the pilot can pass a skill test for the issuance of a first type rating.

Finally, the pilot joins an Operator Conversion Course (OCC) after being hired by an airline. Again, there may be a training gap caused by the difference between licensing requirements and airline operational requirements. The airline has an ability to control pilots’ entry level of competence by selecting only suitable pilots, however, when the entry level is set too high, the airline may experience a pilot shortage. On the other hand, if the entry level is set too low, training failure rate and cost will increase.

Competency gaps may be further aggravated by modular training: such as first training for a stand-alone Private pilot license (PPL) and adding on additional ratings or switching between ATOs.

2.1.2 Regulatory Requirements

Regulatory requirements on pilots training and checking had been evolving rapidly in the 1950s to reflect contemporary technology improvements and the rapid rise in air travel demand. The new licensing standards, as well as the establishment of Federal Aviation Administration (FAA) were first met with a great deal of skepticism (Duckworth, 1960). If the requirements for the issue of a Commercial pilot license in 1962 (§§ 20.40 – 20.45 (United States. Office of the Federal Register [U.S. OFR], 1962)) are compared with their current equivalent (§§ 61.121 – 61.133 (U.S. OFR, 2019)), only minor changes are discovered.

The question is whether regulations published 70 years ago still meet the requirements of the modern aviation industry.

Aircraft certified in the first half of the twentieth century achieved low levels of operational reliability. Engine failure at takeoff experienced from the cockpit of a Douglas DC-3 (manufactured between 1936 and 1942 and today still used for commercial transport in remote areas) is an extremely demanding task. The aircraft requires precise flight control inputs, a swift identification of the malfunction and immediate propeller feathering. Even a small deviation from the prescribed procedures can result in catastrophic consequences. This caused contemporary training to focus on critical technical failures. The primary objective of the training was to train a pilot with deep technical knowledge and precise stick and rudder skills.

The systems employed in modern transport aircraft are very complex; nevertheless, they achieve very high levels of operational reliability. Malfunctions rarely
occur and often do not even influence the safety of the flight. Takeoff with an engine failure experienced by the crew of an Airbus A350 (manufactured since 2014) is almost identical to a routine takeoff without any malfunction. It is practically impossible for the pilot to develop a comprehensive understanding of modern systems since they are so complex. Such an understanding is not even desired. Modern aircraft require a different set of skills; new training objectives evolve in optimum use of automation, energy management of an aerodynamically clean aircraft, etc.

Regulations do not advance as quickly as technological developments. Old regulations cannot simply be canceled or replaced because old aircraft are still in service. Authorities adopt a conservative approach - they set requirements to achieve satisfactory levels of safety under usual circumstances. The design of the regulations is strictly reactive.

National Advisory Committee for Aeronautics (N.A.C.A, 1928) developed methods for the discovery of the root causes of air accidents and published procedures for air accident investigation in 1928. As early as 1937 Peck (1937) noted, that "pilot error" requires deeper analysis than prescribed by N.A.C.A. It was not until 1948 when Davis et al. (1948) published research on pilot errors based on observations on a simple aircraft simulator and produces data usable for improving pilot training. Review of contemporary manuals (such as (Anderson, 1940; Bryan, 1958; Montana Aeronautics Commission, 1946; National Association of State Aviation Officials, AOPA Foundation Inc., 1955)) shows that the training goal was to form an ideal pilot who never makes any mistakes.

In the late 1970s, a fundamental issue in pilot training was recognized: missing education in human factors. A series of major catastrophic crashes (Salt Lake, 1977; Tenerife, 1977; Portland, 1978) revealed sequences of errors leading to the pilots with ideal profiles. Stakeholders accept that it is not possible to train a pilot who never makes any mistakes; it is important to train pilots on effective use of resources, teach pilots to recognize, accept and fix their mistakes (Cooper et al., 1980; Smith, 1979). Airlines adopted Crew Resource Management training programs in the 1980s (Jeppesen, 1996). The benefit of these programs is confirmed by Helmreich (1989) after only several years. The following three decades do not bring any innovation in flight crew training.

2.2 Challenges in Pilot Training

Initial pilot training is based on the principles developed during the first half of the twentieth century. Very strong emphasis is put on manual flying skills. This was
essential in the days prior to autopilots. The manual flying skill does not lose any of its vital importance, but other skills need to be mastered by future airline pilots too. A review of recent accidents (del Balzo, 2014; Funk et al., 1999) shows that mismanagement of automation is cited as one of the leading causes. Today’s general aviation aircraft used in basic training are equipped with avionics as advanced as the avionics in modern transport jets. Rigid training syllabi prevent the use of modern equipment to its full potential. Different levels of automation and reversion to manual flight can be practiced on many general aviation aircraft and save valuable time in expensive full flight simulators during type rating training courses. Type rating training courses should be primarily used to familiarize pilots with type-specific aircraft systems and improve certain competencies; they cannot be used to introduce new competencies since they are relatively short in duration.

European Union (EU, 2011) clearly states minimum hours for each license level. European Union Aviation Safety Agency (EASA, 2023a) recommends learning objectives for theoretical knowledge instruction in the ATPL course. ATOs should develop a training plan for each course, adhering to the principles of Instructional Systems Design (ISD). However, strict regulations and the imperative to minimize costs often lead many ATOs to focus primarily on preparing trainees for a final skill test.

It is a common phenomenon that newly graduated pilots struggle with situational awareness and application of procedures when they start their initial airline experience. However, their preparation can be improved by properly addressing the complexity of airline operations during theoretical knowledge instruction. Goyer (2011) notes that ground school training often becomes just “checking the boxes” and fails to explain the basics of real life operations.

In the past pilots often faced in flight failures and managed to learn proper workload management and decision-making without systematic guidance. Today, pilots can get through their entire career without facing a single major emergency. Technological advancement makes aircraft extremely reliable and leads to complacency. Real life emergencies are always accompanied by an element of surprise. Simulator training can familiarize pilots with system failures. Such familiarity can possibly reduce startle effects for pilots confronted with a known malfunction in flight. However, it is difficult to induce extreme pressure and life-threatening fear in the simulator environment.
2.3 Addressing Challenges

An interesting fact emerged from the Lutte (2017) study: raising experience requirements for first officers (FOQ rule)\(^1\) did not help to deliver pilots with better skills. A more experienced pilot does not necessarily mean a better pilot. Candidates are required to acquire more flight hours after initial training and prior to joining an airline. These hours are often flown without any supervision and can lead to development of bad habits; additional training after joining an airline is often required in order to reverse these bad habits (Lutte, 2017).

Monitoring of the training effectiveness is required to improve training. However, assessing the degree to which training is successful in producing suitably qualified pilots is challenging. The vast majority of training organizations lose track of their students after their final skill test. Simply assessing success rate at skill tests does not provide enough data to adjust training syllabi. Leoff (2018) estimates that “50% of pilots graduating from the current EASA pilot training [are] assessed by airlines [as having] insufficient competencies for employment.” Kalavsky et al., 2017 survey of training methods was able to show that 80% of respondents identified deficiencies in the initial training. However, no attempt was made to differentiate between the various stages of training. The survey still supports the belief that there is significant room for improvement throughout the training.

As stated above, initial pilot training must be fairly versatile because commercial license pilot holders may want to seek a job at crop dusting as well as at an airline; and obviously, each career path will require a slightly different set of competencies. Training organizations could look for feedback from operators which employ their graduates, however, this does not occur. Such an initiative is time-consuming, expensive, hampered by privacy laws and does not bring immediate results. Some major airlines subcontract or even run their own training organization to overcome these challenges. Many airline affiliated training programs were suspended in the early 2000s as parts of cost-cutting measures. In the late 2010s such programs started reappearing and had a great potential of delivering training which was tailored to the needs of airlines. The sudden drop in air travel demand caused by COVID-19 has again suspended these programs.

\(^1\)As a result of the Buffalo Crash, FAA adopted a new first officer qualification (FOQ) rule. From 2014, first officers in the US are required to hold an ATP or R-ATP certificate. This caused a major change to US flight training programs which now often include instructor training and instructor placement to reach the required experience level of 1,500 hours total time for an ATP certificate or 1,000 hours for R-ATP certificate for college graduates who possess a bachelor’s degree with an aviation major from an approved school.
Chapter 2

Multi-crew pilot license (MPL) training is an initiative that allows training organizations and airlines to design a tailored program without relying on any previous training. MPL has met with mixed reactions – it was seen by many stakeholders as a desperate attempt to reduce training time and cost, which however was not true. MPL programs are as long as integrated ATPL programs and are often even more expensive due to the use of Full Flight Simulator.

2.4 Competency-Based Training

Competency-Based Training (CBT) is an approach to education and training that focuses on what trainees can actually do as a result of instruction. Brown (1994) warns there is a lot of ambiguity in defining CBT, however, it is generally agreed that CBT differs from conventional educational approaches, which typically focus on time-based instruction and theoretical understanding, as CBT is centered on competencies: specific skills, knowledge, and attitudes that are directly related to successful performance on the job (Lilly, 1979).

The key features of CBT include a focus on outcomes, flexibility in learning pathways, learner-centered approaches, and a strong connection to industry needs. This approach allows for personalized learning experiences, as trainees can progress at their own pace, demonstrating competency as they achieve mastery of specific skills. The possible benefits of CBT are wide. It enhances the relevance of education to the labor market, improves trainees engagement by linking learning to real-world applications, and provides a clear framework for assessing performance against predefined standards. Additionally, CBT can facilitate lifelong learning and adaptability by focusing on the development of transferable skills. Despite its advantages, implementing CBT poses significant challenges. These include the need for substantial investment in the development of competency standards, training of instructors in CBT methodologies, and ensuring that assessments accurately measure competencies. There is also the challenge of ensuring that CBT does not narrow the focus of education to the immediate needs of employers at the expense of broader educational goals.

CBT offers a focused path for skill acquisition relevant to specific professional requirements. However, its application in various sectors requires a tailored approach that considers the unique challenges and expectations of each field. Brown (1994) agrees with Wheeler (1993) that it is vitally important to follow an appropriate curriculum model when developing CBT. First, competency standards must be defined and describe the job. Second, learning outcomes and assessment criteria
shall be specified and last, chose the methods for training delivery so it is not only effective but also accounts for previous learning.

CBT has been highlighted as a positive influence in professional preparation, particularly in teacher education, despite not being a comprehensive solution to all educational challenges Lilly (1979). The dominance of CBT in medical education has raised questions about its effectiveness, especially concerning measurable accountability and the impact on trainee roles and patient care Brightwell and Grant (2012). McMullen et al. (2023) CBT is widely used in pharmacy education in high income countries. Empirical support exists for competency-based training’s effectiveness in human service programs, offering guidelines for managers to implement CBT effectively (Ricciardi, 2005).

CBT is obviously not a new training philosophy. Although European aviation community faced CBT for the first time in 2006 during implementation of MPL. Australian Civil Aviation Safety Authority introduced the requirement for CBT as early as 1999. Over a decade later, in their groundbreaking analysis of CBT implementation in Australia, Todd and Thomas (2013) concluded that competency-based system was still not well developed.

MPL training program was to be built from scratch with a strong emphasis on CBT. In a survey of experience with MPL Wikander and Dahlström (2016b) discovered that the CBT implementation was only a partial success. Although MPL managed to improve some competencies, it raised additional concerns about the inability to simulate high stress and fear scenarios which can be experienced in real life. Surprisingly, many instructors show resistance to use of CBT. This may be attributed to a general public’s view, that MPL and possibly CBT are meant to reduce training hours – which is in a no way true.

EASA made no investments or even efforts to promote CBT. Regulatory requirements continue to specify the minimum hours to be flown in every training course and also the maximum length of the course. It is unthinkable to introduce a rating, or a license based solely on competency. Minimum hours are used to build the syllabus and estimate training costs. This is important for both training organizations and trainees. Aviation authorities use hours as an assurance that training has been conducted.

However, the minimum hours requirement is in direct contrast to the CBT philosophy. Trainees’ progress should be shown by gaining competencies not by logging hours.
2.5 Evidence-Based Training

IATA realized that rigid regulation can never fully meet the needs of the airline industry and can actually restrict the effectiveness of specific training. A concept known as Evidence-Based Training in aviation instruction has been developed: after several years of collecting evidence from investigation reports, flight data monitoring, observation flights, audits and training the Evidence-Based Training Implementation Guide (IATA, 2013) is published by IATA in 2013. ICAO revised Doc 9868 (International Civil Aviation Organization, 2014) and allowed implementation of EBT to replace recurrent assessment and training. EASA immediately allowed commercial air transport operators to implement a mixed EBT in recurrent training programs. In July 2018, a Notice of Proposed Amendment (EASA, 2018) was published by EASA to allow the full implementation of EBT, replacing recurrent training and checking. EASA finalized legislation for the use of a baseline EBT program in recurrent training and checking in June 2020 and several airlines implemented baseline EBT in 2021. Further efforts, originally planned by EASA and assigned to the UK CAA, to expand EBT to operator conversion course and initial type rating course have so far stalled.

Evidence-based practice was developed in medical science in 1960s. In 1990s, similar methods extended to education Mareš, 2009; Slavík and Siñor, 1993. However, it is still a new concept from the point of view of conservative aviation stakeholders. EBT targets competency development; it focuses on competencies necessary for safe and efficient operations of modern transport aircraft based on evidence gathered from operations, training and research (IATA, 2013). IATA’s threat and error analysis (IATA, 2014) is very detailed and clearly supports implementation of EBT. Such step allows removal of items that are not relevant to a specific operator and addition of items that may pose a bigger threat, thus optimizing training time.

It can be expected that interest in EBT will rise, nevertheless, it would be naive to expect that EBT solves all training related issues. Further research is required to verify suitability of expanding EBT to other types of training.

2.6 Grading

2.6.1 Purpose of grading

Grading is not just about assigning letters or numbers to trainees’ work; it is a complex tool that supports teaching, enhances learning, and plays a critical role in the educational journey. Grading serves as a critical feedback mechanism in education
by providing trainees with insights into their understanding of the training topic and highlighting areas for improvement. Toledo and Dubas (2017) discuss a method for evaluating trainee performance that provides feedback based on standards of learning, enhancing trainees’ information from assessments to diagnose strengths and weaknesses effectively. Similarly, Cain et al. (2022) review the deficiencies of traditional grading systems (i.e., A, B, C, etc), noting that while grades are intended to provide feedback, there are flaws in this approach, suggesting the need for redesigned grading practices to offer more accurate and beneficial feedback. Grades can also serve as a guide for instructors, indicating the effectiveness of their teaching methods and where adjustments may be needed. They can identify topics or concepts that trainees are struggling with, allowing for targeted interventions. Criteria for advancing to higher levels of education, fulfilling prerequisites, or entering specific professional field are often based on grades. In addition, grading systems help institutions maintain standards of education. They provide a measure of the quality of education that students are receiving and can influence the reputation and accreditation of educational institutions.

### 2.6.2 Historical Perspectives on Grading

The reliability of grades assigned by instructors has been a controversial issue since the early 20th century, with studies often criticizing these grades as inconsistent measures of student achievement (Brookhart et al., 2016). Early criticisms emphasized the subjective nature of grades, arguing that they failed to accurately reflect student learning. There has been a significant debate whether grades should be norm-referenced, comparing students against each other, or criterion-referenced, based on predefined standards of learning (Crooks, 1933; Kirschenbaum et al., 1971). While American high schools predominantly adhered to norm-referenced grading to facilitate college admissions through student ranking, a shift towards mastery learning and later, standards-based education, began in some elementary schools (Sawyer, 2013). This evolution in grading practices highlighted a persistent quest for more reliable and meaningful assessments of student achievement, reflecting a broader educational paradigm shift towards aligning grades with clear, measurable learning outcomes.

### 2.6.3 Grading Practices and Their Variability

Grading practices have historically shown significant variability, influenced by a range of factors from grading criteria to instructor perception. The inconsistency
in grading criteria and the absence of clearly defined standards have been identified as primary sources of variability in grades across educational settings (Ashbaugh, 1924; Brimi, 2011). Instructor leniency or severity reflects individual differences in judgment and further contributes to the variability (Silberstein, 1922; Sims, 1933).

The research has consistently highlighted the presence of between-instructor and within-instructor errors (Hulten, 1925), showing how complex is the dynamics of grading practices. While a portion of the variation in instructor-assigned grades can be attributed to traits measured by standardized tests, a significant part remains influenced by other factors (Bowers, 2011). These findings suggest that beyond academic achievement, grades also contain elements of motivation, that cannot be captured by standardized assessments.

The divergence in grading practices is not merely a reflection of subjective judgment but indicates a broader educational challenge. Cizek et al. (1995) found that significant variation may exist among teachers within the same school. Instructors navigate the complex task of grading with varying degrees of guidance, often relying on their professional judgment to balance cognitive and non-cognitive factors in assigning grades with an attempt to achieve fairness and accuracy (Brookhart et al., 2016). The weight given to different factors ranging from cognitive abilities to "academic enablers" such as effort and participation vary greatly (McMillan, 2001).

Brookhart et al. (2016) notes: "many teachers use their understanding of individual student circumstances, their instructional experience, and their perceptions of equity, consistency, accuracy, and fairness to make professional judgments, instead of solely relying on grading algorithms. This indicates that grading practices may differ within a single classroom, just as they do among teachers, with some teachers valuing this variability as a crucial aspect of accurate and fair grading, rather than viewing it as a problem".

Guskey (2009) found variations in the perception of grading between teachers at the elementary and secondary education levels. Elementary school teachers predominantly regarded grading as a communication strategy aimed at students and their parents. On the other hand, teachers in secondary education frequently utilized grading as a mechanism for maintaining discipline and managing the classroom, emphasizing student conduct and task completion.

According to the research by Brookhart et al. (2016), the following findings are presented: First, teachers idiosyncratically use a multitude of achievement and non-achievement factors in their grading practices to improve learning and motivation as well as document academic performance. Second, student effort is a key element in grading. Third, teachers advocate for students by helping them achieve high grades.
Finally, teacher judgment is an essential part of fair and accurate grading.

This high variability, while reflecting the instructors’ perception of trainee performance, also highlights the need for clearer guidelines and support in grading decisions.

2.6.4 The Predictive Value of Grades

The ability of grades to forecast significant educational success has been a focal point of educational research. Empirical studies have consistently demonstrated that grades serve as a robust predictor of critical outcomes such as high school graduation, college admission, and overall academic success (Fitzsimmons et al., 1969; Harwin L. Voss & Elliott, 1966). This predictive capacity extends beyond the realm of primary education, revealing grades as a significant indicator of future success in higher education, particularly within institutions with more inclusive admissions policies (Sawyer, 2013).

Additionally, the correlation between instructor-assigned grades and standardized test scores has been a subject of considerable analysis. Research findings suggest that approximately 25% of the variance in instructor-assigned grades can be attributed to traits comparable to those assessed by standardized tests (Bowers, 2011). This suggests that while there is some overlap between the measurements by grades and standardized tests, a substantial portion of what grades reflect is distinct, based on factors not captured by standardized testing. This distinction emphasizes the wide nature of grades, capturing not only academic achievement but also aspects of student engagement and effort that are critical to educational success.

Contemporary research has positioned teacher-assigned grades as among the most powerful predictors of a student’s risk of failing to graduate from high school (Fitzsimmons et al., 1969; Harwin L. Voss & Elliott, 1966). Subsequent investigations have reinforced the notion that, even when accounting for other variables such as attendance and behavior, grades remain a strong predictor of educational success (Ensminger & Slusarcick, 1992; Morris et al., 1991).

These insights highlight the complex role of grades in the educational system, serving not only as a measure of proficiency but also as a critical tool to identify and support trainees who require additional attention. The high importance of grades in the prediction of educational outcomes emphasizes the need for ongoing research on grading practices and their implications for trainee progress and achievement.
2.6.5 Standards-Based Grading

Standards-Based Grading (SBG) represents a paradigm shift in educational assessment; SBG focuses on evaluating students’ mastery of a subject against clearly defined benchmarks within learning objectives. Unlike traditional grading systems that often connect academic achievement with behavior and effort, SBG aims to provide a more transparent and objective measure of student proficiency in specific content areas. This approach categorizes performance into discrete ordered levels, such as below basic, basic, proficient, and advanced (Brookhart, 2012; Marzano & Heflebower, 2011).

The successful implementation of SBG practices, such as grading on standards and separating achievement grades from learning skills, heavily relies on the comprehension and endorsement of teachers, parents, and students. While many educators express their support for these grading reforms, they also acknowledge utilizing methods that blend effort, improvement, or motivation with academic accomplishments (Cox, 2011). There is a variation among teachers in the adoption of SBG practices, particularly regarding the utilization of common assessments, minimum grading regulations, the acceptance of late submissions without penalties, and permitting students to retake exams and substitute poor scores with the retake results. This variability aligns with the findings of McMunn et al. (2003), indicating that alterations in grading policies do not always translate into changes in grading practices. SBGs have been shown to provide more valuable information (Brookhart et al., 2016). Insights from interviews uncovered that even within SBG environments, some teachers still incorporate non-academic factors (e.g., attendance and participation) into grades (Howley et al., 2000).

2.6.6 Grading in Aviation Training

Grading research is not limited to education. Psychology, Medicine and Pharma face similar challenges and a large number of studies have been published. Instructor grading in aviation is still quite unique. It is very important to understand various training objectives in different stages of training.

The initial, ab initio training can take up to several years. Although the training consists of many hours of academic tuition and acquiring knowledge is essential, Practical skills must be mastered and proper attitude must be built during flight training. Initial training may be compared to other types of vocational training. The distinction is the duration of the training and the fact the pilot may be trained and assessed by dozens of different instructors throughout the course.
Chapter 2

The type rating training course is designed to familiarize the pilot with a new type of a complex aircraft. It usually follows just after the initial training, but experienced pilots attend this type of course when they transition to a different type. Type rating courses have certain prerequisites, but these are in terms of licensing (e.g. multi-engine land and instrument ratings or a specific type rating) not in terms of competencies. A certain level of pilot competencies is expected prior to the start of a type rating, however such level is not clearly specified in regulation or manuals and occasionally not even achieved by the trainees. Type rating training is relatively short and takes several weeks. Generally, training organizations attempt to limit the number instructors who train and assess an individual trainee, however it can still be about ten different instructors.

Recurrent checking is used for the verification that pilots maintain the set standard of competencies and recurrent training may enhance those competencies, however, the primary objective is often to practice those competencies that are seldom used in daily operation.

Current Practices in Aviation Training

Traditionally, maneuvers or tasks are graded. Traditional grading (e.g. (E)xcellent, (A)verage, (B)elow standard, (R)epeat etc.) is deemed somewhat unreliable. Such grading is highly dependent on an individual’s perception. This can be perfectly observed during a type rating course: when a trainee performs the very first engine failure, his or her performance is usually far from ideal. One instructor may grade such task as E – meaning: “well flown for the first time”. Another instructor may grade the very same performance as B – meaning: “flown below standard of a type rated pilot”. Instructor standardization should reduce spread in grading; however, traditional grading lacks clear definition and can only be used for statistical proposes with a great degree of caution.

The Shift Towards Competency-Based Assessment

Using CBTA, grading is focused on measuring to what extent competencies are applied by the trainee. As indicated above, traditional evaluation is highly subjective. CBT requires much better tool to measure trainees’ progress. An evaluation system with clearly defined grades must be established. Each grade represents an observable student behavior, task completion and the magnitude/level of instructor input. Harms, 2015; table 2.1 shows grades definitions based on instructors input.


<table>
<thead>
<tr>
<th>Grades</th>
<th>Instructor intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Master (5)</td>
<td>Not required, inputs for optimization welcome</td>
</tr>
<tr>
<td>Consolidate (4)</td>
<td>Coaching triggers, quick correction of minor errors</td>
</tr>
<tr>
<td>Practice (3)</td>
<td>Assistance temporary required</td>
</tr>
<tr>
<td>Apply (2)</td>
<td>Active instruction</td>
</tr>
<tr>
<td>Describe (1)</td>
<td>Demonstration and active instruction</td>
</tr>
</tbody>
</table>

Table 2.1: Competency-Based Performance Level Descriptors (Harms, 2015)

These grades resemble the Learner Centered Grading system introduced by the FAA Industry Training Standards (FITS), which has been successfully implemented by numerous flight schools across the United States (Summers et al., 2007).

**Regulatory Guidance**

According to ICAO (2013), there are eight core competencies that should be assessed based on behavioral indicators. Initially, IATA (2013) adopted this framework. However, EASA (2020) proposed the addition of a ninth competency, Application of Knowledge, which was subsequently accepted by ICAO and IATA (2024). Additionally, the "behavioral indicators" have been revised and renamed to "Observable behavior (OB)".

EU (2012) mandates that operators implementing EBT establish a competency framework. The EASA Competency Framework, recommended at the AMC level (EASA, 2023b), is such that it is highly unlikely any operator would attempt to significantly modify this framework. Table A.2 shows the current EASA Competency Framework and lists the original sequence IDs published in 2014 and used by both ATOs that provided training records for this work.

Implementing a grading system in a truly CBT environment is challenging. Instructor may be required to grade application of competencies for completion of each individual task. This approach produces an overwhelming number of grades during each training event and places high workload on the instructor. However, it may be used to precisely track trainees progress especially during MPL or type rating courses. Alternatively, application of competencies may be measured throughout the entire training event. This approach has been assessed by IATA (2013) as the preferred approach for EBT, but can also be particularly useful for line flying under supervision.

Grading solely based on competencies may offer limited insight. As such, EASA (2023b) recommends employing up to four levels of grading metrics, as de-
tailed in table 2.2.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 0 (Competent Metrics)</td>
<td>Data metrics providing the information whether the pilot(s) is (are) competent or not</td>
</tr>
<tr>
<td>Level 1 (Competency Metrics)</td>
<td>Quantifiable data from the grading system — numeric grade of the competencies (e.g., 1 to 5)</td>
</tr>
<tr>
<td>Level 2 (Observable Behaviour Metrics)</td>
<td>The instructors record predetermined OBs during the session</td>
</tr>
<tr>
<td>Level 3 (Other Metrics)</td>
<td>The instructors may record other data (e.g., abstract, specific tasks, actions, questions, etc.)</td>
</tr>
</tbody>
</table>

Table 2.2: Grading Metrics (EASA, 2023b)

Aircrew Regulations offer limited guidance on grading practices (EASA, 2023a). In 2019, GM1 to Appendix 3 was introduced, recommending the use of MCC APS grading for ATPL, CPL, and MPL training. However, the guidance provided in GM3 FCL.735.A, which includes an example of a grading system, is notably brief. It suggests a five-point ordinal scale for competency grading, akin to the EBT grading system. While a list of competencies is provided for MPL and MCC APS courses, no such list exists for ATPL and CPL training. Although recurrent training for instructors is mandated annually, the recommendations for standardization in grading practices are broadly defined. More explicit requirements exist for instructors assessing Area 100 KSA, underscoring the importance of standardized grading in this area. Notably, Aircrew Regulations do include GM for KSA grading with two examples: one employs indicators similar to those used in EBT grading, and the other utilizes descriptors. Remarkably, there is an absence of grading guidance for type rating courses.

2.7 Instructor Grading Reliability

IATA defines "Inter-rater reliability" as "The consistency or stability of scores between different raters." and notes it is a key element of EBT implementation (IATA, 2013). The guide provides extensive instructions on how to achieve good inter-rater reliability, but stays short of giving information how to monitor it. EASA decided
not to use the term inter-rater reliability simply because it does not easily translate into all EU languages and uses the term "concordance" with an equivalent meaning. EU (2012) mandates operators utilizing EBT program to establish an instructor standardization and concordance assurance program. EASA (2024) says broadly "In statistics, inter-rater reliability, inter-rater agreement, or concordance, is the degree of agreement among raters" but later expands "Different statistics may be appropriate for different types of measurement. Some options are: joint-probability of agreement, Cohen’s kappa, Scott’s pi and the related Fleiss’ kappa, inter-rater correlation, concordance correlation coefficient, and intra-class correlation."

Gisev et al. (2013) and Tinsley and Brown (2000) note it is important to differentiate between inter-rater reliability and agreement. Inter-rater reliability is sensitive only to the relative ordering of the rated objects. In contrast, inter-rater agreement represents the extent to which the different judges tend to assign exactly the same rating to each object.

Holt et al. (1997) uses two different measures: inter-rater reliability and referent reliability. The distribution of a rater’s assessments is significant as it can constrain the upper limits of inter-rater correlations and agreement. It is crucial that when assessing the same stimuli or a comparable set of stimuli, a rater maintains a consistent average evaluation with other raters. In other words, a rater should not exhibit a systematic bias towards being harsher or more lenient than the group. (Holt et al., 1997). Referent reliability differs fundamentally from inter-rater reliability as it involves comparing each rater’s assessments to an external referent rather than a group norm. This external referent is often referred to as a "Gold Standard" for assessments (Holt et al., 1997) or "Ground Truth".

Baker and Dismukes (2002) emphasizes that without consensus among pilot instructors on observed crew behaviors and their corresponding performance levels, accurate and reliable evaluations of aircrew cannot be achieved. The inconsistency among instructors leads to performance ratings that reflect the evaluators’ perspectives rather than the actual crew performance. This lack of agreement among instructors compromises the assessment of training outcomes and crew proficiency, potentially hindering the identification of performance issues critical to flight safety.
Chapter 3

Dissertation Aims

3.1 Statement of the Problem

Mixed EBT recurrent programs have been implemented and run successfully by many airlines for nearly a decade. EASA has recently allowed the introduction of baseline EBT programs to replace recurrent training and checking. EASA efforts to expand EBT to operator conversion course and initial type rating training course have stalled.

EBT is becoming increasingly popular in aviation. This popularity evokes the state of evidence-based medicine in mid-1990s. Loosely adapting words of Fowler (1997) first published in 1995 it could be said "6 years ago ‘evidence based training’ produced blank looks ... now ... it’s everywhere." Similar initiatives are seen in education; in 1999, Davies (1999) warned "Evidence-based education is not a panacea, but is a set of principles and practices for enhancing educational policy and practice."

Present type rating training courses are based on Operational Suitability Data (OSD) provided by aircraft manufacturers and regulatory requirements (most notably Appendix 9 to Part-FCL (EU, 2011)). Strictly following definition set by Davies (1999), present type rating training courses are indeed "evidence-based". Yet these courses do not meet definition of EBT\(^1\) established by EASA.

It can be argued that the evidence currently used in course design is obsolete or insufficient. It has been shown that evidence needs to be found and properly evaluated before it can be implemented (Rychetník et al., 2002; Švaříček, 2013).

---

\(^1\)(EU, 2012): evidence-based training (EBT) means assessment and training based on operational data that is characterised by developing and assessing the overall capability of a pilot across a range of competencies (competency framework) rather than by measuring the performance in individual events or manoeuvres;
Sufficient amount of data has been collected and interpreted by IATA to support implementation of a baseline EBT program (IATA, 2014). It has not been proven that the same data can be utilized for the implementation of EBT in type rating training.

Objectives of a type rating training course are not clearly defined in regulations. It is generally accepted that upon completion of a course a pilot shall demonstrate the level of theoretical knowledge and the skill required for the safe operation of the applicable aircraft type. Airlines, however, have higher expectations for their pilots; generally in line with ICAO aims of EBT program.\(^2\)

The implementation of EBT in a type rating training course needs to be carefully evaluated. A pilot still needs to learn how to fly a specific type and operate its systems; in order to do so, it may be needed to "measure pilot’s performance in individual events or manoeuvres." A type rating training course is short in duration and provides little room to train new competencies (especially those unknown from previous training and not specific to aircraft type). These facts preclude application of the very same EBT definition which EASA has chosen for recurrent programs, unless major changes are made in course length or initial pilot training. Neither is currently planned. Thus, the EBT approach requires some adjustments to preserve current training goals and bring improvements sought by airlines.

\(^2\)(ICAO, 2013): The aim of [an EBT] programme is to develop and evaluate the identified competencies required to operate safely, effectively and efficiently in a commercial air transport environment whilst addressing the most relevant threats according to evidence collected in accidents, incidents, flight operations and training.
3.2 Research Objectives

The following tasks and hypotheses are to be covered by this work.

Research Task RT1 *Evaluate existing grading methods.*

**Hypothesis H1** A competency-based grading system can be used to effectively track trainees progress during a type rating training course.

Research Task RT2 *Determine whether traditional and competency-based assessments identify the same areas where trainees over- or under-perform in type rating training courses.*

**Hypothesis H2_0** There is no significant difference in identified areas based on the type of grading system (traditional vs. competency-based).

**Hypothesis H2_1** There is significant difference in identified areas based on the type of grading system (traditional vs. competency-based).

Research Task RT3 *Examine suitability of EBT implementation into a stand alone\(^3\) type rating training course and into a tailored\(^4\) type rating training course.*

**Hypothesis H3_0** There is no significant difference in type training data and OCC training data

**Hypothesis H3_1** There is a significant differences in type training data and OCC training data.

Research Task RT4 *Develop methodology for evaluation of instructors concordance.*

---

\(^3\)Organized by an ATO without a link to any operator.

\(^4\)Organized or subcontracted by an AOC.
3.3 Purpose Statement

Although much effort has been put into research on the use of EBT in recurrent training, there are gaps in the research regarding studies that help to determine how evidence-based principles can be applied to type ratings.

The primary objective of this thesis is to support the expansion of EBT into type rating training courses.

Pilot training market is becoming increasingly competitive (Valenta, 2018). Following trends is a selling point - courses labeled "competency-based" are sometimes offered by training organizations without thorough understanding of competency-based training. Although some harm may be done by an improperly implemented competency-based instrument rating course, it should not create an unsafe situation as the course is governed by a conventional skill test. EBT is a popular term and will provide a marketing advantage, however, inappropriate EBT use may have negative implications on safety. EBT application is a recurring process - not just one-off adjustment to syllabi and training records. It needs to be determined for which kind of organization EBT provides both operational and safety benefit. Only those organizations shall be encouraged to implement EBT.

3.4 Implications

A thorough analysis of trainees’ performance throughout the course will be performed, providing a basis for the initial adjustments to the course syllabi. Outcome may identify specific tasks or maneuvers that could potentially be eliminated from the syllabi and others that may require increased focus.

A grading system forms an integral part of any training program. A grading system that offers accurate insights into the effectiveness of training will be scrutinized. Grades can be statistically processed only if instructor objectivity and concordance are assured. EASA dictates requirements for concordance but does not offer any tools or in-depth guidance. This work will propose a detailed scientific approach.
Chapter 4

Methodology

4.1 Data Sources and Analysis: Training Records

4.1.1 Requested Data

A major issue for industry-wide research is access to training records. The industry lacks a common scheme to process and share training data. Privacy laws severely restrict any research efforts. Tracking pilots progress across several institutions from the beginning of initial training to a final line check at the end of an operator conversion course is impossible.

An aviation authority from an EASA member state, two airlines holding EASA Air Operator Certificates (AOC) - one operating third-generation jets and the other operating fourth-generation jets—and three training organizations (offering an instrument rating course both in integrated and modular formats, a third-generation jet type rating course, and a fourth-generation jet type rating course, respectively) have been requested to submit anonymized records of pilot training and examination. As all training records are anonymized it is impossible to determine the background of trainees. However, trainees’ experience can be estimated on the basis of the type of training course or phase of training. Table 4.1 shows estimations used in this study. It is possible that an experienced general aviation flight instructor (more than 500 hours total time) enrolls in an initial type rating course with more than 10 simulator sessions – such career path is rare in Europe and is disregarded by this study.
Chapter 4

<table>
<thead>
<tr>
<th>Type of training course or phase</th>
<th>Experience (total hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instrument Rating</td>
<td>70 - 250</td>
</tr>
<tr>
<td>Instrument time in ATP integrated course</td>
<td>70 - 220</td>
</tr>
<tr>
<td>Type Rating (10 or more sessions)</td>
<td>200 - 500</td>
</tr>
<tr>
<td>Type Rating (less than 10 sessions)</td>
<td>More than 500</td>
</tr>
<tr>
<td>Operator Conversion Course (LIFUS 40 sectors or more)</td>
<td>220 - 700</td>
</tr>
<tr>
<td>Operator Conversion Course (LIFUS less than 40 sectors)</td>
<td>More than 500</td>
</tr>
</tbody>
</table>

*Table 4.1: Estimated experience based on type of training*

This study is focused on the type rating training course for the issue of the first type rating for a transport category jet used in commercial air transport. Training preceding and following this course was considered for analysis. Pilot experience at the beginning of this course is between 200 and 500 hours.

Preceding training is either an ATP integrated training course or modular course up to CPL/ME/IR. Sample training records provided by a training organization approved for an ATP integrated training course were screened. Initially, Phase 4 (instrument flight instruction) was expected to provide relevant information for this study as competencies build during this phase are the mostly used in future career. It was discovered that grading system used in the sample data was highly subjective and grades did not correspond to plain text remarks. Such data were not suitable for statistical processing and were disqualified from this study.

The Multi-crew cooperation training course (MCC) introduces additional complexity as it can be administered in four different versions: it can be combined with initial type rating training, incorporated as part of an ATP integrated course, offered as a stand-alone training course, or conducted under the Enhanced MCC training to airline pilot standards course (APS MCC) framework.

Data from MCC combined with initial type rating training course were considered in this study. Data from MCC included in ATP integrated course were inconsistent and unusable. No data from any stand-alone training course or APS MCC were available. At the time of data gathering no training organization offered an APS MCC; although this course is more expensive, it has potential to significantly improve pilots’ skills and should be included in future studies.

Data from both ATOs providing type rating training courses were used extensively.

32
The type rating course is followed by an Operator Conversion Course (OCC). OCC is formed by theoretical training on company procedures and CRM, flight training (e.g. emergency procedure in a simulator), Line flying under supervision (LIFUS) and a final line check. Both were asked to provide training records. One airline provided insufficient amount of records \((n = 3)\). The other airline provided more records, however, LIFUS records contained only plain text remarks and were not suitable for statistical processing. Total of 31 line check reports were included in this study. This is an insufficient amount of data to make conclusive findings, however, it can serve as a proof of concept whether line check results can or cannot be utilized as evidence for type rating training.

4.1.2 Data Types

Theoretical Knowledge Instruction

Neither training organizations nor airlines are required to track trainees’ progress during theoretical knowledge instruction. The only regulatory requirement is to log training attendance. Some training organizations utilize progress checks. Such checks are usually in a form of quiz questions and are very similar to the theoretical knowledge examination done by aviation authorities. Due to this characteristic, they provide very little information on trainees’ progress.

Carefully assembled tests and oral exams feature much greater potential. Such examination is usually utilized by universities. However, there is no standardized way of reporting exam results.

The industry would benefit from implementing a system for monitoring trainees’ progress in gaining theoretical knowledge. Such initiative is, however, beyond the scope of this study.

Theoretical Knowledge Examination Reports

Very little to no information may be derived from theoretical knowledge examination certificates; as these only indicate a percentage of correctly answered quiz questions which are randomly selected from a database for each subject. Detailed figures of frequently failed subjects and areas would provide some insight into knowledge deficit.

An aviation authority was requested to provide data from ATPL examination; however, this request was denied as the authority only tracks total number of exams and counts for fails or passes.
Chapter 4

No airline- or training organization-supplied data have been considered for further analysis as these examinations are often followed by a debrief with an instructor and must be corrected to 100%.

**Flight Training Records**

Flight training records contain description of trainee's performance during each training event. The original purpose is to serve as a medium to share information between instructors. These records are often very detailed. An analysis of flight training records is the primary focus of this study. The form of records varies across institutions. Several grading methods may be incorporated even within one course. Traditionally, records feature grading of maneuvers or tasks. Assessments of soft skills is often included.

**Flight Test Records**

Training courses are usually concluded by a flight test (skill test, proficiency check or line check).

Grading of skill test and proficiency check is no longer recommended by EASA. A simple Pass or Fail result is shown on the test form with a remarks section. Failed skill test and proficiency check require a ‘failure of test report’. An aviation authority declined to provide de-identified reports.

Line check records follow the format of flight training records. All line check records from the airline operating 3rd generation jet aircraft were analyzed.

### 4.1.3 Processed Data

The integral part of this study is an analysis of type rating training records. This study uses four distinct data sets from two EASA ATOs and one EASA AOC. This data set presents a unique opportunity for research into the efficacy of pilot training programs, offering insights into the development of competencies critical for the safe and efficient operation. Through statistical analysis, it is possible to identify trends, patterns, and predictors of success within the training program, contributing valuable knowledge to the field of aviation training and education. More than 38,000 grades were processed in the MATLAB environment.

**ATO A Type Rating Training Course**

ATO A offers type rating training for a fourth-generation jet, involving a comprehensive syllabus (fig. 4.1). This syllabus comprises classroom instruction, e-learning
modules, and three sessions guided by an instructor on an APT. These components are not assessed and were not included in this study’s evaluation. Scores from the theoretical knowledge examination were not provided in the records received. This study utilized the grades from the Graded Procedures Training on an APT, encompassing nine sessions, and Practical Training on an FFS, including eight sessions and additional sessions for LVO and UPRT as required. Base training or Zero Flight Time Training (ZFTT) grading was not available. Each training session was designed with specific scenarios, simulating both normal and non-normal conditions. The grading data, therefore, reflects a wide array of performance metrics, from basic operational tasks to complex emergency response strategies.

CBTA was utilized and nine competencies were graded during each session, a norm grade was established for each competency in every session. A five-point numerical scale was used.

The study dataset includes 42 trainees and 19 instructors, with the training designed to occur in sessions where trainees are paired up, yet the specific pairing of the trainees is not known or documented within the dataset. Each session is conducted with one instructor. The instructors are trained as standardized by the ATO to ensure a consistent level of instruction across all sessions. No further information on instructors was available. No specific details on the trainees, such as their background, skill level, or any other identifying information were available; however, experience level is estimated between 200 and 500 hours, suggesting this training course was their first multi pilot aircraft.

1Airbus Procedure Trainer (APT), in this context, refers to a Flight Simulation Training Device (FSTD) certified as FTD 1 with touch screen panels.
ATO B Type Rating Training Course

ATO B is approved for a type rating training course for a third-generation jet aircraft. The course syllabus is very similar to ATO A type rating training course. A slight difference can be observed with administration of procedure training: there are only two guided and not graded procedures training sessions and a minimum one self-study procedure training session\(^2\) on cockpit mock up. The practical training comprises 10 sessions on the FTD 2 for trainees without previous experience and 4 sessions for those with prior experience credit. All trainees attend five FFS sessions, which include one LVO session. Base training or ZFTT records were available.

ATO B provided anonymized records from 87 courses spanning the period from 2017 to 2020. Fourteen records were discarded, mostly due to missing or illegible pages. In 2019, there was a switch from traditional grading to Competency-Based Training and Assessment (CBTA) grading, with no adjustments made to the syllabus. This allows for comparison of the two grading systems.

No information about the age, gender, or experience of the trainees was provided. In the traditional grading, there were 52 trainees, all estimated to be first officers with no previous experience (total time 200 - 500 hours). In the CBTA grading, there were 21 trainees, of which 5 were experienced with more than 500 hours.

Training was delivered by a group of 34 instructors. No information about their age, gender, or experience of the trainees was provided. All instructors have been standardized by the ATO according to current EASA regulations. Instructors were trained on CBTA prior to the rollout of the new grading method, however some instructors may have had previous CBTA experience from the MPL training.

Traditional grading utilized a five-point numerical scale with associated labels (5: Above Standard, 4: Standard, 3: Below Standard, 2: Poor, 1: Failed - require repeat) and there was no further guidance on specific use of grades. Individual scenario elements were graded during each session.

There was no change to any scenario during the transition the CBTA grading, however, the training records have been substantially revised. The CBTA grading utilized a five-point numerical scale with descriptors(5: Master, 4: Consolidate, 3: Practice, 2: Apply and 1: Describe), elaborate requirements on the use of grades based on application of competency, task completions and instructor intervention was provided in the training manual. Each competency was graded during individual scenario elements in each session.

\(^2\)Trainees have unlimited access to the cockpit mock up and can utilize the device anytime throughout the course.
AOC B Operator Conversion Course

AOC B, an EASA AOC holder for commercial air transport, primarily operates third-generation jet aircraft. Newly hired pilots are required to complete an Operator Conversion Course (OCC), which includes classroom training, simulator sessions, LIFUS, and a final line check. However, no data from the simulator training were accessible for analysis. LIFUS data, consisting solely of plain text remarks, were not examined. Only reports from failed final line checks were provided, excluding data from successful ones. The analysis of these failure reports aimed to pinpoint the root cause of each failure, categorizing them by either task or competency. AOC B grades various 'skills' during line checks and uses 4-point categorical scale (Excellent, Average, Below Average, Failed and Not Observed). Text remarks were systematically analyzed to categorize them into groups that align with those used in the type rating course. Comparison of outcomes from ATO B type rating courses and AOC B line checks was performed.

4.2 Grading System

4.2.1 Free-Text Remarks

Free-text remarks can provide more elaborate information on the trainee’s progress, however, a large scale analysis is particularly challenging. This approach requires analyst to be thoroughly familiar with pilot training process. Some remarks such as “Improve Situational Awareness” do not present grounds for further analysis. By contrast, other remarks such as “Review holding pattern entry procedures” or “Review rudder trim technique” instantly point to problematic areas. Deficiencies in previous training can be identified by looking for patterns of similar remarks.

Free-text comments were not analyzed due to the inconsistent quality of text recognition, which was affected by the quality of scans and the diversity of handwriting styles. Additionally, extracting words from images within sample parts did not yield any results.

Grading consistency was verified by comparing remarks with grades within one training event and similar remarks from different instructors to assigned grades. For example if an instructor grades a landing as a 5 and notes that "trainee made a long landing" grading is inconsistent and cannot be processed further. Free-text comments were analyzed and compared with grades to verify the consistency of the rating in 25 randomly selected records.
4.2.2 Analyzing Grading

ATO A Type Rating Grading Data

This data set provides a good opportunity to observe how trainees develop their competencies during the course. The basic assumption is that the level of competency increases with each training session. In this study, the linear model of competency acquisition was assumed. Kuhfeld and Soland (2021) warn that the assumption of linear growth is frequently incorrect and several other learning curves have been described (Anzanello & Fogliatto, 2011). The S-curve or Plateau models are suggested by instructor anecdotal evidence; however, these are difficult to capture on a five-point grading scale, thus the linear growth was chosen. The grade value should increase in every session, this is a fixed effect of sessions. Sessions may vary in difficulty of the scenario creating a random effect, additional random effects are individual variability of trainees, instructor variability and possibly competency. Linear mixed-effects model (Bates, 2005) with following formula was used:

\[
value \sim 1 + session + (1|trainee) + (1|trainer) + (1|session) + (1|competency)
\]

It can be argued that competencies should be assessed individually, since the trainee can excel in one and fall short in another, therefore, an alternative formula was also considered: \(value \sim 1 + session + (1|trainee) + (1|trainer) + (1|session)\).

ATO B Type Rating Grading Data

Throughout the course, several tasks and scenario elements were repeated. To facilitate statistical analysis, similar tasks were categorized into groups, as depicted in table A.1. It is important to note that not all tasks were graded during each occurrence and there was a slight change in the tasks graded with the transition to CBTA. Each graded occurrence and common groups are illustrated in fig. 4.2.

To allow comparison of traditional and CBTA grades, a mathematical conversion of grades was carried out. The converted grade for the traditional method was calculated as \(T_{\text{converted}} = T - \text{standard}\). Through a meticulous analysis of the training syllabi, leveraging expert knowledge and comparison with the IATA (2013) Competency Map, the necessary competencies to successfully complete a given task were identified. Consequently, the converted grade for CBTA was calculated as \(C_{\text{converted}} = \frac{1}{9} \sum_{k=1}^{9} w_k(C - \text{norm})\), employing weights for respective tasks as outlined in table 4.2.

Variability of grades of these task groups throughout the course was observed with the aim to identify areas where trainees under- and over- perform. Weighted arithmetic means were determined for each session.
The dynamics of grade improvements during the training course should be explored, with a specific focus on whether these improvements were evident within two distinct grade types. The research questions were twofold: Is there evidence of grade improvement across repeating scenario elements for each grade type? And do grade types exhibit different trends in grades across all sessions, including differences in specific measurements in each session? The use of the linear mixed-effect model for traditional grading is not suitable. To better address these questions, a repeated measures analysis of variance (rANOVA) was conducted, with measurements serving as the within-subject factor (corresponding to the distinct grading occasions during each session) and grade type as the between-subject factor. This leads to the formulation of additional hypotheses: $H_{2a_0}$: there is no significant difference in grade values in each session and $H_{2b_0}$: there is no significant interaction between grade values and grade types.

To further elaborate on individual groups, another hypothesis was formed: $H_{2c_0}$ Group 7 grades from all sessions in CB and T are samples from continuous distributions with equal medians and tested using the Wilcoxon rank sum test.
Table 4.2: Scenario Elements Groups and Competency Weights

<table>
<thead>
<tr>
<th>Group-ID</th>
<th>Group-Name</th>
<th>APK</th>
<th>COM</th>
<th>FPA</th>
<th>FPM</th>
<th>KNO</th>
<th>LTW</th>
<th>PSD</th>
<th>SAW</th>
<th>WLM</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Normal Procedures</td>
<td>0.222</td>
<td>0.167</td>
<td>0.111</td>
<td>0.111</td>
<td>0.222</td>
<td>0.056</td>
<td>0.028</td>
<td>0.028</td>
<td>0.056</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Supplementary Procedures (Memory)</td>
<td>0.211</td>
<td>0.158</td>
<td>0.105</td>
<td>0.105</td>
<td>0.211</td>
<td>0.053</td>
<td>0.053</td>
<td>0.053</td>
<td>0.053</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Supplementary Procedures (Read and do)</td>
<td>0.267</td>
<td>0.200</td>
<td>0.067</td>
<td>0.067</td>
<td>0.133</td>
<td>0.067</td>
<td>0.067</td>
<td>0.067</td>
<td>0.067</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Manual Landing</td>
<td>0.100</td>
<td>0.050</td>
<td>0.000</td>
<td>0.700</td>
<td>0.100</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.050</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Non-Normals - minor</td>
<td>0.200</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
<td>0.150</td>
<td>0.075</td>
<td>0.075</td>
<td>0.100</td>
<td>0.100</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>Non-Normals - complex</td>
<td>0.190</td>
<td>0.095</td>
<td>0.095</td>
<td>0.095</td>
<td>0.143</td>
<td>0.095</td>
<td>0.095</td>
<td>0.095</td>
<td>0.095</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>Non-Normals - one engine inop</td>
<td>0.195</td>
<td>0.098</td>
<td>0.146</td>
<td>0.195</td>
<td>0.098</td>
<td>0.073</td>
<td>0.073</td>
<td>0.074</td>
<td>0.049</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>Manual Flight Maneuvers</td>
<td>0.121</td>
<td>0.030</td>
<td>0.190</td>
<td>0.095</td>
<td>0.190</td>
<td>0.048</td>
<td>0.048</td>
<td>0.048</td>
<td>0.048</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>Automatic Flight Maneuvers</td>
<td>0.057</td>
<td>0.043</td>
<td>0.571</td>
<td>0.000</td>
<td>0.214</td>
<td>0.029</td>
<td>0.029</td>
<td>0.029</td>
<td>0.029</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>LVO</td>
<td>0.190</td>
<td>0.143</td>
<td>0.190</td>
<td>0.095</td>
<td>0.190</td>
<td>0.048</td>
<td>0.048</td>
<td>0.048</td>
<td>0.048</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>Recall Maneuvers</td>
<td>0.100</td>
<td>0.050</td>
<td>0.100</td>
<td>0.000</td>
<td>0.100</td>
<td>0.150</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
</tr>
<tr>
<td>100</td>
<td>CRM (LTM, PSD, WLM)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.333</td>
<td>0.333</td>
<td>0.000</td>
<td>0.333</td>
<td>1</td>
</tr>
<tr>
<td>101</td>
<td>Leadership and Teamwork</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1</td>
</tr>
<tr>
<td>102</td>
<td>Problem Solving and Decision Making</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1</td>
</tr>
<tr>
<td>103</td>
<td>Workload Management</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>1</td>
</tr>
<tr>
<td>104</td>
<td>SITAW</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1</td>
</tr>
<tr>
<td>105</td>
<td>Communication</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>Ignore - group</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
</tbody>
</table>

4.3 Instructor Grading Reliability

The type rating grading data are particularly specific. Trainees typically train in pairs and participate in all sessions. Up to nine different competencies are evaluated in each session, and no session is observed by more than one instructor. Although Type Rating Instructors are required to pass an Assessment of Competence every six years, there is no mandate to document discrepancies in grades assigned.

The dataset comprised five main elements: value, trainer, trainee, session, and competency. Additionally, for ATO B the groupID was considered. Here, value represents the difference between the grade awarded by the trainer to a trainee and the norm for CBTA, or the standard grade in a traditional grading system.

It was impossible to directly apply any of the common inter-rater reliably or agreement measures to this dataset. However, several assumptions about the data were made: the trainees have similar background and comparable level of prior experience, the learning curve and performance may vary, thus the assigned grades will be below, on or above the target set by the syllabus, but the training objective (becoming type-rated after a skill test) was achieved by all trainees. Overall, the distribution of the grades should be relatively uniform.

Given the dataset’s unique structure, conventional inter-rater reliability or agreement measures could not be directly applied. However, several assumptions were made: trainees possess similar backgrounds and levels of prior experience, learning curves and performance vary so may fall below, meet, or exceed the syllabus targets, all trainees achieved the training goal of becoming type-rated after a skill
test. So the distribution of grades is expected to be relatively uniform.

To assist in achievement of research task RT4 the following hypothesis was formulated H4a: *Each instructor assigns grades that come from the same distribution* and was tested using the Kruskal-Wallis test. Instructors whose grades diverged significantly from the distribution underwent further analysis concerning the number of sessions they conducted and the progress of the trainees they trained.
Chapter 5

Results

5.1 Trainee Progress Analysis

ATO A Type Rating Grading Data

The results from the linear mixed-effects model fit by maximum likelihood estimation with formula $value \sim 1 + session + (1|trainee) + (1|trainer) + (1|session) + (1|competency)$ were $AIC = 6493.6$, $BIC = 6540.5$, $LogLikelihood = -3239.8$ and $Deviance = 6479.6$. The fixed effect coefficients with 95% confidence intervals were intercept: $Estimate = 2.1761$, $SE = 0.19368$, $p < 0.00001$ and session: $Estimate = 0.098256$, $SE = 0.016752$, $p < 0.00001$, indicating statistically significant initial grade of 2.2 with 0.1 increment per session. The random effects were trainees estimate of 0.1272 with 95% CI from 0.10077 to 0.16056, trainer estimate of 0.17485, with 95% CI from 0.11852 to 0.25795, session estimate of 0.3671, with 95% CI from 0.26162 to 0.5151, competency estimate of 0.14207, with a 95% CI ranging from 0.087336 to 0.2311 and residual error estimate of 0.40362.

ATO B Type Rating Grading Data

The results from the linear mixed-effects model for ATO B CBTA indicated a less than optimal fit: $(AIC = 16641, BIC = 16693, LogLikelihood = -8313.3, Deviance = 16627)$. Fixed effects: Intercept: $Estimate = 3.2373$, $SE = 0.077562$, $tStat = 41.738$, $DF = 13237$, $p = 0$ and session $Estimate = 0.00046257$, $SE = 0.0030647$, $tStat = 0.15093$, $DF = 13237$, $p = 0.88003$. Random effects: trainees 0.24998, CI from 0.18391 to 0.3398, trainers 0.20879, CI from 0.15154 to 0.28765, session 0.046566, CI from 0.03076 to 0.070492, competencies 0.036579, CI from 0.021355 to 0.062658 and error 0.44917, CI from 0.44378 to 0.45463.
Figures for Both ATOs

Mean grade values for all trainees, across each session and for each competency, are depicted in fig. 5.1 for ATO A, and in fig. 5.2 for ATO B. Additionally, fig. 5.3 shows mean grade values across each session for selected group IDs and fig. 5.4 presents the mean grade values for traditional grades at ATO B.

Figure 5.1: Mean Grade by Session and Competency (ATO A)

Figure 5.2: Mean Grade by Session and Competency (ATO B - CBTA)
5.2 Scenario Elements Evaluation

To allow for initial comparison of CBTA and traditional grades, failure ratios mean combined grades (mean converted grade per session) across CBTA (C) and traditional (T) sessions are shown in fig. 5.5. The error bars represent the standard deviation of the mean combined grades for each session, plotted on the primary vertical axis. The secondary vertical axis illustrates the fail ratio, presented as orange bars for each session.
Failed sessions require additional training. Each additional training is a standalone session with a separate training record. Figure 5.6 shows tasks or scenario elements that were trained during additional sessions.

Session 9 is identified as having the highest failure rate across all sessions. Figure 5.9 illustrates the distribution of all raw traditional grades during these sessions, distinguishing between those sessions that failed (requiring a repeat, depicted in blue) and those that succeeded (depicted in orange). Additionally, all
assigned CBTA grades, along with their corresponding competencies, are presented in fig. 5.8. The distributions of all CBTA grades and specifically FPM CBTA grades are detailed in fig. 5.9 and fig. 5.10, respectively.

**Figure 5.7:** Relative Probability of Assigned Traditional Grades in Session 9 for Failed and Passed Sessions

**Figure 5.8:** All Assigned CBTA Grades in Session 9

![Relative Probability of Assigned Traditional Grades](image1)

![All Assigned CBTA Grades](image2)
In order to reveal problematic groups of tasks, the converted grades were compared between the two grading systems in fig. 5.11.
Chapter 5

Figure 5.11: Comparison of Lesson Mean Converted Grades across Groups

With focus on group 7 (non-normals - one engine inop) in sessions 9, 13 and 15, the repeated measures ANOVA revealed no significant main effect of the grading in different sessions ($F(2, 136) = 0.29559$, $p = 0.74457$) and failed to reject H2a0, suggesting that, on average, grade values did not significantly improve or decline across the three measurements. Furthermore, there was no significant interaction effect between grade type and measurements ($F(2, 136) = 0.32698$, $p = 0.72166$) and failed to reject H2b0, indicating that the trend of grade changes over the measurements did not differ significantly between the grade types. Subsequent multiple comparison tests, conducted to examine pairwise differences between measurements within each grade type, reinforced the rANOVA findings. For CBTA, which was expected to show improvement, all pairwise comparisons yielded p-values significantly above the conventional alpha level of 0.05, indicating no statistically significant improvements or declines between any two measurements (all $p > 0.995$). Similarly, for traditional grades, which was expected to possibly not show any trend, the comparisons did not reveal statistically significant differences among the measurements (all $p > 0.374$). Figure 5.12 shows how converted grades vary across each session and between CBTA and traditional grades.
Chapter 5

Figure 5.12: Group 7 Grade Distribution for CBTA and Traditional Grades

Wilcoxon rank sum test rejects H2c0 hypothesis with \( p = 1.8514 \times 10^{-6} \). Figure 5.13 shows Group 7 grade distribution in all sessions. Additionally, Wilcoxon rank sum test for left-tailed hypothesis Group 7 grades from session 9 median of CB is equal or more than the median of T was used and failed to reject with \( p = 0.9936 \).

Figure 5.13: Group 7 Grade Distribution in All Sessions for Converted CBTA and Combined Traditional Grades
5.3 Comparison of Outcomes from Type Ratings and Line Checks

Originally, groups from table 4.2 were intended to identify the root causes of line check failures. However, during the analysis, several elements were found that did not correspond to these predefined groups. Consequently, three additional groups were established to accommodate these outliers: Group -1 for Operations Manual knowledge, Group -2 for Energy/Descent Path Management, and Group -3 for ATC communication. Figure 5.15 and fig. 5.14 illustrate the relative likelihood of each group contributing to a line check or session failure, respectively.

![Figure 5.14: Distribution of Groups Causing a Session Failure during Type Rating Course](image-url)
5.4 Instructor Grading Reliability

ATO A Type Rating Grading Data

Only 0.08% of sessions required a repeat (were graded fail); fig. 5.16 shows the proportion of sessions marked as failed, as graded by various instructors.

Hypothesis H4a0 was rejected by Kruskal-Wallis test with \( p = 1.1384 \times 10^{-102} \). Figure 5.17 can be used to identify instructors who grade significantly different than other instructors.

Further analysis of instructor 14 whose mean ranks significantly differ from other instructors reveals that this instructor graded only three trainees as shown in fig. 5.19. Black lines indicate mean grades with standard deviation from other instructors, colored lines are individual trainees.

Black crosses in fig. 5.18 indicate mean difference from a norm grade assigned to each competency by all instructors and colored lines indicate individual instructors.
Chapter 5

Figure 5.16: Ratio of Failed Sessions as Graded by Instructors (ATO A)

Figure 5.17: Instructor Ranks (ATO A)
ATO B Type Rating CBTA Grading Data

3.8% of sessions resulted in failure. fig. 5.21 illustrates the percentage of failed sessions as graded by each instructor. Trainer 7 graded 9 sessions, Trainer 17 graded
8 sessions, and Trainer 15 graded 38 sessions. In particular, Trainer 15 failed Trainees 4 and 5 twice, and in three out of the four failures, did not identify any competencies below the norm. Additionally, Trainees 20 and 21 experienced failure twice, each time by Trainers 7 and 17.

Hypothesis H4a for ATO B CBTA grades was rejected with $p$ close to 0. Ranks are shown on fig. 5.22.

Figure 5.23 indicates the mean difference from a norm grade assigned to each competency by all instructors and colored lines indicate individual instructors. Figure 5.24 illustrates the mean difference assigned by selected instructors to each group of scenario elements.

Figure 5.21: Ratio of Failed Sessions as Graded by Instructors (ATO B - CBTA)
Figure 5.22: Instructor Ranks (ATO B - CBTA)

Figure 5.23: Mean Grades Assigned by Instructors to Competencies (ATO B - CBTA)
Figure 5.24: Mean Grades Assigned by Selected Instructors to Groups (ATO B - CBTA)

Figure 5.25: Mean Grades Assigned by Selected Instructors to Groups (ATO B - Traditional)

ATO B Type Rating Traditional Grading Data

Figure 5.26 displays the percentage of sessions that failed, as graded by each instructor, with an overall total of 1.5% failed sessions. Instructor 6 conducted 31 sessions, Instructor 9 conducted 24, Instructor 11 conducted 20, Instructor 12 conducted 14, Instructor 15 conducted 58, and Instructor 21 conducted 18 sessions.

The hypothesis H4a0 for ATO B traditional grades was rejected, with a p-value close to 0. Rankings are depicted in fig. 5.22. Figure 5.25 shows the mean difference
in grades assigned by selected instructors to each group of scenario elements.

Figure 5.26: Ratio of Failed Sessions as Graded by Instructors (ATO B - Traditional)

Figure 5.27: Instructor Ranks (ATO B - Traditional)
Chapter 6

Discussion

6.1 Grading Methods

In case of ATO A the model reveals a statistically significant, positive progression in grades over sessions, suggesting an overall learning effect or improvement in performance inline with the CBTA philosophy. The random effects on grade variability are large. Trainee variability is an inherent part of the system, the absence of extreme values in this data set suggests no deeper research is required. However, should these values increase dramatically, it may indicate the necessity for implementing some form of course admission control. The instructor variability is examined later in this section. Significant session variability suggests that session-specific factors greatly affect grade variability and can be attributed to varying difficulty of sessions, but without a scenario element grading individual areas of focus cannot be found.

Grading competencies based on individual scenario elements can provide detailed insights into both trainee progress and scenario difficulty. This is why IATA (2013) ranked this method highly. ATO B has adopted this approach. There is no distinct trend in competency grades and data suggest that instructors may exhibit grading laziness due to the extensive number of grades (approximately 54) they must record during each session, in addition to providing free-text remarks for competencies below the norm. It appears that instructors often default to grading at the norm, and if performance falls below, they typically identify only one competency for detailed explanation, despite the likelihood of multiple competencies or scenario elements being below the standard. In contrast, ATO A follows the more common practice in EBT programs, grading competencies per session. This method is simpler for instructors but yields limited data for statistical analysis of scenario elements, making it challenging to identify specific problematic areas.

An interesting observation is that in ATO B, FPA and FPM mean grades
slightly decrease with the introduction of engine failures in session 9, however, this does not occur in ATO A, where engine failures are first introduced in session 5 on FTD using autoblight only and later in session 15 with manual flight. LOFT sessions at ATO B display a drop of mean grades, this does not occur at ATO A. This may indicate a better level of CBTA implementation at ATO A.

Neither ATO employs a level 2 metrics of grading Observable behavior, which could enhance competency tracking but still falls short in describing performance on individual scenario elements. A level 3 metrics approach seems essential for type rating training courses, offering valuable data on selected scenario elements without significantly complicating the grading process. This aligns with the methodology (Stainer, 2021) adopted for their EBT program. While free-text remarks may not be useful for statistical analysis, they remain crucial for communication - both between instructor and trainee and among instructors conducting different training sessions - and should therefore be preserved.

A competency-based grading system that uses the EASA Competency Framework can be used to effectively track trainees progress during the type rating training course and hypothesis H1 cannot be disproved. It is important to note, that a task or scenario element monitoring (such as level 3 metrics) must be employed and instructors must consistently apply the VENN grading model, grading not solely on outcomes but also monitor OBs.

### 6.2 Scenario Elements Evaluation

Zero signifies a standard grade in the traditional grading system and the norm-grade in CBTA. Across different sessions, there is no distinct trend observed. Despite significant overlap in the error bars, a subtle shift towards higher grades is noticeable with CBTA grading. This shift could be attributed to CBTA’s more accurate reflection of actual performance. However, the potential for grading leniency must also be considered. In the traditional system, instructors are required to provide comments on trainee performance only if a grade of 1 (fail) is assigned. Conversely, CBTA mandates justifications for grades both below and above the norm. Notably, the failure ratios remain quite low.

Session 9 exhibits the highest failure ratio under both grading systems. This session, which introduces engine failures, is conducted on a FTD Level 1. Engine failures in third-generation jets present significant challenges for pilots with low experience, demanding proficient rudder-stick coordination. The appropriateness of using an FTD for this particular task is debatable and might adversely affect the
training outcome.

Traditional grading incorporates the evaluation of soft skills throughout the session, adhering to recommended CRM grading guidelines and aligning closely with non-technical competencies assessment. Notably, no soft skill (group id >= 100) solely resulted in session failure, with only four instances receiving a grade of 1. In direct contrast, non-technical competencies accounted for six out of eleven additional sessions under CBTA grading. This discrepancy may significantly contribute to the higher fail ratio observed in CBTA grading for session 6. Despite consisting of relatively straightforward tasks, such as cold weather operations and unreliable airspeed, the combination of these tasks creates a complex scenario that excellently showcases competencies. A similar rationale applies to Group 5 (Non-normals - minor), which led to repetitions only in CBTA. While these tasks are simple individually, their aggregation effectively highlights trainees’ weak competencies.

Interestingly, there were no repetitions of Group 8 (Manual Flight Maneuvers) tasks under CBTA grading, despite the anticipated struggles of low-experience trainees with manual handling of third-generation jets — a challenge frequently observed in traditional grading through repeated raw data ILS approaches. This absence in CBTA grading suggests a potential shift in instructor focus towards other competencies. An analysis of Competency 4 (FPM) grades reveals 61 out of 1499 grades below the norm, predominantly (12) in manual landing, yet none resulted in session failure. Ten relate to engine failures. This outcome may be due to the variability of the trainee, necessitating long-term attention to discover more definitive trends.

Similarly to session means, all individual groups have slightly lower means in traditional grading system with significant overlap in standard deviation. However, Group 7 (non-normals - one engine inop) is showing results above 0 in CBTA and below 0. As discussed earlier, engine failures pose a significant challenge to trainees and account for a large number of repetitions, which is even higher in CBTA. Contradictory, this is not reflected in the absolute value of grades.

The findings suggest that there is no evidence of grade improvement during the course for either CBTA and Traditional grading. Moreover, there is no differential trend observed in grade changes between these grade types across the repetitions of scenario elements. The absence of significant differences, both within and between grade types, indicates a consistent pattern of grade values over time. This is somewhat contradicting the CBTA philosophy. The trainee should gain competencies and the associated grade should improve throughout the training course. With the relatively large error associated with the measurements ($MSE = 0.14323$), it is pru-
dent to consider other possible statistical influences: the variable level of difficulty of each session may be significant. However, for Group 7, this should not influence the outcome of the maneuver, although it may affect the subjective perception of the situation by the trainee.

Although hypothesis $H_2$ was statistically disproved, an overarching examination of the data reveals similar areas of underperformance among trainees. The results do not show the existence of a positive shift in the selected median grades from CBTA to traditional system. It was suspected that in Session 9, instructors might overvalue FPM and subsequently rate other competencies higher, even in cases where the session’s overall outcome was a failure; however, this hypothesis was not supported by the evidence. Additionally, no specific areas were identified where trainees consistently excel. In the context of traditional grading, only 8.3% of the evaluations were categorized as ’Above Standard.’ Similarly, within the CBTA system, the highest grade, 5, was awarded in merely 2.4% of cases.

6.3 Comparison of Outcomes from Type Ratings and Line Checks

While unsuccessful line checks make up less than 1% of all cases, an in-depth review of failed line check records has yielded interesting findings. A total of 31 failed initial line checks for first officers were evaluated. The absence of overlapping groups clearly disproves hypothesis $H_3$, because there is a significant difference between type rating training data and OCC training data. Nevertheless, this assertion comes with notable limitations, given that only failed line checks were considered. However, it certainly suggests that airlines can furnish ATOs with informative data.

Notably, while Group 4 (manual landing) did not directly lead to session failures, it consistently received lower grades during the type rating course, identifying it as a potential area of focus. Group -3 (ATC communication) suggests a failure in the training system as early as the initial stages, prior to type rating course. This area, however, offers opportunities for enhancement through simulator practice. Group -2 (energy/descent path management), a skill largely acquired through experience, could see improved trainee threat assessments with thorough briefings, although incorporating such training faces constraints due to the already busy syllabus; adding new elements would inevitably extend the duration of the type rating training. Lastly, Group -1 (Operating Manual knowledge) points to potential issues in OCC classroom training, which could be rectified only in type rating programs tailored specifically to the airline’s operations.
6.4 Instructor Grading Reliability

The hypothesis H4a was rejected for all three datasets. This rejection does not necessarily indicate a low level of grading reliability. However, instructor grading that significantly deviates from the overall distribution warrants further scrutiny.

For ATO A, Instructor 14’s grading significantly differs from that of all other instructors. A closer examination reveals that Instructor 14 graded only 3 sessions, likely comprising one pair in one session and another session where only one grading sheet was available in the dataset. Two of the three trainees often received above-average grades from other instructors, suggesting their performance was consistently strong. Therefore, it appears there is no discrepancy in Instructor 14’s grades.

Instructor 9 has the highest fail ratio but graded only 6 sessions, with one session graded as a fail. Despite this, his/her grade distribution aligns with that of other instructors.

Instructor 11 significantly differs from only 8 instructors and graded 101 sessions. A closer examination of his/her grading reveals that he/she mostly assigns above-norm grades for non-technical competencies (2 and 6 to 9), which may indicate a potential standardization problem.

Instructors 15 to 19 each graded only 2 sessions.

For ATO B CBTA grades, the linear mixed-effects model did not demonstrate a good fit; however, it revealed a random effect of instructor on grades of 0.21, compared to 0.17 at ATO A. This dataset is particularly unique as it includes four trainees (likely forming two pairs) who each failed two sessions. Although unusual, CBTA grading should still reflect their learning journey. A total of 296 sessions were graded by 21 instructors, many of whom graded only a few sessions, making assessments of reliability challenging. Further analysis will concentrate on selected instructors who either graded many fails or handled a significant number of sessions.

Instructor 7 and 17 show high fail ratio, but both instructed a pair of underperforming trainees and their grade distribution appears relatively good.

Instructor 15 frequently assigns relatively high grades yet demonstrates a very high fail ratio, presenting an apparent inconsistency. A detailed examination of their grading shows that, for three failed sessions, no competency was graded below the norm, which fails to justify the session failures and clearly indicates a problem with his/her grading style. Additionally, Instructor 15 occasionally assigns half grades (0.5), a practice not permitted in CBTA grading—two other instructors have also assigned 0.5 grades.

In general, instructors often grade around the norm, which may be attributed to the extensive number of grades they are required to record.
Interestingly, ATO B traditional grades appear to be much better aligned, despite the lack of a clear definition for these grades. This consistency could be due to the long-standing experience with this grading approach and a well-established company culture. Regarding Instructor 15, who assigns relatively high grades in traditional grading yet maintains a high failure ratio, this pattern confirms a standardization issue.
Chapter 7

Conclusions

The growth of the airline industry and the increasing demand for pilots underscore the need for enhancements in pilot training. Effective training is essential for ensuring safe and efficient operations. However, current training methods are often rigid and may not fully meet the needs of airlines (Leoff, 2018).

This work has conducted a comprehensive analysis of various aspects of type rating training, including training syllabus, grading systems and instructor concordance.

The comparative study of traditional and competency-based training and assessment systems highlights their common goal: to pinpoint areas of over- or under-performance among trainees. Competency-based grading, in particular, offers a more detailed and nuanced insight into a trainee’s abilities, suggesting that a move towards competency-based assessments could refine the accuracy of pilot training evaluations. This approach allows for more targeted and effective training, especially in highlighting non-technical skills that traditional grading systems often overlook. Nevertheless, it is important to note that while competency-based grading excels in evaluating non-technical skills, it may omit some aspects of technical skills assessment. As Gontar and Hoermann (2015) propose, pass-fail decisions should encompass both technical and non-technical skills, necessitating the inclusion of task or scenario element monitoring in type rating training records. Moreover, the importance of maintaining free-text remarks as a communication tool is evident.

The evidence suggests that an ATO alone cannot gather sufficient data to adjust the training syllabus effectively. A practical solution is the customization of type rating courses to meet specific airline requirements, a strategy already in use. However, this requires airlines to provide feedback to the ATOs. While standalone type ratings will continue to be in demand, especially among smaller operators with limited financial or coordination capacities, aircraft manufacturers and/or regula-
tory authorities could act as intermediaries in providing additional evidence from initiatives like Data4Safety.

Grading data from two ATOs are analyzed. ATO A demonstrates consistent grading and alignment with trainees’ learning curves. In contrast, ATO B CBTA grades show significant variability, likely due to a recent shift to CBTA from Traditional grading system, underscoring the relevance of this study in monitoring grading practices.

The application of the Kruskal-Wallis test has been notably effective in identifying instructors with inconsistent grading practices. However, this test alone is insufficient as the sole measure of instructor grading reliability. It is necessary to individually examine outliers. A common challenge across all three datasets is the presence of instructors who conducted a relatively low number of sessions, making statistical evaluation difficult, a similar issue identified by Robert W. Holt and Boehm-Davis, 2002. Another limitation is the difficulty in detecting instructors who grade consistently at the norm but occasionally assign random grades above or below the norm, a widespread issue in grading methodologies as discussed by Goldsmith and Johnson (2002).

Furthermore, it has been observed that type rating and operator conversion courses often cover competencies that should have been acquired during initial training, leading to increased costs for airlines and frustration among trainees and instructors. The adoption of EBT principles could optimize training resources. The mastery of CBTA is a crucial first step. While IATA has provided clear definitions of competencies, it is vital for instructors to fully comprehend these definitions to apply effective evaluation methodology. A well-implemented CBTA system is a valuable asset in pilot training, enabling instructors to objectively record trainee performance and deliver focused training. Such a system also serves as a comprehensive statistical tool.
Research Work

Publications

Effects of Airline Industry Growth on Pilot Training
Valenta, V.

Implementing Evidence-Based Practice in Initial Pilot Training
Valenta, V.

Effects of Napping on Pilot Performance: An Experimental Study
Hanáková, L., Valenta, V., Řezníček, A., Matyáš, R., Socha, V.

Design of Wearable Eye Tracker with Automatic Cockpit Areas of Interest Recognition
Socha, V., Vídeňský, J., Kušmírek, S., Hanáková, L., Valenta, V.

Development of Flight Simulation Device for Perception Assessment
Socha, V., Hanáková, L., Kušmírek, S., Malich, T., Gavura, T., Valenta, V., Kavka, M., Piorecká, V., Lališ, A.

Effect of Psychological Training on Pilot’s Performance
Pilots’ performance and workload assessment: Transition from analogue to glass-cockpit
Socha, V., Socha, L., Hanáková, L., Valenta, V., Kušmírek, S., Lališ, A.
Applied Sciences. 2020, 10(15) ISSN 2076-3417.

Workload assessment of air traffic controllers Socha, V., Hanáková, L., Valenta, V.,
Socha, L., Ábela, R., Kušmírek, S., Pilmannová, T., Tecl, J.

Impact of Regulatory Changes on Operating Procedures Design in Commercial Air
Transport Operations
Valenta, V., Hulínská, Š., Puškáš, T.
Journal Of Global Science. 2016, 1(1) ISSN 2453-756X.

List of projects
Development of methods for implementation of evidence based training into initial
professional pilot training programme
principal investigator: Valenta, V.
co-investigators: Absolon, S. - Hulínská, Š. - Němec, V.
2017 - 2018
Studentská grantová soutěž ČVUT - SGS17/155/OHK2/2T/16

Air Traffic Control and Radiotelephony Simulator for Remote Teaching
principal investigator: Valenta, V.
co-investigator: Malich, T.
2021 - 2021
EuroTeQ

Experimental verification of pilots’ resilience against vestibular illusions in terms of
training and practice
principal investigator: Kalivodová, M.
co-investigators: Hanáková, L. - Karapetjan, L. - Socha, V. - Valenta, V.
2021 - 2022
Studentská grantová soutěž ČVUT - SGS21/134/OHK2/2T/16
References


European Union Aviation Safety Agency. (2024). *Ebt manual (2.1).*


International Air Transport Association. (2024). Evidence-based training implementation guide (2nd ed.).


Leoff, T. (2018). Strategic option for pilot training innovation. Workshop on Area 100 Knowledge, Skills & Attitudes: theoretical knowledge courses for the ATPL, MPL and CPL.


Todd, M. A., & Thomas, J. W. M. (2013). Experience, competence, or syllabus? influences on flight hours at licensing of commercial pilots. *International Jour-


Wikander, R., & Dahlström, N. (2016a). The multi crew pilot licence-revolution, evolution or not even a solution?


Acronyms

ALPA  Airline Pilot Association
AMC   Acceptable means of compliance
AOC   Air Operator Certificate
APS MCC  Enhanced MCC training to airline pilot standards course
APT   Airbus Procedure Trainer
ATO   Approved Training Organization
ATP   Airline transport pilot
ATPL  Airline transport pilot license
CBT   Competency-Based Training
CBTA  Competency-Based Training and Assessment
CPL   Commercial pilot license
CRM   Crew Resource Management
EASA  European Union Aviation Safety Agency
EBT   Evidence-Based Training
ERAU  Embry-Riddle Aeronautical University
FAA   Federal Aviation Administration
FFS   Full Flight Simulator
FSTD  Flight Simulation Training Device
FTD   Flight Training Device
IATA  International Air Transport Association
ICAO  International Civil Aviation Organization
IMC  Instrument meteorological conditions
IR  Instrument Rating
ISD  Instructional Systems Design
LIFUS  Line flying under supervision
LVO  Low visibility operations
MCC  Multi-crew cooperation training course
ME  Multi-engine
MPL  Multi-crew pilot license
N.A.C.A  National Advisory Committee for Aeronautics
OB  Observable behavior
OCC  Operator Conversion Course
OM  Operations Manual
OSD  Operational Suitability Data
PPL  Private pilot license
SBG  Standards-Based Grading
SITAW  Situational Awareness
UPRT  Upset Prevention and Recovery Training
ZFTT  Zero Flight Time Training
List of Figures

2.1 Competency Gaps between Succeeding Training Courses .......................... 11

4.1 Type Rating Training Course Flow ...................................................... 35
4.2 Allocation of Graded Groups in Lessons Comparing CBTA and Traditional Grading, Highlighting Common Groups .......................... 39

5.1 Mean Grade by Session and Competency (ATO A) ................................. 43
5.2 Mean Grade by Session and Competency (ATO B - CBTA) ...................... 43
5.3 Mean Grade by Session and Group ID (ATO B - Traditional Grading System) ........................................................................ 44
5.4 Mean Grade by Session (ATO B - Traditional Grading System) ................ 44
5.5 Comparison of Mean Combined Grades and Failure Ratios across CBTA- and Traditional-graded Sessions .................................................. 45
5.6 Groups of Tasks Repeated during Additional Training Sessions ............ 45
5.7 Relative Probability of Assigned Traditional Grades in Session 9 for Failed and Passed Sessions ......................................................... 46
5.8 All Assigned CBTA Grades in Session 9 .................................................. 46
5.9 Relative Probability of All Assigned CBTA Grades in Session 9 for Failed and Passed Sessions ......................................................... 47
5.10 Relative Probability of FPM Assigned CBTA Grades in Session 9 for Failed and Passed Sessions ......................................................... 47
5.11 Comparison of Lesson Mean Converted Grades across Groups ................. 48
5.12 Group 7 Grade Distribution for CBTA and Traditional Grades ............... 49
5.13 Group 7 Grade Distribution in All Sessions for Converted CBTA and Combined Traditional Grades .................................................. 49
5.14 Distribution of Groups Causing a Session Failure during Type Rating Course ........................................................................ 50
5.15 Distribution of Groups Causing a Line Check Failure during Operator Conversion Course ........................................................................ 51
5.16 Ratio of Failed Sessions as Graded by Instructors (ATO A) ....................... 52
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.17 Instructor Ranks (ATO A)</td>
<td>52</td>
</tr>
<tr>
<td>5.18 Mean Grades Assigned by Instructors to Competencies (ATO A)</td>
<td>53</td>
</tr>
<tr>
<td>5.19 Trainees Trained by Instructor 14 (with Influence)</td>
<td>53</td>
</tr>
<tr>
<td>5.20 Trainees Trained by Instructor 14 (without Influence)</td>
<td>53</td>
</tr>
<tr>
<td>5.21 Ratio of Failed Sessions as Graded by Instructors (ATO B - CBTA)</td>
<td>54</td>
</tr>
<tr>
<td>5.22 Instructor Ranks (ATO B - CBTA)</td>
<td>55</td>
</tr>
<tr>
<td>5.23 Mean Grades Assigned by Instructors to Competencies (ATO B - CBTA)</td>
<td>55</td>
</tr>
<tr>
<td>5.24 Mean Grades Assigned by Selected Instructors to Groups (ATO B - CBTA)</td>
<td>56</td>
</tr>
<tr>
<td>5.25 Mean Grades Assigned by Selected Instructors to Groups (ATO B - Traditional)</td>
<td>56</td>
</tr>
<tr>
<td>5.26 Ratio of Failed Sessions as Graded by Instructors (ATO B - Traditional)</td>
<td>57</td>
</tr>
<tr>
<td>5.27 Instructor Ranks (ATO B - Traditional)</td>
<td>57</td>
</tr>
</tbody>
</table>
List of Tables

2.1 Competency-Based Performance Level Descriptors (Harms, 2015) .... 24
2.2 Grading Metrics (EASA, 2023b) ........................................... 25

4.1 Estimated experience based on type of training ...................... 32
4.2 Scenario Elements Groups and Competency Weights ............... 40

A.1 Overview of Scenario Elements and Tasks with Their Allocation to Groups .................................................. 83
A.2 Competency Descriptions (adapted from (EASA, 2023b)) ......... 84

C.1 Distribution of norm Grades throughout the Course ................. 89
C.2 Trainee 35: Percentage of Grades per Competency throughout the Course .................................................. 89
C.3 Session 10: norm Grades per Competency .............................. 90
C.4 Session 10: Percentage of Grades per Competency ................. 90
C.5 Session 10: Percentage of Grades per Competency by Instructor 1 (8 grading opportunities) ................................. 91
C.6 Session 10: Percentage of Grades per Competency by Instructor 8 (8 grading opportunities) ................................. 91
C.7 Session 10: Percentage of Deviation from norm by Instructor 1 (8 grading opportunities) ................................. 92
C.8 Session 10: Percentage of Deviation from norm by Instructor 8 (8 grading opportunities) ................................. 92
C.9 All Sessions: Percentage of Deviation from norm by Instructor 1 (1341 grades) .................................................. 93
C.10 All Sessions: Percentage of Deviation from norm by Instructor 8 (1115 grades) .................................................. 93
List of Software Used


Appendix A

Competency Tables
<table>
<thead>
<tr>
<th>item-name</th>
<th>group-name</th>
<th>group-id</th>
<th>item-id</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ignore</td>
<td>Ignore - group</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Electrical Power Up/Down</td>
<td>Supplementary Procedures (Read and do)</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>FMS and AFDS use</td>
<td>Normal Procedures</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Take off</td>
<td>Manual Flight Maneuvers</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Normal procedures on the ground</td>
<td>Normal Procedures</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Briefing</td>
<td>Normal Procedures</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Inflight flows</td>
<td>Normal Procedures</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Manual flight</td>
<td>Manual Flight Maneuvers</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>ILS</td>
<td>Automatic Flight Maneuvers</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>LNAV/VNAV</td>
<td>Automatic Flight Maneuvers</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>V/S</td>
<td>Automatic Flight Maneuvers</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Autoland</td>
<td>Normal Procedures</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>CRM</td>
<td>CRM (LTM, PSD, WLM)</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>LTW</td>
<td>Leadership and Teamwork</td>
<td>101</td>
<td>101</td>
</tr>
<tr>
<td>PSD</td>
<td>Problem Solving and Decision Making</td>
<td>102</td>
<td>102</td>
</tr>
<tr>
<td>WLM</td>
<td>Workload Management</td>
<td>103</td>
<td>103</td>
</tr>
<tr>
<td>SAW</td>
<td>SITAW</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>COM</td>
<td>Communication</td>
<td>105</td>
<td>105</td>
</tr>
<tr>
<td>ILS raw data</td>
<td>Manual Flight Maneuvers</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>Circle-to-land</td>
<td>Automatic Flight Maneuvers</td>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>Go Around</td>
<td>Manual Flight Maneuvers</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>TCAS</td>
<td>Recall Maneuvers</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>Pilot Incapacitation</td>
<td></td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td>SOP</td>
<td>Normal Procedures</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>Engines start - apu inop</td>
<td>Supplementary Procedures (Read and do)</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>De-icing</td>
<td>Supplementary Procedures (Read and do)</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>Severe Turbulence</td>
<td>Supplementary Procedures (Memory)</td>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td>Airspeed Unreliable</td>
<td>Non-Normals - complex</td>
<td>6</td>
<td>22</td>
</tr>
<tr>
<td>Minor Failure - other</td>
<td>Non-Normals - minor</td>
<td>5</td>
<td>23</td>
</tr>
<tr>
<td>Major Failure - other</td>
<td>Non-Normals - complex</td>
<td>6</td>
<td>24</td>
</tr>
<tr>
<td>RTO</td>
<td>Recall Maneuvers</td>
<td>11</td>
<td>25</td>
</tr>
<tr>
<td>V1-cut</td>
<td>Non-Normals - one engine inop</td>
<td>7</td>
<td>26</td>
</tr>
<tr>
<td>One eng inop landing</td>
<td>Non-Normals - one engine inop</td>
<td>7</td>
<td>27</td>
</tr>
<tr>
<td>One eng inop app</td>
<td>Non-Normals - one engine inop</td>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>One eng inop GA</td>
<td>Non-Normals - one engine inop</td>
<td>7</td>
<td>29</td>
</tr>
<tr>
<td>Engine fail - other</td>
<td>Non-Normals - one engine inop</td>
<td>7</td>
<td>30</td>
</tr>
<tr>
<td>UPRT</td>
<td>Manual Flight Maneuvers</td>
<td>8</td>
<td>31</td>
</tr>
<tr>
<td>Windshear</td>
<td>Recall Maneuvers</td>
<td>11</td>
<td>32</td>
</tr>
<tr>
<td>GPWS</td>
<td>Recall Maneuvers</td>
<td>11</td>
<td>33</td>
</tr>
<tr>
<td>Hydraulics</td>
<td>Non-Normals - complex</td>
<td>6</td>
<td>34</td>
</tr>
<tr>
<td>Flight Controls</td>
<td>Non-Normals - complex</td>
<td>6</td>
<td>35</td>
</tr>
<tr>
<td>Decompression</td>
<td>Non-Normals - complex</td>
<td>6</td>
<td>36</td>
</tr>
<tr>
<td>Evacuation</td>
<td>Non-Normals - minor</td>
<td>5</td>
<td>37</td>
</tr>
<tr>
<td>LVO - Normal</td>
<td>LVO</td>
<td>10</td>
<td>38</td>
</tr>
<tr>
<td>LVO - Non normals</td>
<td>LVO</td>
<td>10</td>
<td>39</td>
</tr>
</tbody>
</table>

Table A.1: Overview of Scenario Elements and Tasks with Their Allocation to Groups
<table>
<thead>
<tr>
<th>ID</th>
<th>Competency</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Application of knowledge (KNO)</td>
<td>Demonstrates knowledge and understanding of relevant information, operating instructions, aircraft systems, and the operating environment</td>
</tr>
<tr>
<td>1</td>
<td>Application of procedures and compliance with regulations (PRO - previously APK)</td>
<td>Identifies and applies appropriate procedures in accordance with published operating instructions and applicable regulations.</td>
</tr>
<tr>
<td>2</td>
<td>Communication (COM)</td>
<td>Communicates through appropriate means in the operational environment, in both normal and non-normal situations.</td>
</tr>
<tr>
<td>3</td>
<td>Aeroplane Flight Path Management, automation (FPA)</td>
<td>Controls the flight path through automation.</td>
</tr>
<tr>
<td>4</td>
<td>Aeroplane Flight Path Management, manual control (FPM)</td>
<td>Controls the flight path through manual control</td>
</tr>
<tr>
<td>6</td>
<td>Leadership and Teamwork (LTW)</td>
<td>Influences others to contribute to a shared purpose. Collaborates to accomplish the goals of the team</td>
</tr>
<tr>
<td>7</td>
<td>Problem Solving and Decision-making (PSD)</td>
<td>Identifies precursors, mitigates problems, and makes decisions</td>
</tr>
<tr>
<td>8</td>
<td>Situation awareness and Management of Information (SAW)</td>
<td>Perceives, comprehends and manages information and anticipates its effect on the operation</td>
</tr>
<tr>
<td>9</td>
<td>Workload Management (WLM)</td>
<td>Maintains available workload capacity by prioritising and distributing tasks using appropriate resources</td>
</tr>
</tbody>
</table>

*Table A.2: Competency Descriptions (adapted from (EASA, 2023b))*
Appendix B

Comments Submitted to EASA

As part of this work, ten comments on NPA 2018-07(B) Update of ORO.FC — evidence-based training subtask were submitted to EASA. Four comments were accepted by EASA, one partially accepted, five were noted, none rejected.


Additionally, 12 comments were submitted to NPAs 2020-02, 2020-14 and 2021-05. Comments and responses can be found in the respective Comment Response Documents published by EASA.
<table>
<thead>
<tr>
<th>Comment</th>
<th>592</th>
<th>Comment by: <strong>Czech Technical University</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>We appreciate EASA work on the evidence-based training subtask. We support voluntary implementation of EBT. Further expansion of EBT to OCC, initial type rating, and eventually even to initial pilot training is deemed desirable.</td>
<td></td>
</tr>
</tbody>
</table>

| Response | Noted |

<table>
<thead>
<tr>
<th>Comment</th>
<th>358</th>
<th>Comment by: <strong>Czech Technical University</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>ARO.OPS.226 point (c)(1)</strong> We suggest a development of GM. Even slight variation in internal policies may cause difficulties for airlines holding several AOCs in different member states and attempting to implement a common EBT programme.</td>
<td></td>
</tr>
</tbody>
</table>

| Response | Partially accepted  
The text has been modified to ensure clarity. See the EASA Opinion. |

<table>
<thead>
<tr>
<th>Comment</th>
<th>359</th>
<th>Comment by: <strong>Czech Technical University</strong></th>
</tr>
</thead>
</table>
|         | **AMC1 ARO.OPS.226(d) point (c)(3) wording ‘relevance of the operator’s approved EBT programme’**  
Terms ‘relevance’ and ‘effectiveness’ are clear and do not require further explanation. |

| Response | Noted |

<table>
<thead>
<tr>
<th>Comment</th>
<th>360</th>
<th>Comment by: <strong>Czech Technical University</strong></th>
</tr>
</thead>
</table>
|         | **AMC1 QARO.OPS.226(d) point (c)(10) wording ‘Continuing standardisation of EBT instructors’**  
Verification of concordance by an authority may be limited to inspecting outputs from the instructor concordance assurance programme. Authority oversight of the instructor concordance assurance programme (and relevant metrics) is required by AMC1 ARO.OPS.226(d) point (c)(11).  
Decay prevention should be addressed by the operator. |

| Response | Accepted  
The new AMC1 ORO.FC.231(a)(4) and GM1 ORO.FC.231(a)(4) have been developed in the AMC & GM to Part-ORO to clarify this point. The authority would need to read the information in the AMC & GM to Part-ORO to avoid duplication of the provision. |

<table>
<thead>
<tr>
<th>Comment</th>
<th>361</th>
<th>Comment by: <strong>Czech Technical University</strong></th>
</tr>
</thead>
</table>
|         | **GM1 to AMC1 ARO.OPS.226(d) point (b)**  
We suggest to include a part of the explanatory note in the GM. Please consider rewording for better clarity as follows:  
(b) *The analysis of the pilot competencies across the domains should also take into account the operator’s experience in the approved EBT programme and the level of difficulty contained within the scenario elements of the programme, which may result in variations of the grading results and those variations are acceptable.* |

| Response | Accepted |

86
comment 363 comment by: Czech Technical University

Our study of grading has identified a 5-level scale as ideal for the purpose of pilot evidence based training. However, use of 'numerical grades' (1 to 5) indicate lower accuracy than use of 'description/named grades' (e.g. Ideal, Effective, Satisfactory, Acceptable, Unsafe). Instructors using 'numerical grades' may tend to 'instructor-evaluator laziness'. This discrepancy can possibly be mitigated by robust instructor training.

response Noted

comment 364 comment by: Czech Technical University

Although we understand the concern, we believe, that operators should be discouraged from use of alternative grading system in a long term.

response Noted

comment 365 comment by: Czech Technical University

This methodology provides high quality data for further analysis and shall be promoted.

response Noted

comment 366 comment by: Czech Technical University

For better clarity: Consider re-ordering table Outcome/How many/How often/How well from grade 5 to 1 to match order in previous tables.

response Accepted

comment 362 comment by: Czech Technical University

FCL.740(b)(1) allows an AOC to provide a refresher training for an expired type rating. However, a refresher for an expired IR can only be done by an ATO. We suggest to allow an AOC to provide both IR and TR refresher. Please consider amending FCL.625(c)(1) as follows:

'go through refresher training at an ATO, or an AOC approved for such refresher, to reach the level of proficiency needed to pass the instrument element of the skill test in accordance with Appendix 9 to this Part; and'

response Accepted
Appendix C

ATO A Grading Data

IATA (2024) published a edition 2 of the Evidence-Based Training Implementation Guide in January 2024 and notes that some of the inter-rater reliability metrics suggested by EASA are not suitable for aviation training data, a finding that aligns with this work. IATA, 2024 suggests the use of thresholds published by Landis and Koch (1977). This is certainly a very efficient approach when during standardization training instructors observe the same content. Goldsmith and Johnson (2002) acknowledges that a primary data source for evaluating an assessor is the calibration sessions. A crucial aspect of these sessions is that several evaluators observe and assess the same performance of aircrew. However there are certain risks when preparing these sessions. It is commonly agreed that experts have conflicting options and it is difficult to establish ground truth. Wong et al. (2021) warns that cherry-picked examples or removing conflicting options, because it destroys the statistical properties of random sample. The cross-replication reliability (xRR) developed by Wong et al. (2021) may present even better insight when comparing results from standards instructors preparing the content and instructors trained for standardization.

There is no method suggested by IATA to assess the reliability of live grading data and the approach suggested by this work for type rating training can possibly be extended to the EBT program. This appendix provides selected ATO A grading outputs in a similar tabular format as recommended by IATA, 2024.
### Table C.1: Distribution of norm Grades throughout the Course

<table>
<thead>
<tr>
<th>Competency / Grade</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00%</td>
<td>5.88%</td>
<td>23.53%</td>
<td>17.65%</td>
<td>52.94%</td>
<td>0.00%</td>
</tr>
<tr>
<td>2</td>
<td>0.00%</td>
<td>5.88%</td>
<td>5.88%</td>
<td>88.24%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>3</td>
<td>5.88%</td>
<td>5.88%</td>
<td>11.76%</td>
<td>23.53%</td>
<td>52.94%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4</td>
<td>47.06%</td>
<td>0.00%</td>
<td>11.76%</td>
<td>29.41%</td>
<td>11.76%</td>
<td>0.00%</td>
</tr>
<tr>
<td>5</td>
<td>0.00%</td>
<td>5.88%</td>
<td>11.76%</td>
<td>64.71%</td>
<td>17.65%</td>
<td>0.00%</td>
</tr>
<tr>
<td>6</td>
<td>0.00%</td>
<td>5.88%</td>
<td>11.76%</td>
<td>82.35%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>7</td>
<td>0.00%</td>
<td>5.88%</td>
<td>0.00%</td>
<td>88.24%</td>
<td>5.88%</td>
<td>0.00%</td>
</tr>
<tr>
<td>8</td>
<td>0.00%</td>
<td>5.88%</td>
<td>11.76%</td>
<td>64.71%</td>
<td>17.65%</td>
<td>0.00%</td>
</tr>
<tr>
<td>9</td>
<td>0.00%</td>
<td>5.88%</td>
<td>11.76%</td>
<td>76.47%</td>
<td>5.88%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

### Table C.2: Trainee 35: Percentage of Grades per Competency throughout the Course

<table>
<thead>
<tr>
<th>Competency / Grade</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00%</td>
<td>5.88%</td>
<td>17.65%</td>
<td>17.65%</td>
<td>58.82%</td>
<td>0.00%</td>
</tr>
<tr>
<td>2</td>
<td>0.00%</td>
<td>5.88%</td>
<td>5.88%</td>
<td>88.24%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>3</td>
<td>0.00%</td>
<td>5.88%</td>
<td>11.76%</td>
<td>29.41%</td>
<td>52.94%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4</td>
<td>52.94%</td>
<td>0.00%</td>
<td>11.76%</td>
<td>17.65%</td>
<td>17.65%</td>
<td>0.00%</td>
</tr>
<tr>
<td>5</td>
<td>0.00%</td>
<td>5.88%</td>
<td>11.76%</td>
<td>47.06%</td>
<td>35.29%</td>
<td>0.00%</td>
</tr>
<tr>
<td>6</td>
<td>0.00%</td>
<td>5.88%</td>
<td>11.76%</td>
<td>76.47%</td>
<td>5.88%</td>
<td>0.00%</td>
</tr>
<tr>
<td>7</td>
<td>0.00%</td>
<td>5.88%</td>
<td>0.00%</td>
<td>82.35%</td>
<td>11.76%</td>
<td>0.00%</td>
</tr>
<tr>
<td>8</td>
<td>0.00%</td>
<td>6.25%</td>
<td>12.50%</td>
<td>62.50%</td>
<td>18.75%</td>
<td>0.00%</td>
</tr>
<tr>
<td>9</td>
<td>0.00%</td>
<td>5.88%</td>
<td>11.76%</td>
<td>70.59%</td>
<td>11.76%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Competency / Grade</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>-------------------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>1</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>2</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>3</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>5</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>6</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>7</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>8</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>9</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

*Table C.3: Session 10: norm Grades per Competency*

<table>
<thead>
<tr>
<th>Competency / Grade</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>2.38%</td>
<td>97.62%</td>
<td>0.00%</td>
</tr>
<tr>
<td>2</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>73.81%</td>
<td>26.19%</td>
<td>0.00%</td>
</tr>
<tr>
<td>3</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>4.76%</td>
<td>92.86%</td>
<td>2.38%</td>
</tr>
<tr>
<td>4</td>
<td>0.00%</td>
<td>0.00%</td>
<td>40.48%</td>
<td>50.00%</td>
<td>9.52%</td>
<td>0.00%</td>
</tr>
<tr>
<td>5</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>76.19%</td>
<td>23.81%</td>
<td>0.00%</td>
</tr>
<tr>
<td>6</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>73.81%</td>
<td>26.19%</td>
<td>0.00%</td>
</tr>
<tr>
<td>7</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>83.33%</td>
<td>16.67%</td>
<td>0.00%</td>
</tr>
<tr>
<td>8</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>78.05%</td>
<td>21.95%</td>
<td>0.00%</td>
</tr>
<tr>
<td>9</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>80.95%</td>
<td>19.05%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

*Table C.4: Session 10: Percentage of Grades per Competency*
Table C.5: Session 10: Percentage of Grades per Competency by Instructor 1 (8 grading opportunities)

<table>
<thead>
<tr>
<th>Competency / Grade</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>2</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.50%</td>
<td>12.50%</td>
<td>0.00%</td>
</tr>
<tr>
<td>3</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4</td>
<td>0.00%</td>
<td>0.00%</td>
<td>25.00%</td>
<td>75.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>5</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.50%</td>
<td>12.50%</td>
<td>0.00%</td>
</tr>
<tr>
<td>6</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.50%</td>
<td>12.50%</td>
<td>0.00%</td>
</tr>
<tr>
<td>7</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>8</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.50%</td>
<td>12.50%</td>
<td>0.00%</td>
</tr>
<tr>
<td>9</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.50%</td>
<td>12.50%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

Table C.6: Session 10: Percentage of Grades per Competency by Instructor 8 (8 grading opportunities)

<table>
<thead>
<tr>
<th>Competency / Grade</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>2</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.50%</td>
<td>12.50%</td>
<td>0.00%</td>
</tr>
<tr>
<td>3</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.50%</td>
<td>12.50%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>5</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>6</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>7</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>8</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>9</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Competency / Grade</td>
<td>-5</td>
<td>-4</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>-------------------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>1</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>2</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.50%</td>
</tr>
<tr>
<td>3</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>4</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>25.00%</td>
</tr>
<tr>
<td>5</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.50%</td>
</tr>
<tr>
<td>6</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.50%</td>
</tr>
<tr>
<td>7</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>8</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.50%</td>
</tr>
<tr>
<td>9</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.50%</td>
</tr>
</tbody>
</table>

Table C.7: Session 10: Percentage of Deviation from norm by Instructor 1 (8 grading opportunities)

<table>
<thead>
<tr>
<th>Competency / Grade</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>2</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.50%</td>
<td>12.50%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>3</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.50%</td>
<td>12.50%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>5</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>6</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>7</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>8</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>9</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

Table C.8: Session 10: Percentage of Deviation from norm by Instructor 8 (8 grading opportunities)
<table>
<thead>
<tr>
<th>Competency / Grade</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>92.74%</td>
<td>7.26%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>2</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>91.94%</td>
<td>8.06%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>3</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>99.19%</td>
<td>0.81%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>96.77%</td>
<td>2.42%</td>
<td>0.00%</td>
<td>0.81%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>5</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>96.77%</td>
<td>2.42%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>6</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.81%</td>
<td>96.77%</td>
<td>2.42%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>7</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.81%</td>
<td>98.39%</td>
<td>0.81%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>8</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.81%</td>
<td>95.12%</td>
<td>4.07%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>9</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.81%</td>
<td>96.77%</td>
<td>2.42%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

*Table C.9: All Sessions: Percentage of Deviation from norm by Instructor 1 (1341 grades)*

<table>
<thead>
<tr>
<th>Competency / Grade</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>86.58%</td>
<td>13.42%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>2</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>83.89%</td>
<td>16.11%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>3</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.67%</td>
<td>89.93%</td>
<td>9.40%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>77.18%</td>
<td>22.82%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>5</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>80.54%</td>
<td>18.79%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>6</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>81.21%</td>
<td>18.12%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>7</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>88.59%</td>
<td>11.41%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>8</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>77.85%</td>
<td>22.15%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>9</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>83.22%</td>
<td>16.78%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

*Table C.10: All Sessions: Percentage of Deviation from norm by Instructor 8 (1115 grades)*