This project has received funding from the European Union's Horizon Europe Programme

RESEARCH PROJECT EASA.2022.HVP.01

D-3.3.1 STANDARDISED METRICS AND METHODS FOR INSTRUCTOR CONCORDANCE ASSURANCE

Digital transformation - Case studies for aviation safety standards – Data Science Applications (DATAPP)

An Agency of the European Union

Disclaimer

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Union Aviation Safety Agency (EASA). Neither the European Union nor EASA can be held responsible for them.

DELIVERABLE NUMBER AND TITLE:	DATAPP D-3.3.1 Standardised metrics and methods for instructor concordance assurance
CONTRACT NUMBER:	EASA.2022.HVP.01
CONTRACTOR / AUTHOR:	CAAi/ Tim Ramsdale, Di Lintott, Chris Whitehurst, Pol Niño
IPR OWNER:	European Union Aviation Safety Agency
DISTRIBUTION:	Public

APPROVED BY:	**AUTHOR**	**REVIEWERS**	**MANAGING DEPARTMENT**
	CAA International (CAAi)	**B. Briguglio - CAAi**	**CAAi**
		Andrada Bujor - ALG	**ALG**
		Alex Olivera - ALG	
		Francisco Arenas	**EASA**
		SPT 0012 Task Force	

DATE:31 July 2024

SUMMARY

Problem area

Digitalisation is reshaping the aviation business at quick pace, bringing efficiency and wider opportunities to manage information. The deployment of digital solutions throughout the air transport industry is a fact and brings significant changes to the traditional working processes, business models, standards and regulations.

EASA faces new challenges on what the required changes in safety standards and regulations are needed in response to the introduction of innovative solutions and processes. Anticipating what is to come in the industry in the field of data science applications is key to make sure safety levels are maintained without slowing innovation down.

The objective of this project is to identify and assess relevant changes to the existing aviation safety standards to support the deployment of the digital solutions under three case studies:

- Case Study 3: Flight training data for EBT/CBTA (Evidence-Based Training / Competence-Based Training and Assessment)
- Case Study 4: Digital fuel management
- Case Study 5: Flight data models for safety

The project aims to provide a comprehensive evaluation of benefits, constraints, standardisation and deployment issues, including the recommendations for adjusting safety regulations and related standards, and how new digital technologies could contribute to addressing the identified issues.

Description of work

The present document relates to 'D-3.3 Training Materials, part of the EASA.2022.HVP.01- Horizon Europe Project. The aim is to provide guidelines promoting best practices for the management of pilot grading data with the objective of improving instructor concordance. The focus of this report is:

1. Best-practice for standardised metrics and methods to assess agreement and alignment for instructor concordance.
2. Definition of a mechanism to assess the evolution of concordance over time, and to prevent/mitigate against over- or under-grading.

This document also contains data protection guidelines, which although generic in nature, are highly relevant for training organisations when handling training related data.

Results and Application

The report delves into one of the solutions proposed in the context of the project, providing further details and a series of recommendations that could be applied to achieve an effective implementation of the solution. All of this is collected and provided in the form of training materials. Such training materials are intended to be used at EASA's discretion, for instance by including it in dissemination documents or in guidance material to help in the potential implementation of the solutions by the stakeholders. Thus, the output of this document provides additional information to EASA to support their decision on the evolution of the solutions proposed in the context of this project.

CONTENTS

ABBREVIATIONS

ACRONYM	DESCRIPTION
AltMoC	Alternative Means of Compliance
AMC	Acceptable Means of Compliance
ATO	Approved Training Organisation
APK	Application of Procedures competency, superseded by PRO
CAA	Civil Aviation Authority
CA	Competent Authority
CAP	Civil Aviation Publication
CAT	Commercial Air Transport
CBTA	Competency Based Training and Assessment
COM	Communication competency
EASA	European Union Aviation Safety Agency
EBT	Evidence Based Training
EBTI	EBT Instructor
EVAL	Evaluation
FPA	Flight path management – automation competency
FPM	Flight path management – manual control competency
FSTD	Flight Simulation Training Device
GDPR	General Data Protection Regulation
GM	Guidance Material
IATA	International Air Transport Association
ICAO	International Civil Aviation Organisation
ICAP	Instructor Concordance Assurance Programme
IE	Instructor Evaluator
IQR	Inter-Quartile Range
KNO	Application of knowledge competency
KPI	Key Performance Indicator
LTW	Leadership and teamwork competency
NAA	National Aviation Authority
OB	Observable Behaviour
OM	Operations Manual
PRO	Application of procedures and compliance with regulations competency
PSD	Problem solving and decision making competency
SARPs	Standards and Recommended practises
SAW	Situation awareness and management of information competency
SFI	Synthetic Flight Instructor
SPI	Safety Performance Indicator
TRI	Type Rating Instructor
WLM	Workload management competency

1. Introduction

1.1 Background

Digitalisation is reshaping the aviation business at quick pace, bringing efficiency and wider opportunities to manage information. The deployment of digital solutions throughout the air transport industry is a fact and brings significant changes to the traditional working processes, business models, standards and regulations.

In its role of EU Aviation Safety Regulator, EASA faces new challenges on what the required changes in safety standards and regulations are needed in response to the introduction of innovative solutions and processes. Anticipating what is to come in the industry in the field of data science applications is key to make sure safety levels are maintained without slowing innovation down. For that, identifying the key main applications in that area in the form of case studies, allows to better picture us in what is to come and will allow translating that future into recommendations for standardisation and regulations.

This project aims at evaluating a series of changes applied to aviation products, processes and operations resulting from the deployment of new digital solutions with a focus on measuring the impact on safety standards and regulatory materials as well as to prepare their evolutions. The project is built upon three case studies allowing to develop a comprehensive investigation of the key changes at stake:

- Case Study 3: Flight training data for EBT/CBTA. The case study will encompass the development of comprehensive guidelines for moving towards the implementation of EBT and CBTA concepts.
- Case Study 4: Digital fuel management. The project will encompass the in-depth analysis of the benefits and constraints associated to state-of-the-art digital solutions for fuel management, considering the current safety issues reported, as well as the preparation of comprehensive documentation to support the proposed evolution of standards and regulatory requirements.
- Case Study 5: Flight data models for safety. The proposed case study will investigate the development of comprehensive data models 'bridging' between the flight data sources and their use for the operator's safety-relevant processes and for industry-wide data exchange program.

1.2 Scope of the report

This report represents one of the deliverables under the task "T-3.3 Training material" of "Digital Transformation – Case Studies for Aviation Safety Standards" project (EASA.2022.HVP.01- Horizon Europe Project). D-3.3 is complemented by 5 (five) individual deliverables covering the different case studies of the project, presented in Table 1-1 below.

Table 1-1 List of deliverables complementing D-3.3

Deliverable		Title	Case Study
D – 3.3	**D-3.3.1**	**Standardised metrics and methods for instructor concordance assurance**	**CS3 - EBT**
	D-3.3.2	Untapped benefit of fuel reduction schemes: Reviewing the NPA-2016-06(A) economic impact assessment	CS4 - Fuel
	D-3.3.3	Recommendations on assurance framework for analytical development and approval of fuel schemes	
	D-3.3.4	Training requirements for FDM analyst	CS5 - FDM
	D-3.3.5	Development of industry-agreed FDM algorithms and logics	

Within the context of the DATAPP project, the D-3.3.x deliverables are intended to provide dissemination material designed to concretise some of the solutions proposed during the project, particularly those that could represent potential quick-wins. Such materials are intended to be used at EASA's discretion, for instance by including it in dissemination documents or in guidance material.

This deliverable “D-3.3.1 Standardised metrics and methods for instructor concordance assurance” provides proposals for amending the EBT Manual which will supplement AMC and Guidance material and give advice on best practices for the management of EBT training data particularly regarding instructor concordance.

The present document is structured as follows:

- Section 2 as a brief introduction with recommendations for incorporation of training material into the EBT Manual.
- Section 3 comprises the training material which addresses the solutions and conclusions extracted under Case Study 3 Flight training data for EBT/CBTA.
- Section 4 sets out recommendations for data protection.

2. Creation of Training Material

2.1 Approach to the creation of Training Material

Recommendations within this document are undoubtedly complimentary to some of the existing content of the EASA EBT MANUAL V2.2, however training material addressing the objectives is provided from a holistic perspective rather than attempting to identify how EASA may promulgate this additional material.
The primary objectives were specified as:

1. To develop best-practices for standardised metrics and methods to assess agreement and alignment for instructors' concordance.
2. To define a mechanism to assess the concordance evolution over time, preventing from the appearance of overgrading or undergrading.

During the collation of the training material, it became clear that the objectives are complimentary and that success in adopting the best practices will lead to the prevention of concordance drift with time. Therefore, both objectives are addressed in one set of training material, rather than two.

2.2 Recommendations

The training material is presented as a single proposal for content sub-divided into sections by subject heading. Two options are suggested:

- **Option 1 – Recommended**
  Incorporate the whole content as a separate document “Best practice guidance for achieving instructor concordance”. Supplementary material such as explanatory videos may be produced to enhance the guidance.
- **Option 2**
  Extract text from this document and incorporate it into existing publications such as the EBT Manual.

3. Training Material

3.1 Overview

The key to the success of any training programme is the competence of the instructors and their ability to effectively evaluate pilot performance. In an EBT programme, there are primarily two outputs from the EBT instructors:

1. The actual enhancement of standards provided directly to the pilots; and
2. The data output from the instructors via the grading system which is used to provide the training system with feedback on pilot competency.

The purpose of grading is partly to provide useful measurements of individual performance. If the grading of pilots is not accurate and therefore the data produced by the instructors is not valid, the ability of the organisation to identify individual and collective training needs, to review, develop and improve the training programme and to feed information to the Safety Management System may be compromised.

There are inherent risks associated with the judgement of Grades 1, (NOT COMPETENT) and 2, (minimum acceptable level of competence), which determine whether a pilot may conduct line operations or may require tailored or additional training respectively. Lack of concordance in determining these grades is associated with a potential safety risk and should be considered when assessing if the levels of concordance are acceptable. Concordance assurance metrics should focus on these risks.

The purpose of instructor concordance training and the Instructor Concordance Assurance Programme, (ICAP) is to ensure that instructors are properly trained and identify areas of weak concordance to drive improvements in the quality and validity of the grading system.

Effectively these activities also provide mitigation against the risks associated with invalid training data.

Instructor concordance is therefore critical to the success of EBT and the improvement in pilot performance over time.

The ability of a competent authority inspector to assess the suitability of the EBT programme relies in part on the inspectors understanding of instructor concordance data, especially the metrics which operators use to determine concordance and the criteria which an operator uses to assess the acceptability of agreement and alignment of instructors' assessments.

Whilst the competent authority may recommend criteria to measure concordance standards, care should be exercised due to the potential variability of data, especially within immature EBT operators. If the metrics described in this document are used in the ICAP and inspectors review the ICAP data or report with the operator at least once every cycle, acceptable concordance standards should be assured.

3.2 Type Rating and Synthetic Flight Instructor Training

The quality of instructors is critical to the success of an EBT programme. The selection and training of new instructors before they embark on EBTI training is important in the context of concordance.

Principles of competency-based training and assessment should be embedded wherever possible into all parts of the Type Rating Instructor (TRI) and Synthetic Flight Instructor (SFI) courses, including the pilot competency framework, the evaluation and training of key observable behaviours and root cause analysis linked to competencies. Language use to describe pilot performance should be consistent with the competency framework.

Compliance with FCL.920 must be demonstrated by the Approved Training Organisation (ATO), however reference should also be made to the EBT Instructor Competency Framework described in GM3 ORO.FC.146(c).

3.3 EBT Programme suitability

EBT regulations ORO.FC.146 and ORO.FC.231 point (a)(4) mandate an Instructor Concordance Assurance Programme, it's purpose and how it should function, but the metrics by which it should be judged, minimum KPIs and fundamentally how it can be assessed as suitable and effective are left to the operator to determine and the competence authority to agree.

EBT Approval requires operators to have demonstrated 2 years of an instructor concordance assurance programme within Mixed EBT. [AMC1 ORO.FC.231(a)(1)]. This programme should be accepted as suitable *and* effective before the Approval is granted.

3.4 EBT Instructor Training

The EBT Instructor – Initial standardisation programme, [AMC1 ORO.FC.146(c)] isn't required to include dedicated concordance training, however the instructor should be trained to "*evaluate performance using a competency-based grading system"*.

GM1 ORO.FC.146(c) provides guidance related to concordance training, however the Controlled Event concordance method described below should also be used during EBTI Initial training.

3.5 Concordance

3.5.1 Introduction to Concordance

Concordance must be addressed specifically by an operator on at least an annual basis, through the provision of recurrent instructor standardisation courses and the ICAP, as well as individual instructor concordance verification and standardisation observations, (although the latter is not required by regulation).

The annual recurrent EBT instructor standardisation course, which must include concordance training, may be combined with the triennial refresher course for TRI/SFI revalidation, however the operator must ensure that the course is compliant with both the OPS and FCL requirements. AMC2 ORO.FC.146(c) does not specify the content of the recurrent concordance training and GM1 ORO.FC.146(c) gives only limited guidance, but the recurrent concordance training event will include elements of the ICAP, (see ORO.FC.231 and the associated AMC and GM).

When these requirements are combined with the content of the instructor refresher course i.e. AMC1 FCL.940.TRI(a)(1)(ii), (a)(2)(ii), (b)(1)(ii), (b)(2)(ii); FCL.940.SFI(a)(2), (e)(1), the challenge for the operator in combining the two courses to create a single compliant and effective training course will be significant.

It is recommended that operators do not combine these courses until the EBT programme, especially the ICAP, is mature.

GM1 ORO.FC.231(a)(4) makes reference to *"complex groups of instructors"*. Operators which fall in this category may need to adopt additional concordance measures, either by conducting more training events or by observing instructors grading live events on a more frequent basis.

Sub-contracted instructors potentially constitute a higher risk in terms of concordance, especially if they conduct training for multiple organisations and should be avoided by EBT operators.

It states in AMC1 ORO.FC.231(a)(4) that *"The operator should verify the concordance of the instructors …. for a sufficient number of competency-grade combinations."* The question is what constitutes a sufficient number.

For the 'Controlled Content Method', concordance exercises conducted on an annual basis as part of the recurrent standardisation, instructors should grade the pilots using the operator grading system and competency framework, i.e. all competency-grade combinations are available. The instructors should evaluate the pilot performance in the same way as if it were a live event such as an EVAL phase or a Line Evaluation.

Concordance analysis of whole instructor group grading data, the 'Simulator Evaluation Data Method', will use all the competencies, however there will undoubtedly be limited data for the extreme grades: Grades 1 and 5 on a five-point scale; Grade 1 on a four-point scale and Grade 2 on both scales. Analysis should therefore focus on Grades 3 and 4 from which the grading behaviour of the instructors can be inferred across the whole grading scale.

3.5.2 Grading Systems

A 5-point Grading System should be adopted for EBT, [AMC1 ORO.FC.231(d)(1)]. Its usage should be encouraged due to the wider range of evaluation possibilities it offers, but also for the personal satisfaction and motivation it can provide to pilots. However, AMC2 ORO.FC.231(d)(1) directs operators seeking to develop an alternative grading system to apply for an AltMoC and the EBT Manual *SECTION I: EBT Implementation Material* suggests that a 4-point scale, maintaining the equivalence of Grades 1,2 and 3, but combining Grades 4 and 5 into a single grade, may be accepted by competent authorities.

Operators should develop grading guidance to help the instructors determine the grade of the pilots they assess. [AMC1 ORO.FC.231(d)(1)]. Also, instructors should evaluate the performance of pilots by determining a grade for each competency *using a methodology defined by the operator.* The VENN model includes word pictures for each grade within each competency, but also allows the grade descriptions to be simplified once instructors become familiar with the system.

Operators may simplify the word pictures of grades so that a single definition may be given for each grade. An example is:

- Grade 3: Company Standard - The competency was applied correctly, regularly demonstrating most of the relevant behaviours when required, which resulted in a safe operation.
- When grading EVAL phases, as well as recording and classifying all observed pilot behaviour, instructors should be encouraged to identify key OBs linked to the root cause of the performance of each competency. These are important in maintaining instructor concordance as they are the key to determining ***why and how*** a pilot performed at the assessed grade. Examples of the number of OBs that should be recorded are in the Table 3-1 below using the standard Venn Model abbreviated word pictures.

Table 3-1 Observable behaviours using standard Venn model

Grade	Outcome	How well?	How many?	How often?	OBs to select
5	Enhanced safety, effectiveness and efficiency	in an exemplary manner	Almost all	Always	Those demonstrated in an exemplary or effective manner which positively contributed to the outcome. Minimum none, maximum 2.
4	Safe	effectively	Most	Regularly	
3	Safe	adequately	Many	Regularly	None (Company standard grade)
2	Not Unsafe	minimally acceptable	Some	Occasionally	Those that were missing, (which could have significantly affected the outcome) or those which were the root cause. Minimum 1, maximum 2.
1	Unsafe situation	ineffectively	Few	Rarely	

These should also be recorded in the training database to provide information to measure training system performance and develop the training programme.

ORCA is the recommended method for the conduct of the grading, [AMC3 ORO.FC.231(d)(1)].

In the early stages of the evolution of the EBT programme, instructors should be encouraged to record their observations[1] in plain language rather than trying to identify OBs, which would likely be distracting and add to workload. Classification links the instructors notes to the OBs. Similarly, the debrief should use the pilots' own language rather than trying to force the discussion around the OBs which may not be wholly familiar.

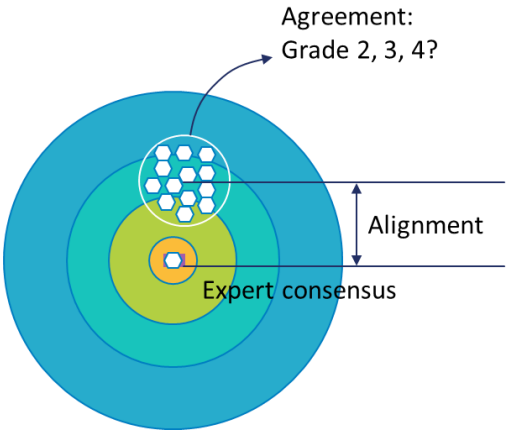
Concordance is likely to be enhanced if the instructors (and the pilots) are encouraged to routinely use the phraseology of the OBs, for example, rather than use the phrase "descent management", the phrase "intended flight path" should be used. This directly links the discussion to OBs 3.2, 3.4, 4.2, 4.5 and also 3.3 and 4.4.

There may be software available in the market, that are intended to be use in the simulator. For example, to display the planned syllabus, or for the recording of pilot behaviours etc. However, these systems should be considered in the context of a detailed evaluation, and when the operator is mature enough to understand the trades of introducing software in the simulator and has established criteria about what are good and bad practices when conducting a simulator session. When not properly implemented, they are likely to lead to distraction, higher workload and, most importantly, carry the risk of forcing concordance by putting pressure on an instructor to identify OBs whilst watching the pilots perform. It is recommended that operators seek expert assistance in this area and the use of third parties may be necessary.

3.5.3 Agreement and Alignment

The operator's ICAP should include data analysis demonstrating instructor-group assessment homogeneity (agreement) and instructor assessment accuracy (alignment). These concepts can be illustrated using a target analogy.

Figure 3-1 Representation of agreements and alignment (source: EASA DATAPP Workshop on 14th November 2023)



[1] Data protection must be ensured. The instructor's observation notes should not be stored in the operator's database. Although not recommended, it could be stored if they are properly de-identified first. The risks of storing such observations are amongst others: (in addition of a possible violation of the GDPR), it may be a non-structured and/or non-standardised data, it may contain sensible information of the crew and/or instructors themselves, etc. The Intention of these recordings are to save them for the use in the debriefing of the session and then they should be deleted after the instructor has completed the grading (level 0, 1, 2 and/or 3 as required)

Each white dot within the white circle represents an instructor's grade assigned to a pilot competency and the bullseye represents the 'expert consensus'.

Agreement: the instructors are in close agreement regarding the grade, whatever it might be.

Agreement is paramount at the Level 0 grading metric level, (Overall Grade 1) i.e., is the pilot competent to conduct line operations? Grade 2 agreement is significant as it potentially determines whether extra training is required.

Alignment levels identify whether the instructors' grading is aligned to the operator's or regulator's standards.

3.5.4 ICAP Methods

There are two complimentary methods to which reference is made in the regulations:

- Method 1, Instructor-Group Assessment Homogeneity - The Simulator Evaluation Data Method
- Method 2, Instructor Accuracy - The Controlled Content Method

The Simulator Evaluation Data Method has the advantage of using real grading data from live events. It is primarily aimed at assessing agreement and identifying outliers who may be biased but, as the number of instructors increases, it may be used as an indication of alignment, (measured by the spread of the Interquartile Range - the grading behaviour of the large population of instructors is likely to tend towards an expert consensus.)

It can also be utilised to gather data on OBs most commonly identified which is also a measure of agreement.

The Controlled Event Method addresses both agreement and alignment, directly engages with all instructors and involves training to calibrate instructors and improve concordance. It can use training data from previous events to target grading of individual competencies and specific behaviours.

Both methods are described below, creating guidance on the best practices for the methods to assess concordance within the ICAP and identify standardised metrics to assess agreement and alignment.
An effective concordance process using a combination of both these methods will also address individual instructor concordance, [GM1 ORO.FC.231(a)(4)].

3.5.4.1 Instructor Concordance Assurance Method 1: The Simulator Evaluation Data Method

AMC1 ORO.FC.231(a)(4) states that Complex operators should include an ICAP-specific data analysis, demonstrating instructor-group assessment homogeneity (agreement) and that concordance should be verified once every cycle and that the operator should establish procedures to address those instructors who do not meet the standards required.

This method addresses these requirements and includes elements of GM1 ORO.FC.231(a)(4): a continuous assessment of concordance; agreement may be inferred from instructors who have observed the same content; procedures for small groups of instructors.

This method involves the analysis of all grading data from the EVAL phase over a defined period. It compares instructors' grading behaviour against their peers and assesses agreement amongst the instructor community, based on the observations of the same content in a homogenous pilot group, by analysing the grading data for all competencies. Agreement can be inferred from this analysis and potential outliers identified and managed.

Operators should verify concordance at least once every cycle, [AMC1 ORO.FC.231(a)(4)] and therefore this analysis should include 2 EVAL phases within the 2 modules of the cycle. It is recommended that the analysis takes place after every module.

NOTE 1: Operators will require appropriate software to collate and analyse large amounts of data. Spreadsheet programmes can be utilised to provide effective analysis. Also, the mathematical process involved in the

conversion of the basic grading data, (categorical data) into continuous data and the process for the detailed statistical analysis of the data is beyond the scope of this guidance. It is recommended that operators seek expert assistance in this area and the use of third parties may be necessary.

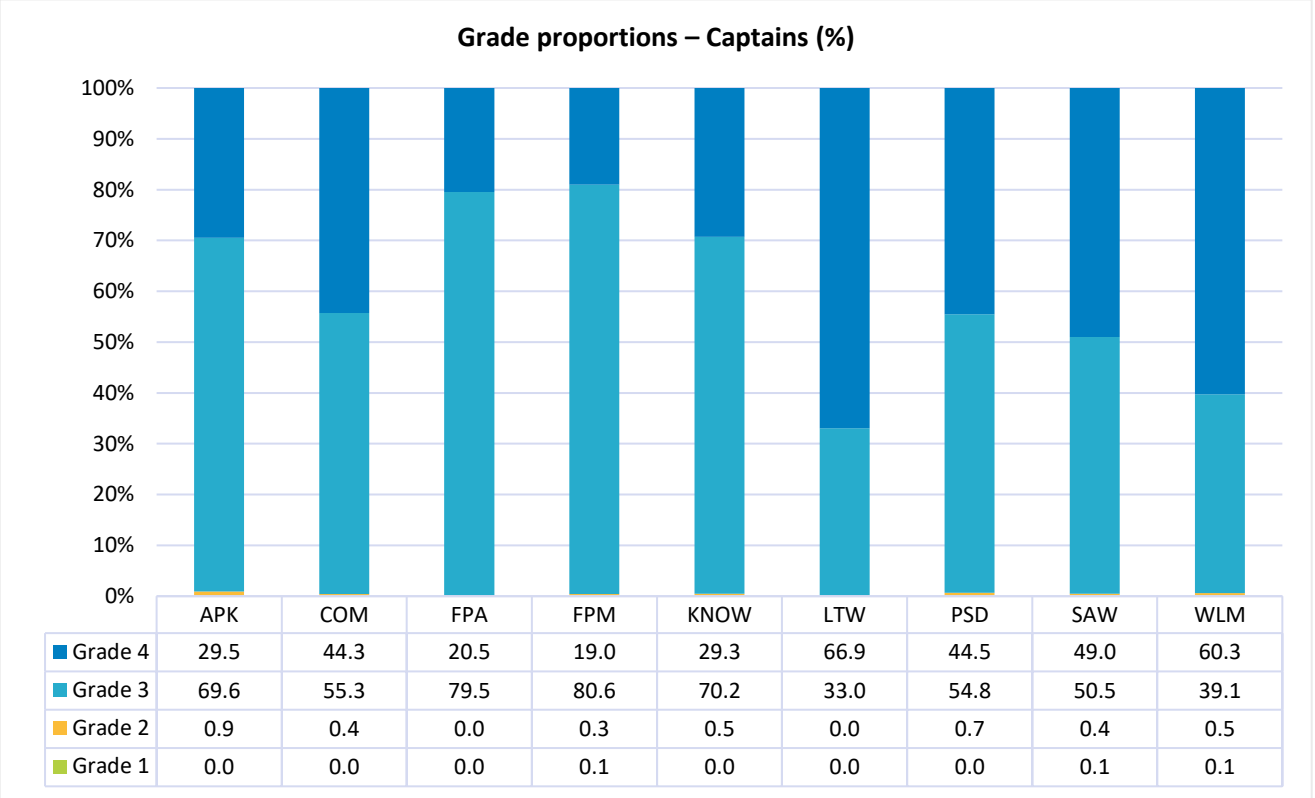
The Simulator Evaluation Data Method - process
Step 1: 'Clean' the data

Remove unwanted data, e.g., incomplete sets of data, corrupted data, data from instructors no longer part of the programme etc.

The data needs to be conditioned to facilitate analysis. Processes written in a programming language such as 'Python' which is a general purpose programming language may be used.

This categorical data, (number of grades 1 to 5 for each competency) may then be summarised and visualised in histogram form, as in the example below.

Figure 3-2 Example of grade proportions



	APK	COM	FPA	FPM	KNOW	LTW	PSD	SAW	WLM
Grade 4	29.5	44.3	20.5	19.0	29.3	66.9	44.5	49.0	60.3
Grade 3	69.6	55.3	79.5	80.6	70.2	33.0	54.8	50.5	39.1
Grade 2	0.9	0.4	0.0	0.3	0.5	0.0	0.7	0.4	0.5
Grade 1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.1

In this example, there are very few Grades 1, 2 and 5, which is likely to be typical of most mature operators. Most of the data, therefore, is Grades 3 and 4.

(Note that this operator is using an adapted competency model based on a previous iteration of the standard competency framework, hence APK is present, and PRO is not.)

This summarises the data for all instructors. If the same analysis and pictorial representation were to be created for each instructor, it could give an indication of potential bias, for example if an instructor assessed a greater percentage of pilots as Grade 4 and 5, it could indicate bias, (overgrading). This method would only be suitable for an operator with a small number of instructors. GM1 ORO.FC.231(a)(4) quotes '10' as an example of a small number, however it is up to the operator to demonstrate to the competent authority that this method is

effective for ensuring instructor concordance. For small numbers of instructors, this method alone could adequately address agreement and discussion of instructors grading behaviour of real EVALs may be incorporated into the recurrent standardisation training, however it is recommended that the anonymity of potentially biased instructors is preserved.

For larger numbers of instructors, further analysis will be required.

Simply calculating the averages of pilot grades, (mean or median), does not adequately capture the distribution or variability of instructor grading behaviour and is an unreliable measure of instructor concordance. It should be avoided, and a more sophisticated method described below should be used.

Step 2: Convert to a continuous data distribution

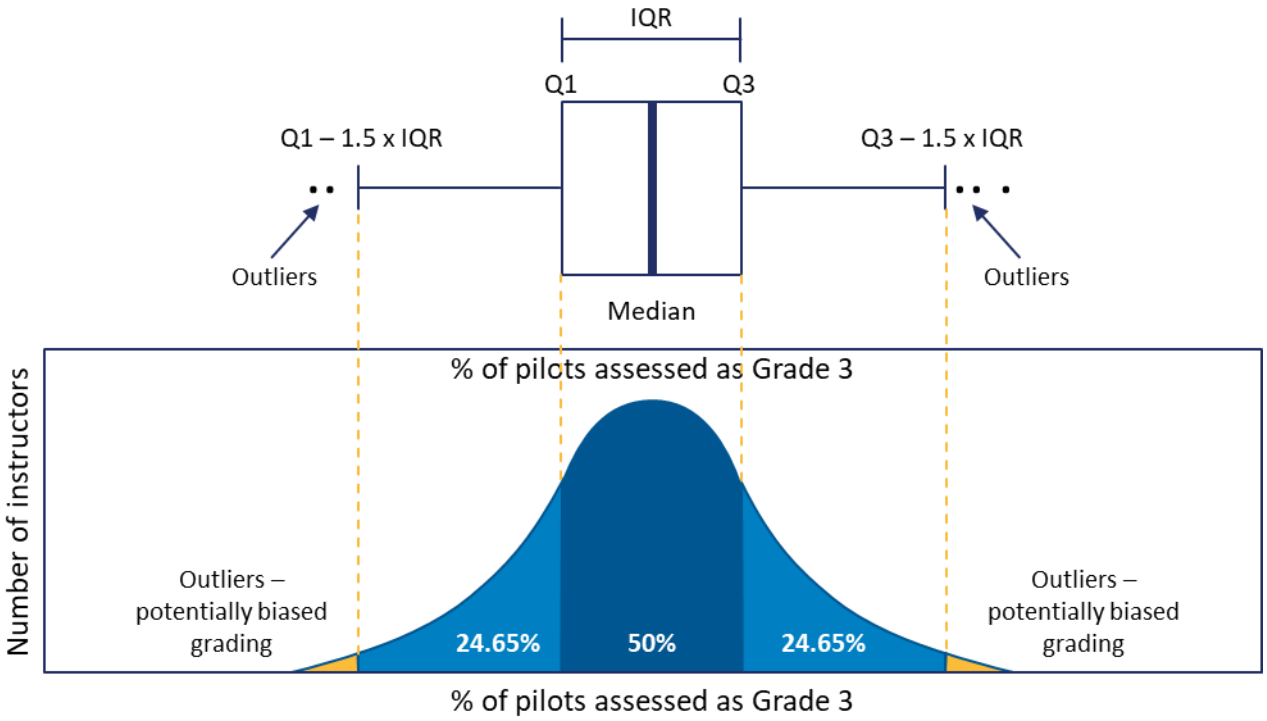
Convert categorical data to continuous data which will prepare the data for accurate and meaningful interpretation. Continuous data is a type of numerical data that can take any value within a given range and, in this context, means that, for a given grade, instructors may assess any percentage of the pilot population, (from 0 and 100%) as performing at that level of competence.

Instructors who have graded less than 30 events, (i.e., pilot pairs in the EVAL phase) should be discarded because, in statistical terms, they may distort the analysis. Operators may reduce this figure if it excludes a significant proportion of instructors but should not reduce the number below 20.

The analysis should be performed for every competency and every grade, resulting in 45 possible combinations, each of which may be visualised as in the graphic below, which is a combination of a distribution curve and a box plot for a single competency at Grade 3.

Identification of potential bias will not require detailed interpretation of every combination but should initially focus on Grades 3 and 4, i.e., 18 combinations (This is discussed in more detail below.)

Figure 3-3 Example of Grade 3 distribution



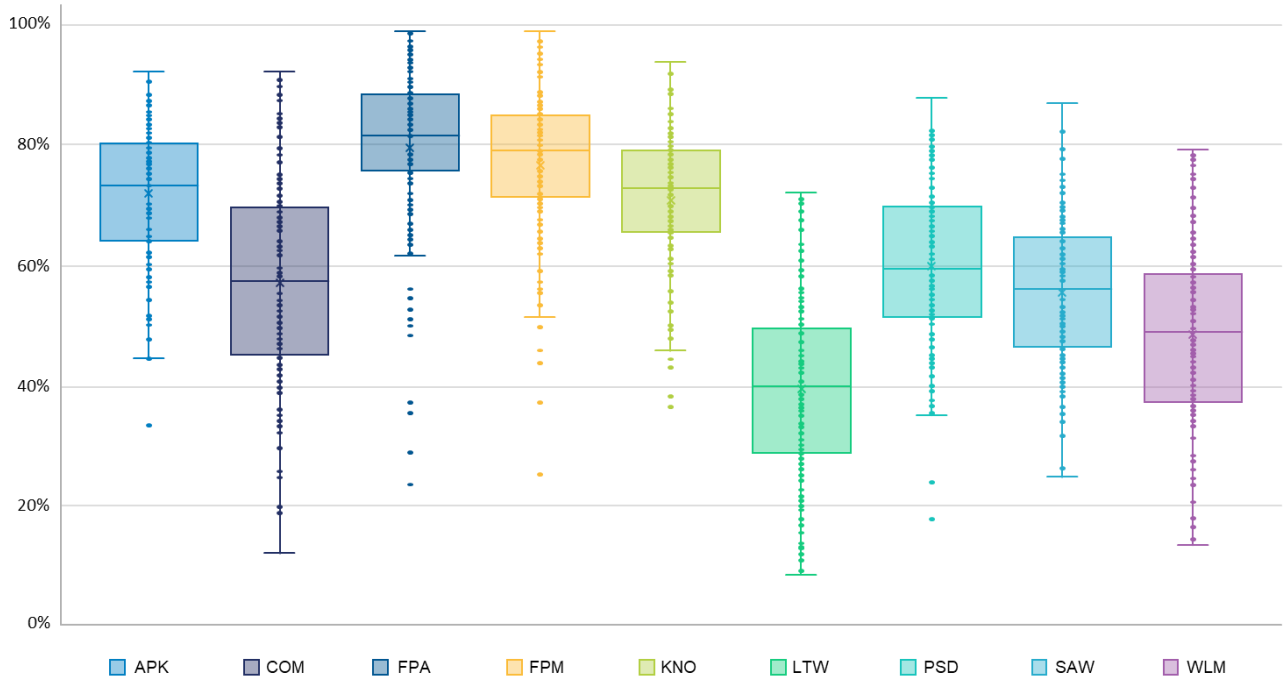
It is unlikely that the distribution will be a normal distribution as in the graphic and it will be skewed to the right or left depending on the grade being analysed.

The main purpose is to identify outliers, i.e., *potentially* biased instructors. The use of 1.5 x the Inter-quartile range, (IQR) is a widely accepted robust statistical way of detecting outliers and identifying these individuals.

In a normal distribution as in the graphic, it can be seen that 97.3% of the data lies within the 1.5 IQR range. The outliers will potentially be quite a small number of instructors, even in a large organisation. For data not following a normal distribution such as instructor grading behaviour, it is possible that outliers will be even fewer in number and concordance analysts may wish to vary the range considered to 1.4 IQR, 1.3 IQR, etc to capture more potential bias.

The results discussed above may also be visualised using a box plot summarising all competencies for each grade. An example is reproduced below.

Figure 3-4 Instructor outliers per competency – Grade 4 (Note APK is present in place of PRO)



Outliers are clearly visible in the box plot.

Also of note is the variation in IQR, (the coloured box). The competencies FPA, FPM and KNO have a smaller IQR which is an indication of concordance. These technical competencies are easier to assess as the OBs may be related to more binary evidence, e.g., OB 4.5 *"Maintains the intended flight path … ",* can be judged using heading, speed, altitude etc. The more non-technical competencies such as COM in this example, has a much broader IQR which may indicate less concordance and training for grading of this competency may need to be included in the next recurrent training event.

Similar variance is reflected in the spread of 1.5 IQR, i.e., the data points between the 'whiskers'.

The next step focuses on those instructors whose grading results lie outside 1.5 x the IQR. As mentioned previously, the majority of the data will be for Grades 3 and 4 and these should be the focus of the analysis, at least initially, (18 Competency / Grade combinations). If bias is confirmed using this data, it can be inferred that the bias, whether overgrading or undergrading will occur across all grades for the same competency.

It's possible that a single instructor may be biased in one or more competencies and may overgrade *and* undergrade in different competencies. Bias, particularly where combinations of overgrading and undergrading

exist may be due to problems classifying and assessing performance using the OBs. This should be addressed in any subsequent recalibration exercises.

Step 3: Peer Review normalisation to confirm bias

Each outlier relates to the percentage of pilots assessed at the grade being considered. For example in the box plot above the clear outlier for the APK indicates that the instructor assessed approximately 34% of pilots as Grade 4, compared to the instructor group median of about 73%.

Outliers are only an indication of *potential* bias; the bias needs to be confirmed or rejected.

In this example, those pilots assessed as Grade 4 for APK by this instructor need to be identified. The grading of those pilots by other instructors needs to be considered – what percentage of other instructors who assessed these same pilots awarded a Grade 4? The comparison of the two percentage figures gives an indication of whether the bias can be confirmed.

The bias may be overgrading or undergrading, identified by the grades awarded by the instructor comparison group. As a rule, the bias is more likely to be towards overgrading.

If the comparison group is large, the likelihood of their grading being accurate increases and therefore any differential between the suspected biased instructor and the comparison group becomes more significant. Whilst there are statistical methods for determining statistically significant difference, at this stage in the evolution of EBT, a training manager's judgement is acceptable.

Once the bias is confirmed, the key OBs identified by the biased instructor should also be compared to the instructor comparison group. Significant discrepancies should be noted and addressed in the next step, to confirm whether the biased instructor is using an effective method for classifying the observed behaviour as OBs.

Note that each pilot is only assessed during an EVAL once per module and therefore this process can only be achieved based on completion of more than one module; the second EVAL phase will be assessed by different instructors, allowing the peer group comparison to be made.

The process will never be 100% reliable for confirming bias and therefore other factors may need to be considered, such as the characteristics of the pilot population being considered - is there more likely to be variations in pilot performance due to low experience levels for example. Also, the second EVAL may be more difficult or easier. In addition, it has to be considered whether the bias may be due to racial, cultural, religious or gender factors.

As the ICAP matures, the identification of metrics to measure the success of this method will evolve.

Suggested operator metrics are:

- The number of outliers, (as a percentage of instructors), confirmed as biased. This should reduce with time for a stable instructor team;
- The number of instructors in the comparison group;
- % bias per competency, measured for each grade. (Captains and First Officer grades may be separated for this purpose);
- Trends year-on-year.

Step 4: Bias confirmed - Actions to be taken with biased instructors

The data supporting the analysis should not be shared with the instructor – there is a risk of bias in the opposite direction. Similarly, it's better if the instructor is not told that they are biased – better to observe their grading

behaviour directly and recalibrate if required. Operators should develop their own process for this recalibration but the following are offered as guidance:

1. The only reliable method to confirm that an instructor's grading is accurate and valid is to observe the individual conducting a real event.
2. It's recommended that a standardisation observation is conducted on an annual basis for every instructor. This pool of instructors who conduct these observations should be small and be members of the training standards team. These events afford an ideal opportunity to monitor instructors whose grading behaviour is suspected of being biased. Note that this is not a regulatory requirement but is recommended as best practice for the standardisation of any group of instructors or evaluators.
3. Grouping[2] of biased instructors for concordance training sessions, including the 'Controlled Event' concordance. Grading calibration may be specifically addressed during these sessions.
4. Consideration should be given to exploring the diversity of the pilot workforce when creating the video scenarios to be used for the 'Controlled Event' concordance training. The pilot actors should be a selection of different ages, genders, racial and religious backgrounds.

3.5.4.2 Instructor Concordance Assurance Method 2: the Controlled Event Method

AMC2 ORO.FC.146(c) specifies that concordance training must be conducted annually and GM2 ORO.FC.146(c) suggests that this training may include grading the same controlled content followed by: a subsequent comparison of intra-group variance and alignment of root-cause analyses between instructors.

AMC1 ORO.FC.231(a)(4) states that complex operators should include an ICAP-specific data analysis, demonstrating instructor-group assessment homogeneity (agreement), instructor assessment accuracy (alignment) and that concordance should be verified once every cycle. Also, the operator should establish procedures to address those instructors who do not meet the standards required.

This method addresses these requirements and includes elements of GM1 ORO.FC.231(a)(4): *Instructor assessment accuracy (alignment) may be inferred from comparing instructor assessments with an 'assessment standard' consisting of correctly identified competency(-ies) and correctly identified grade levels. Neither the competency(-ies) nor the grade level(s) may be communicated in advance to the instructors. The assessment standards may be set by consensus of a standards group, to guard against individual biases.*

This method addresses all these requirements and, together with Method 1, the Simulator Evaluation Data Method, covers the requirements for assuring instructor concordance.

Note that this method addresses both agreement and alignment.

The Controlled Event Method - process
The creation of a controlled event

Ideally, a controlled event is achieved using recording of a simulator session similar to an EVAL phase. The recording does not have to be the whole session, but may be a precis, in the order of 40 minutes long and should include elements of crew briefings so that the instructors may start to understand the character of the pilots.

[2] Grouping does not mean to group all bias instructor together for a training session because they may learn wrongly form each other and they may realised there is something wrong with their grading putting unnecessary pressure. A good practice is quite the opposite, to take one or two non-concordant instructors in a group of fully concordant instructors so the non-concordant instructor learns from the group.

The objectives when creating the controlled event should specify which OBs will feature, across a range of competencies – one or two per competency. (Feedback from previous concordance events and real pilot training data should be used to identify which OBs will be the focus).

The scenario may include minor technical failures but should avoid major system failures with limited decision-making options. It should be realistic and reasonably challenging. It should also involve participation of both pilots in roughly equal amounts to allow demonstration of sufficient number of behaviours to facilitate accurate grading.

This will also allow the scenario to be used for standardising Line Evaluators. The team who creates the scenario should not be involved in defining the assessment standards, i.e., the grades which should be assigned to the pilots.

The assessment standard, comprising grades for all relevant competencies, together with the key observable behaviours which underpin the grades should be created by a standards team, ideally numbering between five and ten experienced EBTIs.

The following points provide guidance on the management of the training:

Instructor grading of the pilots

- The operator should limit the number of instructors for each training event to 10-15;
- A method recording grades and OBs should be provided to ensure that Instructor grading results are anonymous. The success of the training relies partially on the freedom for instructors to share opinion on the performance of the pilots.
- Instructors should be briefed to assess all competencies, even if it is likely that one or more may not actually be observed.
- Instructors should identify key OBs which underpin the grade, either positively influencing the outcome or not present / not effective in terms of the outcome.
- Both pilots should be graded.
- The anonymous instructor grading should be summarised and shared with the group before the discussion exercise.
- This grading constitutes the concordance assurance.

Agreement discussion

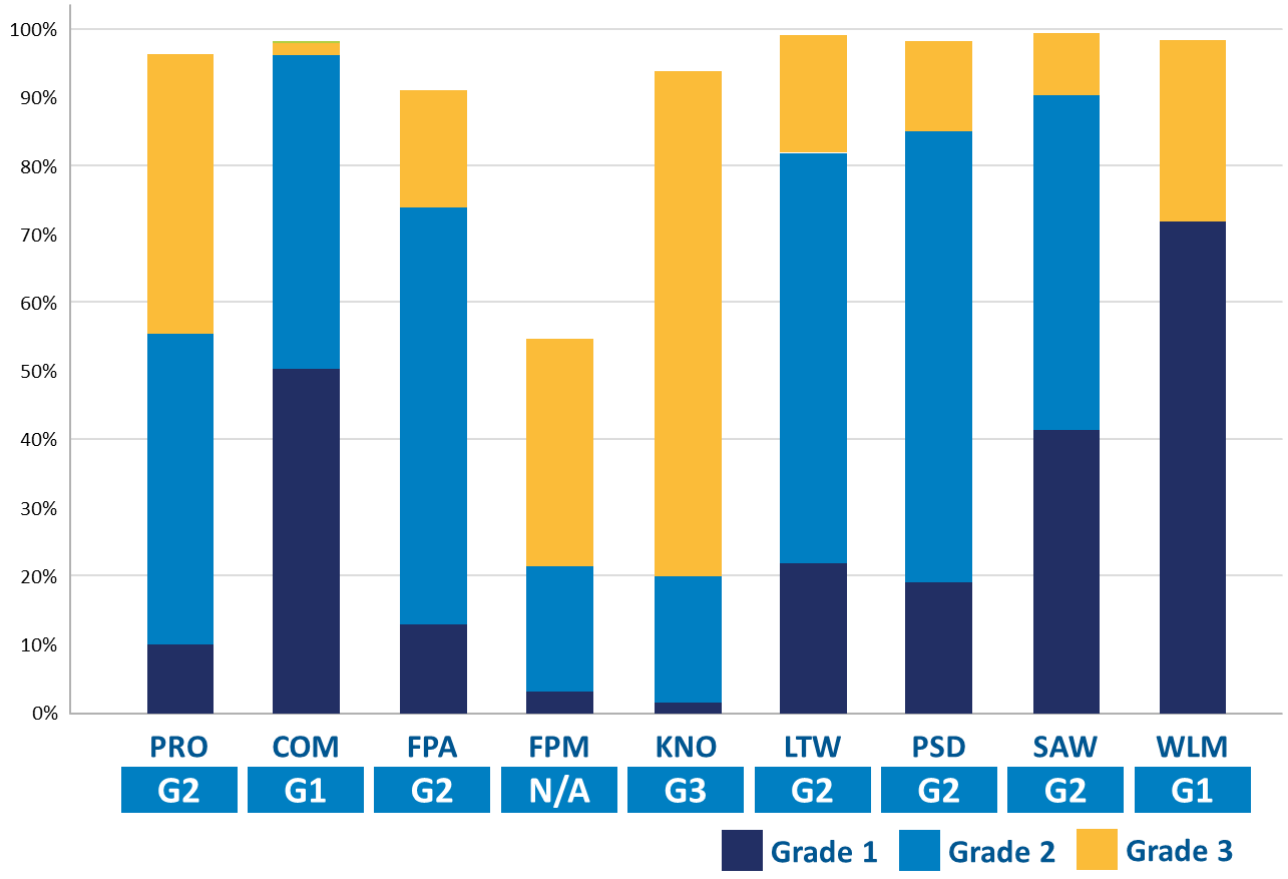
- The initial discussion should focus on agreement between the instructors rather than alignment with the operator's assessment standard:
- The discussion should be facilitative in nature and focus on identifying the root cause of pilot behaviour.
- Anonymity should be preserved by the tutor, maintaining a non-judgemental environment and improving the chance of agreement by encouraging open exchange of opinion between the instructors.
- Company definition of each grade should be used to inform the discussion and reflection on individual assessments.
- Other techniques which may assist instructors in enhancing their grading behaviour should be included, such as:
    - Instructor judgement is key – don't ignore your professional perceptions.
    - Don't talk yourself out of a grade by looking for OBs and forcing the grade to fit them.

- If struggling to identify which OBs are the root cause of underperformance, consider the question “what would you have to do to fix the problem?” This may help steer the grading behaviour.
- Don’t ignore performance as PM. The OBs for FPM and FPA competencies are still relevant.
- The grades are not negotiable during the debrief, but consider the questions which could be asked in order to lead to key OBs and root cause.

Alignment

- Once the agreement discussion is complete and hopefully consensus achieved, the operator assessment standard may be shared and discussed.
- The objective is to calibrate grading in line with operator expectations.
- The agreement discussion and the alignment process constitute concordance training and forms the basis of future improvements in concordance.

Figure 3-5 An example of the results of this concordance exercise over a series of training events



The grades along the base of the graphic represent the assessment standard or ‘expert consensus’.

Note that OBs are not included in this summary, and these also need to be compared with the assessment standard.

Examples of conclusions which may be drawn from the summary:

- Levels of agreement are indicated by the percentages stated. When looking at combinations of grades for each competency, e.g., 2 and 3 or 1 and 2, there is broad agreement, even though there isn't agreement between the individual grades.
- PRO – Only 46.4% agreed with Grade 2. 56.9% of instructors assessed Grade 1 or 2 which is aligned with the standard in terms of risk, i.e., either not competent or further training is likely to be required.
- COM – only 50.4% graded 1 which is probably below alignment KPIs but 97.3% graded 1 or 2. The risk associated with the performance was clearly identified.
- KNO – 73.9% of instructors were aligned with the grade 3 assessment standard, which would probably be acceptable.

It's apparent that the overall concordance cannot be judged by a single metric and must be viewed through a combination of metric lenses.

When judging the acceptance of the concordance assurance data, risk should be considered, i.e. what is the risk of a pilot returning to line operations when the overall grade should have been either "Not Competent" or Individual Tailored Training should have been triggered because of the number of Grade 2s.

Metrics

The following metrics are suggested for measuring concordance:

1. Alignment and agreement with expert consensus of overall result (Level 0 grading metrics: competent or not competent in EVAL/ completed or not completed in training);
2. Alignment with grading of at least one grade 2, i.e., identification of a potential risk with pilot performance.
3. Agreement levels in grading for each competency.
4. Alignment levels in grading for each competency.
5. Alignment and agreement for identification of key OBs (level 2 grading metrics) for grade 4 and 5, (positive) and grade 2s, (missing or ineffective). These indicate the likelihood of identifying root causes.

Key Performance Indicators (KPIs)

Operators should specify what is acceptable for each metric. If a metric KPI is not met, the reason must be established to the satisfaction of the operator and the competent authority.

The metrics above are broadly in order of priority and the most important is the Competent or Not Competent judgement. A target of 95% agreement and alignment would be appropriate.

4. Data Protection Recommendations

Organisations processing personal data must comply with data protection legislation guided by the six data protection principles.

4.1 The six data protection principles

In summary, the six data principles require that personal data is:

(a) processed lawfully, fairly and in a transparent manner in relation to individuals

You must ensure that you do not do anything with the data in breach of any other laws. Lawfulness also means that you don't do anything with the personal data which is unlawful in a more general sense. This includes statute and common law obligations, whether criminal or civil. If processing involves committing a criminal offence, it will obviously be unlawful. However, processing may also be unlawful if it results in:

- A breach of a duty of confidence.
- Your organisation exceeding its legal powers or exercising those powers improperly.
- An infringement of copyright.
- A breach of an enforceable contractual agreement.
- A breach of industry-specific legislation or regulations; or
- A breach of the Human Rights Act 1998.

You must use personal data in a way that is fair. This means you must not process the data in a way that is unduly detrimental, unexpected or misleading to the individuals concerned.

You must be clear, open and honest with people from the start about how you will use their personal data.

(b) collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes

- You must be clear about what your purposes for processing are from the start.
- This requirement aims to ensure that you are clear and open about your reasons for obtaining personal data, and that what you do with the data is in line with the reasonable expectations of the individuals concerned.
- You can only use the personal data for a new purpose if either this is compatible with your original purpose, you get consent, or you have a clear obligation or function set out in law. ('purpose limitation').

(c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed, where:

- **Adequate** – sufficient to properly fulfil your stated purpose.
- **Relevant** – has a rational link to that purpose, and

- **Limited** to what is necessary – you do not hold more than you need for that purpose ('data minimisation').

and:

- You should identify the minimum amount of personal data you need to fulfil your purpose and you should hold that much information, but no more.
- This is the first of three principles about data standards, along with accuracy and storage limitation.
- The accountability principle means that you need to be able to demonstrate that you have appropriate processes to ensure that you only collect and hold the personal data you need.

(d) accurate and, where necessary, kept up to date

- Every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay. You should take all reasonable steps to ensure the personal data you hold is not incorrect or misleading as to any matter of fact.
- You may need to keep the personal data updated, although this will depend on what you are using it for.
- If you discover that personal data is incorrect or misleading, you must take reasonable steps to correct or erase it as soon as possible.
- You must carefully consider any challenges to the accuracy of personal data. ('accuracy').

(e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed

- Personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes subject to implementation of the appropriate technical and organisational measures required by the GDPR in order to safeguard the rights and freedoms of individuals. You must not keep personal data for longer than you need it.
- You need to think about – and be able to justify – how long you keep personal data. This will depend on your purposes for holding the data. ('storage limitation').

(f) processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures ('integrity and confidentiality')

The security principle goes beyond the way you store or transmit information. Every aspect of your processing of personal data is covered, not just cybersecurity. This means the security measures you put in place should seek to ensure that:

- The data can be accessed, altered, disclosed or deleted only by those you have authorised to do so and that those people only act within the scope of the authority you give them.
- The data you hold is accurate and complete in relation to why you are processing it; and

- The data remains accessible and usable, i.e., if personal data is accidentally lost, altered or destroyed, you should be able to recover it and therefore prevent any damage or distress to the individuals concerned.

Information Security Risk Management

Expanding on the "Security Principle", organisations shall identify, assess, treat, and manage information security risk, to support the achievement of its planned objectives in alignment with their overall Risk Management Framework:

- The information security risk assessment process should consider business assets, their vulnerabilities, and the threats to those assets. These criteria should be assessed to capture the risk to the organisation and should form the basis of the information security risk assessment process.
- Risks should be monitored on an ongoing basis.

Data Sharing

In addition to considering whether data sharing achieves a benefit and is necessary, organisations must consider their overall compliance with data protection law when sharing data.

It is recommended that as a first step organisations carry out a Data Protection Impact Assessment (DPIA), even if they are not legally obliged to carry one out. Carrying out a DPIA is an example of best practice, allowing you to build in openness and transparency.

A DPIA will help organisations assess the risks in their planned data sharing and determine whether they need to introduce any safeguards. It will help them assess those considerations and document them. This will also help to provide reassurance to those whose data they plan to share.

It is good practice to have a data sharing agreement.

Data sharing agreements set out the purpose of the data sharing, cover what happens to the data at each stage, set standards and help all the parties involved in sharing to be clear about their roles and responsibilities.

Having a data sharing agreement in place helps you to demonstrate you are meeting your accountability obligations under the EU GDPR.

5. References

[1]. EASA EBT Manual V2.2

[2]. EASA OPS and FCL Regulations current on 14/07/2024

European Union Aviation Safety Agency

Konrad-Adenauer-Ufer 3
50668 Cologne
Germany

DATAPP (Digital Transformation - Case Studies for Aviation Safety Standards – Data Science Applications) | EASA (europa.eu)

Mail EASA.research@easa.europa.eu
Web www.easa.europa.eu

An Agency of the European Union