EUROPEAN UNION AGENCY FOR RAILWAYS

Moving Europe towards a sustainable and safe railway system without frontiers.

ACCOMPANYING REPORT TO THE OPINION OF THE EUROPEAN UNION AGENCY FOR RAILWAYS

for

the European Commission

regarding

ERA/OPI/2022-12: detailed methods for the common safety methods for assessing the safety level and the safety performance of railway operators at national and Union level

Contents

1. Introduction

This report accompanies Agency Opinion 2022-12 regarding detailed methods for the 'common safety methods for assessing the safety level and the safety performance of railway operators at national and Union level' (CSM ASLP). It explains the development process behind the detailed assessment methods and gives additional insights into the strengths and limitations of the respective methods.

The analyses were discussed with the Group of Analysts (GoA) Plenary Working Party on 5 May 2022. The final proposal for the detailed assessment methods was endorsed by the GoA Plenary Working Party on 19 October 2022. The process behind the development of the proposal and Opinion can be read in the Opinion.

Besides a few minor editorial changes, the analysis presented below has been fully drafted by GoA Subgroup C and should therefore be understood as a GoA output. This includes the following annexes:

- Annex 1: Detailed methods for the assessments of operators.
- Annex 2: Minor consequential changes of the recommendation ERA REC1219 concerning the assessment of safety levels.
- Annex 3: Minor consequential changes of the recommendation ERA REC1219 concerning the assessment of safety performance.

The Opinion largely builds on the GoA subgroup C analysis and output as presented here.

2. Analysis

2.1. Safety level assessment

2.1.1. Objective and background

The objectives for the detailed method for the safety level assessment were (a) to assess the extent to which a railway operator is reducing safety risks to fulfil the requirement of maintaining and continuously improving railway safety and (b) to identify railway operators with significantly higher or lower risks with the aim to support the definition of possible action plans, where needed.

Statistical inference shall be used to provide harmonised assessments and to reduce assessment errors to the greatest possible extent.

The assessment will provide for the first time a Europe-wide assessment of the safety levels of operators. It thereby complements the common safety method for assessment of achievement of safety targets (Commission Decision 2009/460/EC), which assesses safety levels of countries.

2.1.2. Scope

The assessment principles specify that the SL assessment consists of two separate tests. First, in accordance with the CSM ASLP Recommendation ERA1219 Annex III Part A 2.3.(a), an assessment whether safety levels have not started to deteriorate, improved or deteriorated and second, in accordance with Annex III Part A 2.3.(b), an assessment whether the safety levels are higher or lower than the level of similar railway operators.

The assessments shall be performed separately for the frequency of events and their severity. The SL assessment on frequency shall be performed on each Category A and B event type. The SL assessment on severity shall only be performed on Category A event types, as it focuses on the FWSI of accidents, and category B event types relate to incidents, not accidents.

The test shall be performed separately for each type of operation that is performed by the railway operator. The distinction can be made because the operator reports on the type of operation that was performed when an event occurred and reports the total volumes by type of operation.

As shown in Table 1, as an example, for an operator that conducts three types of operations there will be 12 distinct 'groups' for which assessments are performed. Moreover, there are about 26 Category A event types and approximately 75 Category B event types. This means that for this single operator about 750 individual assessments shall be performed.

Operator XYZ (example of operator with three types of operations)	SL assessment in accordance with Annex III Part A 2.3.(a)		SL assessment in accordance with Annex III Part A 2.3.(b)	
Type of operations	*Frequency (Event A & B)*	*Severity (Event A)*	*Frequency (Event A & B)*	*Severity (Event A)*
RU-1 : Operating passenger trains	x	X	x	x
RU-2 : Operating high-speed trains	x	X	x	x
RU-5 : Operating terminals	x	X	x	x

Table 1: Example of SL assessments for one operator

An assessment takes place every 3 months, where the assessed period is the latest available one-year period.

The reference period for the SL assessment in accordance with Annex III Part A 2.3.(a) is the three-year period prior to the assessed period. The benefit of the floating period is that, generally, fewer events occur, so that the reference period becomes increasingly 'safer'. This implicitly promotes continuous improvement. The downside of a floating reference period is that slow-paced deteriorations are not as easily spotted. The CSM ASLP GoA shall reflect on whether and how a dynamic reference period can be introduced.
The reference period for the SL assessment in accordance with Annex III Part A 2.3.(b) is the same as the assessed period. The operational volumes for the reference period are derived from all operators that perform the type of operation under assessment.

2.1.3. Assessment methods

Safety level assessments are common for every mode of transport. In the railway sector many public administrations structurally assess the status and evolution of railway safety. A review of methods applied in the railway sector has been performed to understand their respective strengths and weaknesses. A summary is provided in Figure 1.

SAFETY LEVEL ASSESSMENT METHODS		Weaknesses					Strengths			
Methods	**Key reference**	**Small count bias**	**Fixed baseline**	**Unclear definitions & taxonomy concerns**	**Considers severity**	**Impact of probability distribution on assessment**	**Enables fair comparison of operators (i.e., normalization possible)**	**Possibility to aggregate operator results on national and EU level**	**Quantification of the degree of change**	**Utility (in terms of feasibility & usefulness)**
CSM CST	2009/460/EC									
Bayesian	NSA CH									
Compound Poisson	Siemens/TUV									
Cox/Poisson	Evans / Siemens/TUV									
Rate-ratio test	Siemens/TUV									
Negative binomial regression	Liu									

Not assessed
Negatively assessed
Neutrally assessed
Positively assessed

Figure 1: Review of safety level assessment methods in the railway sector

Subgroup C members reflected on the findings and concluded that Bayesian inference and the rate-ratio test were the most adequate candidates for further development. After subsequent tests and discussions, Bayesian inference was selected for the following reasons:

- It provides a more intuitive way of presenting results, namely in terms of probability, rather than by interpreting p-values.
- It allows for the quantification of the degree of the change under different probability levels.
- NSA Switzerland (NSA CH) gained considerable experience with Bayesian inference for safety assessments. Its usability and insightfulness were positively evaluated.

Based on these findings, two distinct implementations of Bayesian inference were further developed for the assessment of the frequency of occurences and for the assessment of the severity of occurences.

A) Assessment of the frequency of occurences

The method is detailed in Annex 1 to this reports. This section provides additional information on how to understand the formulas and interpret the results.

In essence, the method evaluates the probability of an event to occur in the assessed period, while considering the frequency of such events occuring in the reference period. For that, the method considers the respective size of operations in both periods, the prior probability of an event to occur, and the posterior probability. Using the proposed Bayesian approach, it is possible for each probability level to determine the degree with which a change occurred.

The probability levels on what constitutes sufficient evidence for a change can vary depending on the context. Building on Jeffreys (1961) and the experiences from NSA CH the probabilities in Table 2 are proposed to perform the assessments.

SL Class	Probabilities	SL assessment in accordance with Annex III Part A 2.3.(a)	SL assessment in accordance with Annex III Part A 2.3.(b)
1	90% - 100%	Strong evidence for deterioration	Strong evidence for a lower level
2	75% - 90%	Moderate evidence for deterioration	Moderate evidence for a lower level
3	25% - 75%	No evidence for improvement or deterioration	No evidence for a lower or higher level
4	10% - 25%	Moderate evidence for improvement	Moderate evidence for a higher level
5	0% - 10%	Strong evidence for improvement	Strong evidence for a higher level

Table 2: SL Assessment framework

After performing the analysis, an operator can visualise the results as depicted in a few examples below. The y-axis indicates the probability levels, the x-axis the factor with which a change occurred. The black S-curve provides results for different probabilities. The red dot is fixed on factor 1, which indicates for what SL class evidence is found.

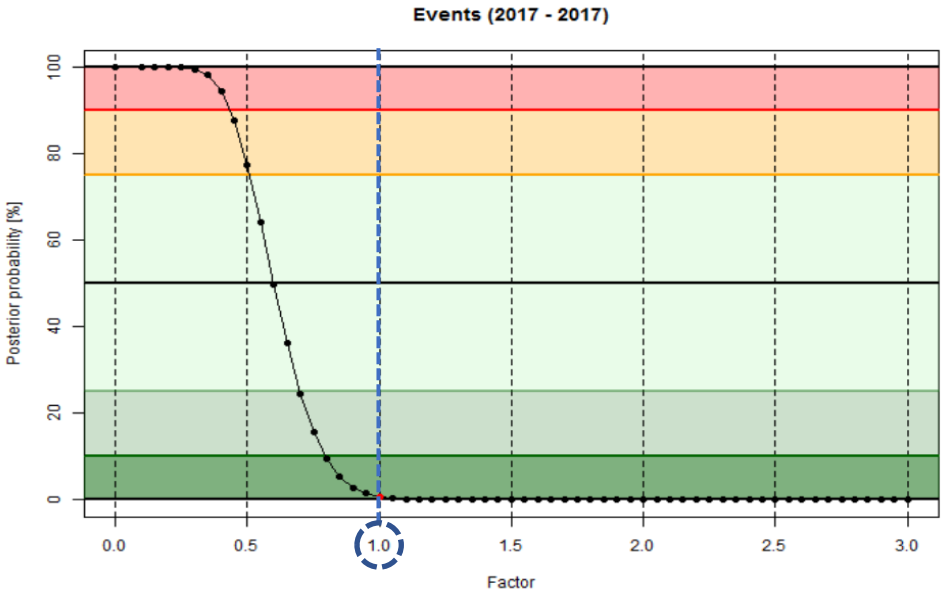


Figure 2: Visualisation of the outcomes of the SL estimation – example 1

The interpretation of the results in Figure 2 is that, as the red point lays in the dark green area (with a probability of almost 0%), there is strong evidence for improvement. In addition, the black s-curve crosses the border between dark green area and green area at approximately 0.80, implying that there is strong evidence for an improvement of the safety level by at least 20%[1].

In Figure 5 the red point is in the light green area, meaning that there is insufficient evidence to prove any deterioration or improvement.

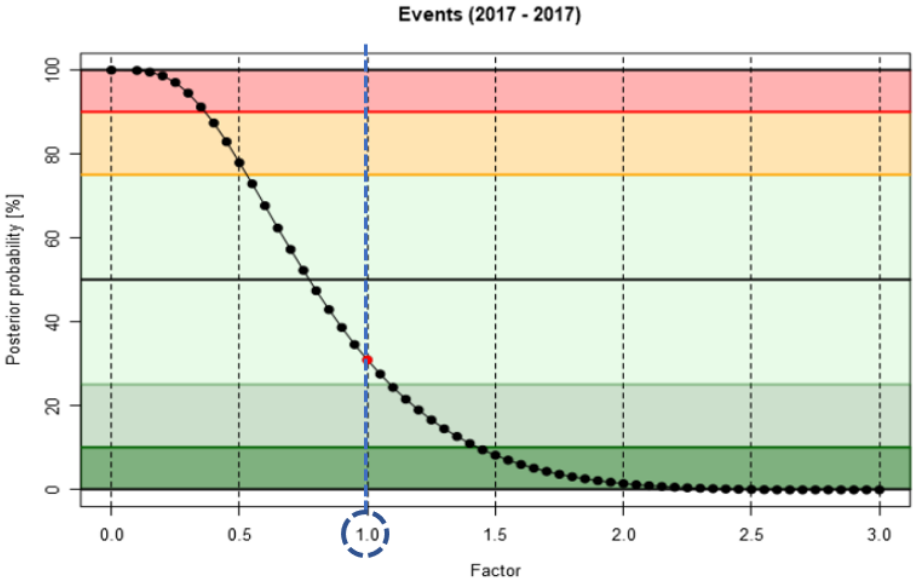


Figure 3: Visualisation of the outcomes of the SL estimation – example 2

In Figure 4 the red point is in the dark red area, meaning that there is strong evidence for a deterioration. As the curve crosses the red line at factor 1.06, there is strong evidence that the safety levels deteriorated by at least 6%.

[1] *Andrášik, R. (2020), 'Evaluation of safety level in public transport based on Bayesian inference', NSA CH*

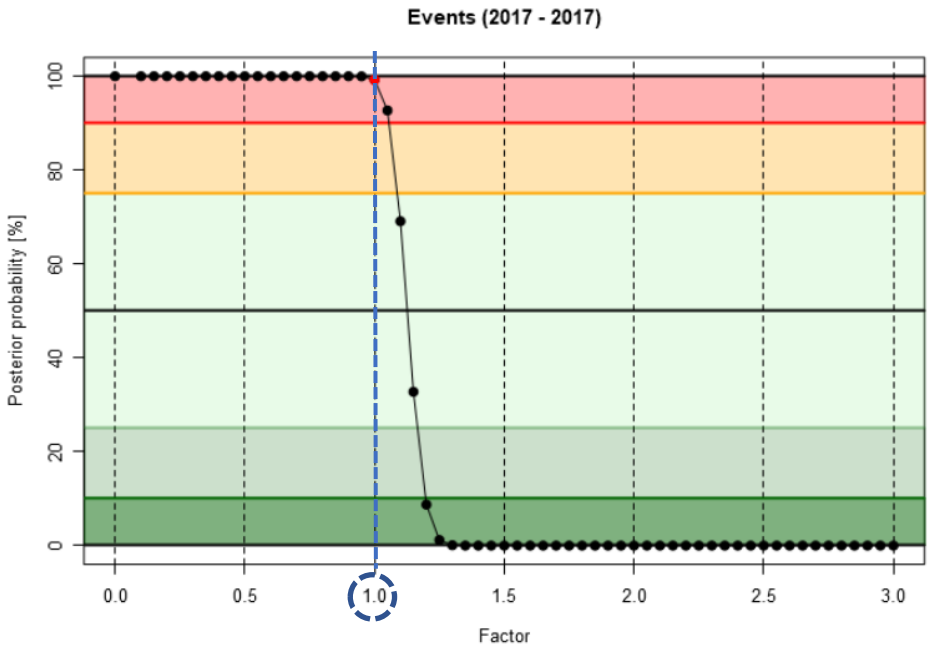


Figure 4: Visualisation of the outcomes of the SL estimation – example 3

The assessment also allows for the temporal analysis on how safety levels evolve, showing the direction and the degree of the change. An example is provided in Figure 5.

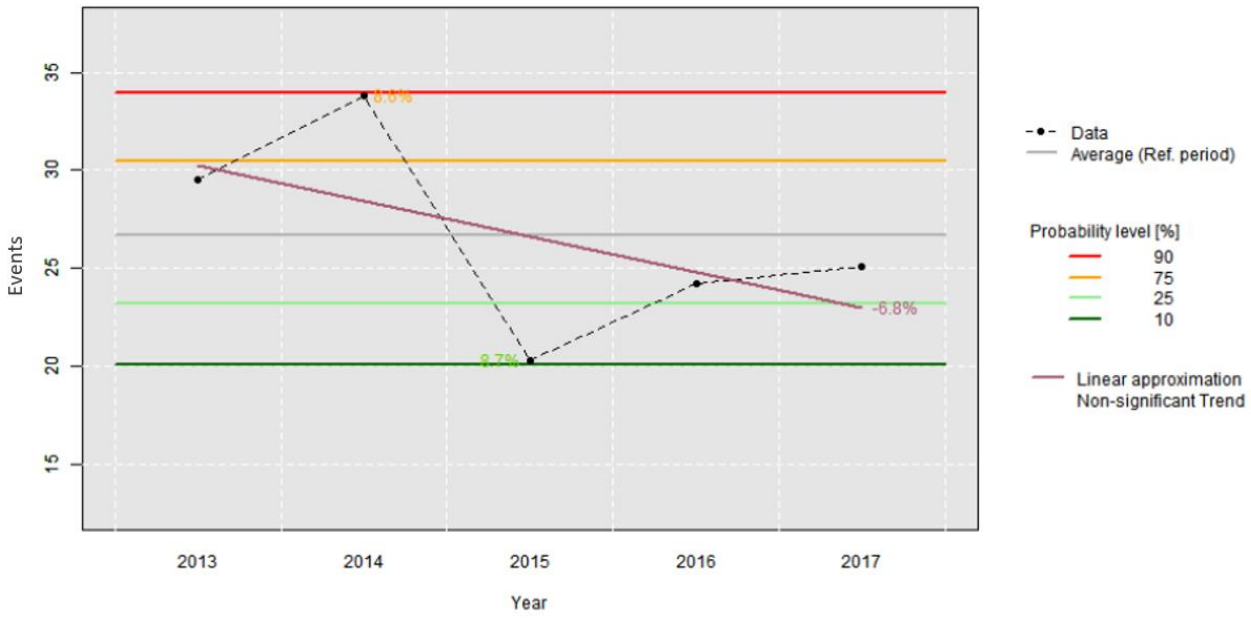


Figure 5: Temporal visualisation of the results of an SL assessment

The combination of showing both the direction and the degree of the change is believed to be a key strength of the proposed method.

B) Assessment of the severity of events

Beyond the assessment of the frequency of events, GoA Subgroup C proposes that an analysis is performed on the severity of accidents in terms of FWSI, allowing a more comprehensive view on the evolution of safety on the European rail network.

The analytical challenge is that events with severe injuries or fatalities are rather rare, leading to limited data that can be used to determine whether a true change occurred. Moreover, unlike frequency data, severity data is typically distribution-free, which requires a specific non-parametric test to be applied.

As indicated before, GoA Subgroup C prefers Bayesian tests for reasons of understandability and comprehensiveness. For non-parametric distributions it was proposed to apply the Bayesian version of a Wilcoxon test. Simulations on accident data by NSA CH confirmed the suitability of the test. Yet other non-parametric tests had a higher statistical power when applied to low sample sizes.

A review took place to understand the comparative power of the test. The reviewed tests include the Kolmogorov-Smirnov Test, Kuiper test, Cramer-Von Mises, Anderson-Darling, frequentist Wilcoxon test, Wasserstein test, and the DTS test. The DTS test had the highest power and would therefore be most apt to identify whether a real change in the distribution of the severity of accidents occurred. More information on this test can be retrieved from Dowd (2020)[2].

Based on this understanding, GoA Subgroup C opts for a two-stepped approach. First, the DTS test shall be applied to identify whether the assessed distribution differs significantly from the reference period distribution. If not, no further test shall be performed, and the assessment is that there is no evidence for a change (i.e. SL class 2 according to Table 2). However, if a change is detected, a Bayesian Wilcoxon Test shall be performed to determine the direction and degree of the change. For the assessment of the results the same probabilities are used as mentioned in Table 2.

Based on the results, similar visualisations as to Figure 2 and Figure 5 can be created.

C) Safety level score (SLS)

The assessment results of the methods under section A) and B) provide a very detailed insight into the trend and comparative safety level of each operator on a granular level, namely the level of an event type. To further support the identification of safety-related improvement needs and opportunities, and to facilitate the aggregation of results, a so-called safety level score (SLS) is proposed.

A safety level score shall be set for each event type and for both the SL assessment in accordance with CSM ASLP Recommendation ERA1219 Annex III Part A 2.3.(a) and the SL assessment in accordance with Annex III Part A 2.3.(b). The SLS considers the degree of the change for each SL class (i.e. moderate or strong evidence for any change), and is weighted by the volume of operations and the average severity of an accident. As severity is considered, a SLS can only be determined for Category A event types. In case a railway operator performs multiple types of operations, as defined by Annex IV Part C, safety level scores shall be determined per type of operation.

As said this allows for the comparison and aggregation of scores, leading to overview tables such as shown below.

A similar exercise shall be performed for Category B event types. While such event types do not necessarily have a severity, some Category B events result into a Category A event. As such, it is possible to derive an average severity for Category B events, which will likely be substantially lower than the average severity of Category A events.

[2] Dowd, C. (2020), 'A New ECDF Two-Sample Test Statistic', Cornell University, https://arxiv.org/abs/2007.01360

	Whether a safety level estimated for a railway operator has not started to deteriorate, has improved or has deteriorated			Whether a safety level estimated for a railway operator is higher or lower than the level of similar railway operators		
TYPE OF OPERATION X	Safety Level	Assessment of the Safety Level		Safety Level Difference	Assessment of safety level difference	
	[Risk/Mio. Tkm]	Assessment result	SL Score	[Risk/Mio. Tkm]	Assessment result	SL Score
Group	0.855	Strong Evidence of Improvement	-0.202	not applicable	not applicable	not applicable
Operator 1	0.87	Strong Evidence of Deterioration	10.494	0.015	No Evidence	0
Operator 2	0.28	Strong Evidence of Deterioration	5.3	-0.575	Strong Evidence of higher SL	-7.1
Operator 3	0.524	Strong Evidence of Deterioration	2.5	-0.331	Moderate Evidence of higher SL	-7.6
Operator 4	1.203	Moderate Evidence of Deterioration	23.6	0.348	No Evidence	0
Operator 5	0.105	Moderate Evidence of Deterioration	15.2	-0.75	Strong Evidence of higher SL	-12.3
Operator 6	2.8	No Evidence	0	1.945	Strong Evidence of lower SL	25.7
Operator 7	0.678	No Evidence	0	-0.177	Moderate Evidence of higher SL	-5.2
Operator 8	1.654	Moderate Evidence of Improvement	-1.7	0.799	Moderate Evidence of lower SL	9.6
Operator 9	0.013	Moderate Evidence of Improvement	-21.6	-0.842	Strong Evidence of higher SL	-12.4
Operator 10	2.703	Strong Evidence of Improvement	-6.7	1.848	Strong Evidence of lower SL	14.3

Figure 6: Example of an SLS score overview, comparing several operators

While the SLS goes beyond the assessment itself, it constitutes an important tool for the GoA to identify risk areas and to prioritise its proposals for railway safety improvements.

D) Aggregation of operator's safety level at national and Union levels

Initially, the CSM ASLP Recommendation opted that the safety level aggregated at national and Union level *'shall be estimated with averages of individual railway operators' safety levels weighted by their respective volume of operation'*.

GoA Subgroup C proposes that no ex-ante aggregation of operator safety levels occurs, but that the safety levels for countries and the Union are calculated the same way as for operators. By doing so, no distortion takes place due to the event allocation rules, and it provides for the greatest level of analysis on the national and Union levels. The tables below show the types of assessments that are to be performed.

Country ZYX	*SL assessment in accordance with Annex III Part A 2.3.(a)*		*SL assessment in accordance with Annex III Part A 2.3.(b)*		*SL Score in accordance with Annex III Part A 2.3.(a)*		*SL Score in accordance with Annex III Part A 2.3.(b)*	
Type of operations	*Frequency (Cat A & B)*	*Severity (Cat A)*	*Frequency (Cat A & B)*	*Severity (Cat A)*	*Frequency (Cat A & B)*	*Severity (Cat A)*	*Frequency (Cat A & B)*	*Severity (Cat A)*
IM-1 : Operating railway lines	x	x	x	x	x	x	x	x
IM-2 : Operating terminals	x	x	x	x	x	x	x	x
RU-1 : Operating passenger trains	x	x	x	x	x	x	x	x
RU-2 : Operating high-speed trains	x	x	x	x	x	x	x	x
RU-3 : Operating freight trains	x	x	x	x	x	x	x	x
RU-4 : Operating dang. goods trains	x	x	x	x	x	x	x	x
RU-5 : Operating terminals	x	x	x	x	x	x	x	x

Figure 7: Assessments on the country level

EU	*SL assessment in accordance with Annex III Part A 2.3.(a)*		*SL assessment in accordance with Annex III Part A 2.3.(b)*		*SL Score in accordance with Annex III Part A 2.3.(a)*		*SL Score in accordance with Annex III Part A 2.3.(b)*	
Type of operations	*Frequency (Cat A & B)*	*Severity (Cat A)*	*Frequency (Cat A & B)*	*Severity (Cat A)*	*Frequency (Cat A & B)*	*Severity (Cat A)*	*Frequency (Cat A & B)*	*Severity (Cat A)*
IM-1 : Operating railway lines	x	x			x	x		
IM-2 : Operating terminals	x	x			x	x		
RU-1 : Operating passenger trains	x	x			x	x		
RU-2 : Operating high-speed trains	x	x			x	x		
RU-3 : Operating freight trains	x	x			x	x		
RU-4 : Operating dang. goods trains	x	x			x	x		
RU-5 : Operating terminals	x	x			x	x		

Figure 8: Assessments on the Union level

By performing these assessments, similar visualisations as to Figure 2 and Figure 5 can be created on country and Union levels. A comparison of countries could moreover be done in a similar fashion as shown in Figure 6.

2.2. Safety performance assessment

2.2.1. Objective and background

The method is detailed in Annex 1 of this report. The sections below provide additional information on how to understand the method and why they were developed as such.

The general objective is to assess, based on the self-estimations provided by each railway operator, the extent to which a railway operator fulfils the requirement of maintaining and continuously improve railway safety in the domain of risk control measures.

The self-assessment consists of an evidence-based maturity level evaluation on four dimensions:

- 'Planning' risk control measures (Area P)
- 'Setting up and operating' of risk control measures (Area D)
- 'Monitoring' of risk control measures (Area C)
- 'Reviewing and adjusting' of risk control measures (Area A)

On each dimension a maturity level from 1 to 5 should be provided in accordance with CSM ASLP Appendix B. The self-evaluation is an annual exercise.

Assessments

The CSM ASLP Recommendation requests the development of three assessments on whether the operator's SP:

a) Is stable. Reference period = Year n-1
b) is better or worse than the level of similar railway operators. Reference period = Year n
c) has improved or deteriorated compared to the past. Reference period = Previous 5 years since certification or authorization

For each assessed objective and assessment period referred to in section 3.1 the Agency shall determine the situation applicable to the assessed operator, allowing the following categorisation:

SP class	SP assessment in accordance with section 3.1 a) and c)	SP assessment in accordance with section 3.1 b)
1	Strong evidence for deterioration	Strong evidence for lower performance
2	Moderate evidence for deterioration	Moderate evidence for lower performance
3	No evidence for improvement or deterioration	No evidence for lower or higher performance
4	Moderate evidence for improvement	Moderate evidence for higher performance
5	Strong evidence for improvement	Strong evidence for higher performance

Assumptions

The safety performance self-estimation framework implies several limitations to how safety performance can be assessed. This section specifies the key assumptions that shape the subsequent methodological choices.

General

- Experiences with SL assessment exist. SP assessment is novel. There is no similar assessment in place across Europe or, to our knowledge, the world. Neither in rail nor other modes of transport. No lessons can therefore be drawn from prior experiences.

Levels

- The dimensions measure distinct activities but are interdependent.
- The maturity levels are ordinal values with unequal distance. Hence, a drop from level 5 to 3 differs from a drop from 3 to 1
- It is expected that the results of the self-evaluation for most operators remain stable over time. Hence, a drop or increase by 1 level is more likely than changes by 2 or more levels.

Grouping

- Creating a composite SP indicator is troubled by the different weights associated with each dimension amongst stakeholder groups and across geographies.
- An aggregation of levels may lead to a loss of information. Internal consistency and other analyses on actual SP data would clarify whether a composite indicator is meaningful.
- Operators can be grouped along the lines of the categories in CSM ASLP Appendix D – Part B.
- Group results are probably non-normally distributed.

From the assumptions above the following can be derived:

- The focus of the assessment will lay on the direction of the change, not its magnitude.
- The possibilities to apply inferential tests are limited.
- Only after self-reports come in, analyses on interdependencies, distributions and patterns can be conducted. These insights can be used for a review of the SP assessment method in due time.

2.2.2. Assessment methods

Each assessment will be done through a two-step approach.

Step 1 – Assessment by dimension

The direction of the change will be assessed for each dimension. The nature of the assessment is different for each test.

Area result	Objective assessed		
	a)	b)	c)
Lower	Maturity level in assessed period < maturity level in reference period	50% of operators in the group have a higher maturity level than the operator in the assessed period	Maturity level in assessed period < maturity level mode in reference period
Stable/Average	Maturity level in assessed period = maturity level in reference period	The result of the assessment is neither lower nor higher	Maturity level in assessed period = maturity level mode in reference period
Higher	Maturity level in assessed period > maturity level in reference period	50% of operators in the group have a lower maturity level than the operator in the assessed period	Maturity level in assessed period > maturity level mode in reference period

The following additional rules apply when applying the framework:

- For objective b):
  An operator belongs to each group in accordance with Appendix D - Part B for which the operator indicates that operations were performed in the assessed period. The assessment shall be applied separately for each group to which the operator belongs.
- For objective c):
  If there is no single maturity level mode in the reference period, the mode will be determined by including the maturity level in the assessed period. If there is still no mode, the highest most frequently reported maturity level will be selected.

Step 2 – Overall assessment

After the four dimensions are assessed, the combined results are evaluated to assess the final safety performance assessment category, using the framework below.

Assessment of area results	Number of areas in which the result was obtained			
	1 area	2 areas	3 areas	4 areas
Lower	Class 3	Class 2	Class 1	Class 1
Stable/Average	Class 3	Class 3	Class 3	Class 3
Higher	Class 3	Class 4	Class 5	Class 5

The following additional rules apply:

› To prevent the flagging of false positives, a change on one dimension is considered to be insufficient evidence of a change in safety performance.
› If SP Class 1, 2, 4, or 5 is noted, the safety performance assessment will never be Class 3.
› If Class 2 and Class 4 are both noted, the safety performance assessment shall be both categories.

Examples

SP assessment a) year-by-year

	P	D	C	A
Year n	2	5	5	3
Year n-1	3	5	2	3
Dimension Assessment	Lower	Stable	Higher	Stable
Overall assessment	No evidence for improvement or deterioration (Class 3)			

SP assessment b) group

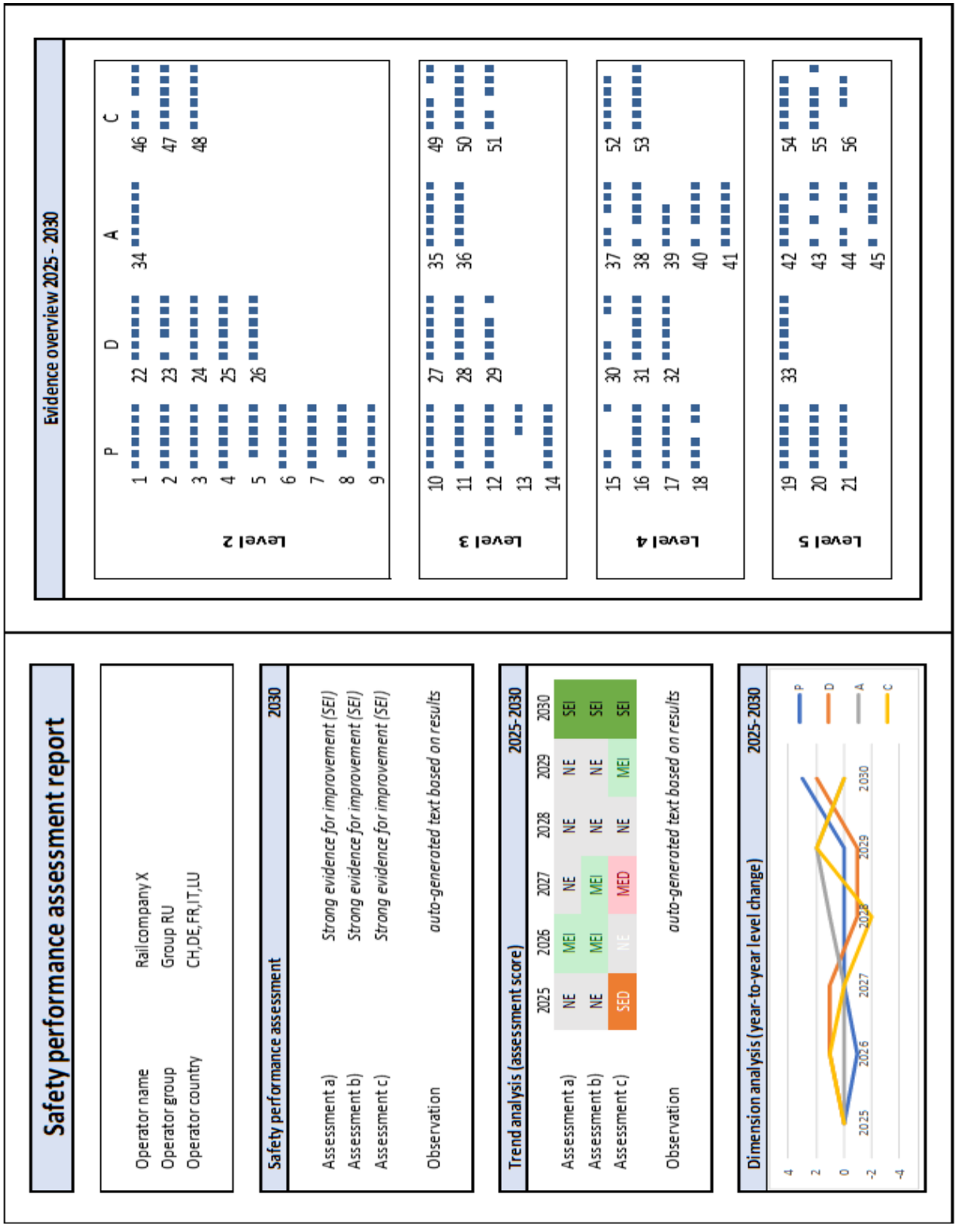
Dimension		P	D	C	A
OPERATOR: Maturity level in year n		2	2	3	5
Group: Distribution of levels Operator level highlighted in blue	1	40%	10%	10%	0%
	2	**50%**	**80%**	80%	0%
	3	10%	10%	**10%**	10%
	4	0%	0%	0%	10%
	5	0%	0%	0%	**80%**
Group: Cumulative proportions of levels Operator level highlighted in blue	1	40%	10%	10%	0%
	2	**90%**	**90%**	90%	0%
	3	100%	100%	**100%**	10%
	4	100%	100%	100%	20%
	5	100%	100%	100%	**100%**
Comparison: operators with lower level		40%	10%	90%	20%
Comparison: operators with same level		50%	80%	10%	80%
Comparison: operators with higher level		10%	10%	0%	0%
Dimension assessment		average	average	stronger	average
Overall assessment		No evidence for lower or higher performance (Class 3)			

SP assessment c) trend

	P	D	C	A
Year n	2	5	5	4
Year n-1	3	5	2	3
Year n-2	3	5	2	3
Year n-3	2	2	1	5
Year n-4	4	5	1	1
Year n-5	4	5	1	3
Mode	4	5	1	3
Dimension assessment	Lower	Stable	Higher	Higher
Overall assessment	Moderate evidence for improvement (Class 4)			

Implementation of the method

GoA Subgroup C asserts that providing solely the assessment results would not give the picture needed to properly interpret the safety performance of an operator. Instead, the ISS should ensure that the results are presented so that interlinkages between the assessments and evidence are clarified, and trends are highlighted. An example of such an overview is provided below.



3. Conclusions

The Agency wants to thank the GoA Subgroup C members and all those that contributed to the development of the detailed SL and SP assessment methods.

The Agency shall build on the GoA Subgroup C proposal and issue its Opinion on the detailed assessment methods.

ANNEX 1 - Detailed methods for the assessment of operators

Annex III – PART C

DETAILED METHODS FOR THE ASSESSMENTS OF OPERATORS

PART C1 - DETAILED METHODS FOR THE SAFETY LEVEL ASSESSMENT OF OPERATORS

1. Methodology for the safety level assessment of operators

1.1 Scope

The safety level assessment consists of an assessment on the frequency of events and an assessment on the severity of accidents.

For each railway operator the assessment on the frequency of events, as detailed in section 1.2, shall be applied to each Category A and Category B event type.

For each railway operator the assessment on the severity of accidents, as detailed in section 1.3, shall be applied to each Category A event type.

A separate assessment shall be performed for each type of operation, as defined by Annex IV – Part C, on which the railway operator reported.

For each three-month calendar period and each completed year an assessment of the achieved safety level shall be performed. The assessed period is the latest available one-year period.

The reference period for the SL assessment in accordance with Annex III Part A 2.3.(a) is the three-year period prior to the assessed period.

The reference period for the SL assessment in accordance with Annex III Part A 2.3.(b) is the same as the assessed period. The operational volumes for the reference period are derived from all operators that perform the type of operation under assessment.

1.2 Assessment on the frequency of events

The assessment on the frequency of events is performed using Bayesian Inference.

1.2.1 Safety level estimation

The total number of events is denoted as n_a for the assessed period and n_r for the reference period.

The shape parameter values are set as $\alpha = 1, \beta = -(\log_2(1-\eta))^{-1}$, where the exposure η is determined by formula

$$\eta = \frac{\text{operational volumes in the assessed period}}{\text{operational volumes in the assessed and reference periods}} \quad (1)$$

The prior f(p) is set by a gamma distribution

$$f(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \quad (2)$$

The posterior probability density function is

$$f(p|n_a, n_r) = \frac{P(n_a, n_r|p) f(p)}{Z} \quad (3)$$

With likelihood function

$$P(n_a, n_r|p) = \binom{n_a + n_r}{n_a} p^{n_a}(1-p)^{n_r} \quad (4)$$

And normalizing constant

$$Z = \int_0^1 P(n_a, n_r|p) f(p)\, dp \quad (5)$$

The posterior probability density function is used to calculate the following probability

$$P(p > \eta|n_a, n_r) = \int_n^1 f(p|n_a, n_r)\, dp \quad (6)$$

Finally, the degree of the change, called the minimal change of frequency (MCF), can be determined by the formula

$$P(p > k\eta|n_a, n_r) = \int_{kn}^1 f(p|n_a, n_r)\, dp \quad (7)$$

which is valid for any factor k ∈ (0, 1⁄η).

1.2.2 Safety level assessment

The estimation shall be assessed using the following probabilities and associated assessments.

SL Class	Probabilities	SL assessment in accordance with Annex III Part A 2.3.(a)	SL assessment in accordance with Annex III Part A 2.3.(b)
1	90% - 100%	Strong evidence for deterioration	Strong evidence for a lower level
2	75% - 90%	Moderate evidence for deterioration	Moderate evidence for a lower level
3	25% - 75%	No evidence for improvement or deterioration	No evidence for a lower or higher level
4	10% - 25%	Moderate evidence for improvement	Moderate evidence for a higher level
5	0% - 10%	Strong evidence for improvement	Strong evidence for a higher level

The MCF shall be provided for each assessment where an improvement or deterioration is noted. In case there is strong evidence for an SL class, the MCF shall also be given for the moderate evidence SL classes.

1.3 Assessment on the severity of accidents

The estimation on the severity of accidents is performed in two steps.

First, for each railway operator a DTS test shall be performed to determine whether there is a significant difference between the severity of the accidents in the reference and assessed period.

Second, for those cases where a significant difference was detected, a Bayesian test shall be applied to classify the direction and degree of the change.

1.3.1 Safety level estimation

The severity distributions are denoted as F_a for the assessed period, F_r for the reference period, and F_c for the combined distribution of F_a and F_r. The symbol 'n' stands for the total number of events for F_c.

Step 1

For each event type a DTS test shall be performed on the FWSI count.

$$DTS = \int_{x \in R} \left(\frac{|F_a(x) - F_r(x)|}{\sqrt{2F_c(x)\left(1 - F_c(x)\right)/n}} \right)^p \quad (8)$$

The next step shall be taken if the p-value of the DTS test is lower than 0.05. If not, no further test shall be performed, and it is concluded that no evidence is found for a change in severity of accidents for the assessed event type (i.e. SL class 3).

Step 2

A Bayesian version of the two-sample Wilcoxon signed-rank test for paired data shall be conducted.

First, the combined dataset from the assessed and reference periods shall be sorted and assigned ranks R_1, … , R_n. Then, transform the ranks to quantiles (inverse-normal rank transformation) as

$$1 \to \Phi^{-1}\left(\frac{1}{2n}\right), 2 \to \Phi^{-1}\left(\frac{3}{2n}\right), \ldots, n \to \Phi^{-1}\left(\frac{2n-1}{2n}\right) \quad (9)$$

The prior is set by

$$\mu_a, \mu_r \sim Uniform(\mu_{min}, \mu_{max})$$
$$\sigma_a, \sigma_r \sim Uniform(0, \sigma_{max})$$

where the hyperparameters are defined as

$$\mu_{min} = \frac{1}{2}\left(\Phi^{-1}\left(\frac{1}{2n}\right) + \Phi^{-1}\left(\frac{3}{2n}\right)\right) \quad (10)$$

$$\mu_{max} = -\mu_{min}$$
$$\sigma_{max}(n) = -\Phi^{-1}\left(\frac{1}{2n}\right) \quad (11)$$

The posterior probability density function is proportional to the likelihood and the prior

$$f(\mu_a, \mu_r, \sigma_a, \sigma_r \mid n_a, n_r) \propto P(n_a|\mu_a, \sigma_a)P(n_r|\mu_r, \sigma_r)f(\mu_a)f(\mu_r)f(\sigma_a)f(\sigma_r) \quad (12)$$

where n_a and n_r stand for the inverse-normal rank transformed data from assessed and reference periods, respectively.

Gibbs sampling is used to sample from the posterior distribution and make inference about parameters $\mu_a, \mu_r, \sigma_a, \sigma_r$ as follows:

1. Set initial values $\mu_a = 0, \mu_r = 0, \sigma_a = 1, \sigma_r = 1$. Set $k = 1$.
2. Let $j = k \bmod 4$. To obtain the next sample, update the j-th parameter while keeping the others unchanged. Randomly draw the parameter from either $Uniform(\mu_{min}, \mu_{max})$ in the case of μ_a, μ_r, or $Uniform(0, \sigma_{max})$ in the case of σ_a, σ_r.
3. Set $k = k + 1$.
4. Repeat steps 2 and 3 until sufficient samples were drawn.

Subsequently, the probability of interest is calculated by

$$P(\mu_a > \mu_r|n_a, n_r) = \int_{\mu_a > \mu_r} f(\mu_a, \mu_r, \sigma_a, \sigma_r \mid n_a, n_r)\, \mathrm{d}\theta \quad (13)$$

where θ is a shortcut for all the four parameters $\mu_a, \mu_r, \sigma_a, \sigma_r$.

Finally, the extent of a shift in the medians, called the minimal change of severity (MCS), can be evaluated through the probability

$$P\left(\mu_a > \mu_r^{shift} \mid n_a, n_r^{shift}\right) = \int_{\mu_a > \mu_r^{shift}} f\left(\mu_a, \mu_r^{shift}, \sigma_a, \sigma_r^{shift} \mid n_a, n_r^{shift}\right) \mathrm{d}\theta \quad (14)$$

where the superscript *shift* means *after shifting the original data*. The above probability can be calculated for any value of the *shift* ranging from -max{ FWSI in the reference period } to max{ FWSI in the assessed period}.

1.3.2 Safety level assessment

The table under section 1.2.2 shall also be used to assess the estimation under section 1.3.1.

2. Safety level score

For each event type a safety level score shall be determined to support the identification of safety-related improvement needs and opportunities. The safety level score equally facilitates the comparison of safety levels between railway operators and countries.

A safety level score shall be determined for the SL assessment in accordance with Annex III Part A 2.3.(a), separately for Category A and Category B event types, and for the SL assessment in accordance with Annex III Part A 2.3.(b), only for Category A event types. In case a railway operator performs multiple types of operations, as defined by Annex IV Part C, safety level scores shall be determined per type of operation.

The safety level score for the assessment on the frequency of events is performed as

$$SLS\ F_{SL\ class} = MCF_{SL\ class} \times \frac{N_{RP}}{V_{RP}} \times S$$

The safety level score for the assessment on the severity of accidents is performed as

$$SLS\ S_{SL\ class} = MCS_{SL\ class} \times \frac{N_{RP}}{V_{RP}} \times S$$

With

$MCF_{SL\ class}$	–	Minimal Change of Frequency for a given credibility level CL [%] derived from the assessment on the frequency of events
$MCS_{SL\ class}$	–	Minimal Change of Severity for a given credibility level CL [%] derived from the assessment on the severity of accidents
N_{RP}	–	Number of events in Reference Period
V_{RP}	–	Volume of transport of the railway operator in the Reference Period
S	–	Severity factor derived from mean severity (in FWSI) per event in the reference period, considering all reported events in ISS.

The safety level score shall be determined for the SL classes 1, 2, 4, and 5 as defined in section 1.2.2. This results in an overview of safety level scores for each event and sub event type.

Subsequently, the safety level score per event type is determined by selecting the score that is highest, either a) The safety level score of the event type or b) the sum of the safety level score of all related subevents.

The SL score for the railway operator is set as the highest value after the following calculations:

- the sum of the SL scores for all event types for fatalities (ESF) and severity (ESS) per SL class.
- The overall SL score for the operator for fatalities (ESF) and severity (ESS) per SL class (i.e. without calculating SL scores per event type). In this situation, S is set as the mean severity of all accidents in the reference period.

3. Assessment of safety level at national and Union levels

The methodology for the safety level assessment of railway operators as outlined in section 1 shall also be applied to each country. The differences are as follows:

- The word railway operator shall be understood as ‘country’.
- For each country, the assessments shall be performed separately for each type of operation as defined by Annex IV Part C.

The methodology for the safety level assessment of railway operators as outlined in section 1 shall also be applied to the European Union. The differences are as follows:

- The word railway operator shall be understood as 'European Union'.
- SL assessments in accordance with Annex III Part A 2.3.(b) shall not be performed.
- the SL assessments shall be performed separately for each type of operation as defined by Annex IV Part C.

Based on the assessments, the safety level scores at national and Union levels shall be determined.

4. Information on the implementation of the safety level assessments

A script that implements the assessments under section 1 shall be made publicly available by the Agency.

The ISS shall enable that the assessments under sections 1.2 and 1.3 can be repeated for each railway operator and event type excluding those events where the unauthorised presence of a third-party was the sole cause of the event.

PART C2 - DETAILED METHODS FOR THE SAFETY PERFORMANCE ASSESSMENT OF OPERATORS

1. Methodology for the safety performance assessment of operators

1.1. Assessments

The safety performance assessment of a railway operator consists of an assessment for each of the four areas established by Appendix B – Part C, followed by the assessment of the combined results for all four areas. This two-stepped assessment applies to each of the three objectives specified by Appendix C – Part B.

1.2. Methodology for assessing the levels by area

The method for assessing the areas under each objective is specified in the table below. The result is that the area safety performance is either lower, stable/average, or higher.

Area result	**Objective assessed**		
	a)	**b)**	**c)**
Lower	Maturity level in assessed period < maturity level in reference period	50% of operators in the group have a higher maturity level in than the operator in the assessed period	Maturity level in assessed period < maturity level mode in reference period
Stable/Average	Maturity level in assessed period = maturity level in reference period	The result of the assessment is neither lower nor higher	Maturity level in assessed period = maturity level mode in reference period
Higher	Maturity level in assessed period > maturity level in reference period	50% of operators in the group have a lower maturity level than the operator in the assessed period	Maturity level in assessed period > maturity level mode in reference period

The following additional rules apply:

- For objective b):
  A railway operator belongs to each group as defined by Appendix D - Part B for which the railway operator indicates that operations were performed in the assessed period. The assessment shall be applied separately for each group to which the railway operator belongs.
- For objective c):
  If there is no single maturity level mode in the reference period, the mode will be determined by including the maturity level in the assessed period. If there is still no mode, the highest most frequently reported maturity level will be selected.

1.3. Methodology for assessing the overall results

The overall safety performance assessment shall be performed using the combined results for the four areas, as assessed per objective according to section 2.3.

The table below prescribes how the assessment category, as specified under Annex III – Part B 3.2, is determined by establishing the number of times a certain result has been obtained.

Area result	Number of areas in which the result was obtained			
	1 area	2 areas	3 areas	4 areas
Lower	Class 3	Class 2	Class 1	Class 1
Stable/Average	Class 3	Class 3	Class 3	Class 3
Higher	Class 3	Class 4	Class 5	Class 5

The following additional rules apply to come to a final safety performance assessment:

- If Class 1, 2, 4, or 5 is noted, the safety performance assessment will never be Class 3.
- If Class 2 and Class 4 are both noted, the safety performance assessment shall be both classes.

2. Aggregation of Operators' safety performance at national and Union levels

All railway operators that reported on operations in a country shall be included in the aggregation at national levels.

The aggregation shall be performed for three groups as shown in the table below based on the entity codes as defined by Appendix D – Part B. There shall be a weighted and unweighted assessment for each of the three groups. The respective weighting factors are shown in the table below.

Aggregation group	Entity codes included	Weighting factor (Volume)
IM	IM-1	Number of train-kilometres
RU	RU-1, RU-2, RU-3, RU-4	Number of train-kilometres
Terminal operator	IM-2, RU-5	Number of railway vehicles processed in terminals

The results of the aggregation are a weighted and unweighted overview per group showing the proportion of railway operators with a certain maturity level, split by area.

The following formula shall be applied to determine the unweighted proportion of railway operators with maturity level 'i' of the total number of operators 'n' per country:

$$Share_i = \frac{\sum Operators_i}{\sum Operators_n}$$

The following formula shall be applied to determine the weighted proportion of railway operators with maturity level 'i' of the group total 'n' per country. It does so by taking the sum of the volumes of all railway operators with maturity level 'i', divided by the sum of the weighting factor for all railway operators within the group:

$$Share_i = \frac{\sum Volume_i}{\sum Volume_n}$$

The same method shall be applied to determine the values on the Union level.

ANNEX 2 - Minor consequential changes of the recommendation ERA REC1219 concerning the assessment of safety levels

Annex III – PART A

ASSESSMENT OF SAFETY LEVELS

Original

3. Applicable reference values and periods of time

3.2. Safety level assessment results

For each assessed objective and assessment period referred to in section 3.1 the Agency shall use the detailed process and criteria described in Appendix C – Part C and shall determine which of the following possible situations is applicable to the operator:

(a) Strong evidence for deterioration

(b) Moderate evidence for deterioration

(c) No evidence for improvement or deterioration

(d) Moderate evidence for improvement

(e) Strong evidence for improvement

Each assessment shall be accompanied by the consideration of statistical uncertainties in accordance with section 6.

Modification

3. Applicable reference values and periods of time

3.2. Safety level assessment results

For each assessed objective and assessment period referred to in section 3.1 the Agency shall use the detailed process and criteria described in Appendix C – Part C and shall determine which of the following possible situations is applicable to the operator:

SL class	SL assessment in accordance with 2.3 a)	SL assessment in accordance with 2.3 b)
1	Strong evidence for deterioration	Strong evidence for a lower level
2	Moderate evidence for deterioration	Moderate evidence for a lower level
3	No evidence for improvement or deterioration	No evidence for a lower or higher level
4	Moderate evidence for improvement	Moderate evidence for a higher level
5	Strong evidence for improvement	Strong evidence for a higher level

Each assessment shall be accompanied by the consideration of statistical uncertainties in accordance with section 6.

Original

5. Generic formula applied for individual railway operator's safety level estimation

5.1. Allocation of occurrences to involved railway operators

5.1.2. The following methods apply to the allocation of the counting of an occurrence to the category of railway operators responsible for the prevention or mitigation of the deemed cause of the accident occurrence.

5.1.3. The following counting rules apply:

(a) In case only one deemed cause – Cat. B event type – is identified. In this case the counting of the occurrence for the safety level estimation is allocated to the railway operator involved in the occurrence that is responsible for the part of the system which is deemed to have caused the occurrence.

(b) In case several combined causes – several Cat. B event types – are identified. In this case the counting of the occurrence for the safety level estimation is allocated in the applicable proportion(s) to the railway operator(s) involved in the occurrence that are responsible for the part(s) of the system which are deemed to have caused the occurrence.

(c) In case the cause(s) - Cat. B event type(s) - are not identified or there is a disagreement between the involved operators. In this case the counting of the occurrence for the safety level estimation is equally shared between the involved railway operator(s).

Modification

5. Generic formula applied for individual railway operator's safety level estimation

5.1. Allocation of occurrences to involved railway operators

5.1.2. (sic) The following methods apply to the allocation of an occurrence to the category of railway operators responsible for the prevention or mitigation of the deemed cause of the accident occurrence:

(a) where only one deemed cause (one category B event) of an occurrence is identified, the occurrence is allocated for the purpose of the safety level estimation to the railway operator involved in the occurrence that is responsible for the part of the system which is deemed to have caused the occurrence.

(b) where several combined causes (several category B events) of an occurrence are identified, the occurrence is allocated for the purpose of the safety level estimation once to each railway operator involved in a cause.

(c) where the causes (category B events) are not identified or there is a disagreement between the involved operators, the occurrence is allocated for the purpose of the safety level estimation to all involved railway operators.

ANNEX 3 - Minor consequential changes of the recommendation ERA REC1219 concerning the assessment of safety performance

Annex III – PART B

ASSESSMENT OF SAFETY PERFORMANCE

Original

3. Applicable reference values and periods of time

3.2. For each assessed objective and assessment period referred to in section 3.1 the Agency shall determine the situation applicable to the assessed operator by implementing the detailed method of Appendix C - Part C, allowing the following categorisation:

(a) Probable performance deterioration
(b) Potential performance deterioration
(c) Stable performance
(d) Potential performance improvement
(e) Probable performance improvement

Modified

3. Applicable reference values and periods of time

3.2. For each assessed objective and assessment period referred to in section 3.1 the Agency shall determine the situation applicable to the assessed operator by implementing the detailed method of Annex III - Part C, allowing the following categorisation:

SP class	SP assessment in accordance with 3.1 a) and c)	SP assessment in accordance with 3.1 b)
1	Strong evidence for deterioration	Strong evidence for lower performance
2	Moderate evidence for deterioration	Moderate evidence for lower performance
3	No evidence for improvement or deterioration	No evidence for lower or higher performance
4	Moderate evidence for improvement	Moderate evidence for higher performance
5	Strong evidence for improvement	Strong evidence for higher performance