Improving Incident Command Assessment Tool: Too Complex to Be Informative?

Stella Polikarpus*
Estonian Academy of Security Sciences †
stella.polikarpus@sisekaitse.ee

Reet Kasepalu
Estonian Academy of Security Sciences
reet.kasepalu@sisekaitse.ee

ABSTRACT

The Effective Command Behavioural Marker Framework (EC) is central to incident command training and assessment in Estonia, yet its implementation in practice, specifically how the assessment instrument is used by assessors, has not previously been examined using large-scale empirical data. This work-in-progress analyses 1,558 formal assessments across two command levels collected over nine years. We examine five-point and four-point scale utilisation across criteria and the implications of collapsing rarely used scale categories. Results show pronounced central-tendency scoring and limited use of extreme scale values, suggesting that parts of the current scale granularity and criterion set add little discriminative information in use. These findings indicate that assessment instrument complexity may constrain assessor judgement and reduce the transparency and formative value of feedback. We therefore propose an assessment approach that preserves the original conceptual structure while proposing clearer behavioural anchors and improved scoring defaults, to be validated prospectively in live assessments.

Keywords

Effective Command Behavioural Marker Framework (EC), Incident Command Assessment, Dynamic Decision-making Assessment, The Collaborative Authoring Process Model (CAPM).

INTRODUCTION

Rescue incident commanders in Estonia have been trained and assessed using virtual reality (VR) for more than 20 years (Põder et al. 2015; Polikarpus, Ley, et al. 2020). VR-based training provides a safe, low-risk environment not only for training but also for assessing command behaviours (Wheeler et al. 2024). To develop valid VR-based simulation scenarios and implement situational awareness (SA) assessments for rescue incident commanders, four foundational components are required (Polikarpus, Sarmiento-Márquez, et al. 2023):

1. The Effective Command Behavioural Marker Framework (EC) as the theoretical model to assess dynamic decision-making skills (Lamb et al. 2021).
2. The XVR On Scene VR environment as the technical platform (XVR Simulation 2023).
3. The Collaborative Authoring Process Model (CAPM) as the process for scenario development (Polikarpus, Ley, et al. 2021).
4. Certified assessors are needed to ensure methodological consistency.

Together, these elements form an integrated ecosystem through which the Estonian Academy of Security Sciences designs, delivers, and evaluates high-fidelity dynamic decision-making assessments within virtual simulation-based training (Polikarpus 2024; Polikarpus, Kasepalu, and Sarmiento-Márquez 2026). The assessment method has been well received by trainees (Polikarpus, Ley, et al. 2020).

*corresponding author
† While EASS is long time user of EC, framework itself is explained: https://www.effectivecommand.com

The EC framework has been used in Estonia for approximately ten years (Polikarpus, Ley, et al. 2020). Overall incident command assessment outcomes are communicated using a three-colour traffic-light scheme (red, amber, green), while the underlying summative assessment is based on a five-point coloured rating scale applied to behavioural criteria. Colour use has been shown to enhance visual communication and influence users' emotional responses (Singh 2006). The traffic-light scheme has two functions as a formative feedback mechanism and an organisational monitoring tool (Lamb et al. 2021). Currently, the summative assessment output is derived from 72 criteria organised into eight behavioural subsections (Q1 Perception [SA1], Q2 Comprehension [SA2], Q3 Prediction [SA3], Q4 Decision-making, Q5 Plan, Q6 Communication, Q7 Command & Control, and Q8 Review), each rated on a five-point scale, yielding 360 individual ratings per assessment.

Simplifying rater-based assessment instruments can improve assessment quality by reducing cognitive load on assessors and enabling more consistent and interpretable ratings of the performance of the trainee (Paravattil and Wilby 2019). When rating tasks become more complex, thereby increasing cognitive load for assessors, this can reduce the quality of their evaluations (Tavares et al. 2016).

This work-in-progress study analyses how the EC assessment instrument is enacted in practice in VR-based incident command training, focusing on whether its five-point scale and 72 criteria meaningfully differentiate performance. The requirement to generate up to 360 ratings per assessment may impose an unnecessary burden (*rater burden*) on assessors who provide formative feedback across multiple assessments within a single day (Polikarpus, Kasepalu, and Sarmiento-Márquez 2026). At the same time, the resulting volume of ratings may overwhelm trainees and reduce the clarity and interpretability of the formative feedback they receive, suggesting that both assessors and trainees could benefit from a more streamlined assessment approach.

Using 1,558 formal assessments collected over nine years, this study makes three contributions: i) it provides empirical evidence of scale compression and central-tendency scoring across subsections; ii) it analyses indications of redundant rating scale, supporting a candidate four-point recoding aligned with the operational colour logic; and iii) it proposes a modified short-form candidate (five criteria per subsection) that preserves rank-order correspondence with the full assessment instrument while reducing rating burden, to be validated in live assessments. Importantly, these analyses are based on simulated transformations of existing ratings and therefore do not assume that assessors would respond identically under a modified instrument. Rather, they provide an initial structural feasibility check. The proposed modifications require prospective validation in live assessment settings to examine potential changes in assessor behaviour.

Accordingly, the study addresses the following research questions:

- RQ1: How is the five-point rating scale used across assessment criteria and command levels?
- RQ2: Does collapsing the five-point rating scale to a four-point scale maintain reliability and conceptual coherence?
- RQ3: Which criteria show the highest discriminative potential, and what is lost or gained in a short-form version of the instrument?

Use of Effective Command Behavioural Marker Framework (EC) in Estonia

Since 2016, Estonia has transformed the training and assessment of rescue incident commanders by replacing traditional multiple-choice tests with a student-centred, VR-based approach (Polikarpus, Ley, et al. 2020). Training and assessment days are conducted in a safe yet realistic VR environment designed to practice and evaluate dynamic decision-making, while ensuring that tactical-level commanders nationwide meet the qualification standard (Polikarpus, Ley, et al. 2020; Polikarpus, Ley, et al. 2021).

The EC, originally developed in the United Kingdom and later expanded for international use, has been implemented in Estonia since 2016 (Polikarpus, Ley, et al. 2020; Lamb et al. 2021). EC serves as the theoretical foundation for incident command assessment and is structured around the SPAR model, comprising Situation Awareness, Planning, Action, and Review (Launder and Perry 2014). During a typical assessment session, two certified assessors observe a commander's performance within a virtual scenario and the following structured debrief, recording their judgements on a five-point colour-coded scale (Polikarpus, Ley, et al. 2020) (see Fig. 1).

Key design decisions in the development of the EC instrument have historically been made without direct involvement from Estonian stakeholders. For instance, the original binary scoring system used prior to 2016 was replaced by a five-point rating scale in 2017 (see Fig. 2), and a "not applicable" (N/A) option was subsequently introduced to account for criteria that could not be observed in certain scenarios (see Fig. 1). The default rating level is set at the amber category (3).

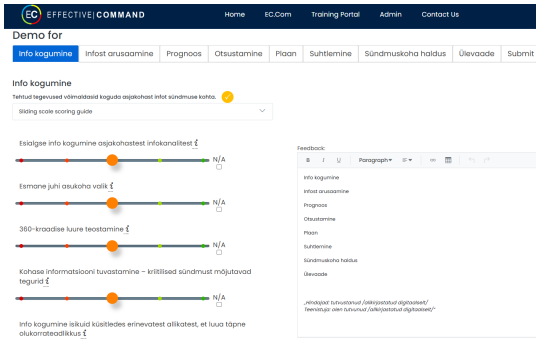

Figure 1. Sample Screenshot from EC Perception Subsection in Estonian Language 2026

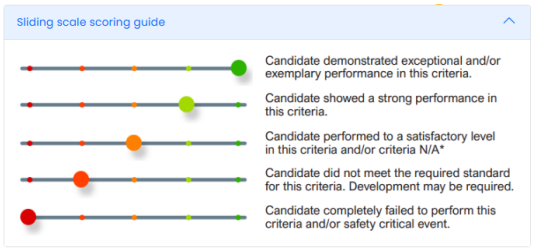

Figure 2. Screenshot of EC Scoring Guide

To support local adaptation and scenario authenticity, assessors in Estonia developed the Collaborative Authoring Process Model (CAPM), a five-step co-creation process enabling an assessors' team to design, test, and validate VR scenarios reflecting realistic operational challenges (Polikarpus, Ley, et al. 2021). The training and assessment day has been evaluated positively, with approximately 90% of Estonian commanders reporting that VR-based EC assessments represent a meaningful improvement in their professional attestation (Polikarpus, Ley, et al. 2020). The same study showed that the structure of the assessment day supports the trainees' basic psychological needs for autonomy, competence, and relatedness, which in turn contribute to high levels of engagement.

The EC evaluates incident commanders across eight behavioural subsections (Q1–Q8), each comprising nine criteria scored on a five-point scale (maximum 45 points per subsection) (Lamb et al. 2021). Scores reflect behavioural quality and safety using three anchor points: 5 (exceeds expectations), 3 (satisfactory), and 1 (safety-critical deficit) (see Fig. 2). Subsection scores are converted into a traffic-light system: Green (> 70%), Amber (55.5–70%), and Red (< 55.5%). Overall assessment outcomes are derived from subsection results according to predefined rules, including the presence of scores critical to safety and the number of red or green subsections (Lamb et al. 2021). Since 2019, in Estonia the meaning of colour-coded overall assessment results are: green indicating excellent decision-making competence, amber indicating competence at the occupational qualification threshold, and red indicating competence below the threshold (Polikarpus, Kasepalu, and Sarmiento-Márquez 2026).

METHOD

The initial EC dataset comprised 1,817 assessment records. Records related to demonstrations, training, monitoring, or attendance activities (n=114; 6.3%) were excluded because assessor certification could not be verified. In addition, 130 assessments from 2016 (7.1%) were removed due to a prior change in the scoring system (Polikarpus, Ley, et al. 2020). Screening of records from 2017 to January 2026 identified two certificates (0.1%) with scale-entry errors, nine records (0.5%) with missing identifiers, and four certificates (0.2%) issued in a non-standard language of assessment, which were all excluded. In total, 15 records (0.82%) were removed due to technical or assessor-related issues. The final dataset comprised 1,558 formal first- and second-level incident command assessments, each representing a single certificate-based assessment with all 360 criteria recordings.

The dataset included 1,399 first-level and 159 second-level assessments, representing 570 unique individuals. Commanders were assessed between one and six times during the study period, with three assessments per individual being the most common.

Data screening was conducted in Microsoft Excel, and statistical analyses were performed using JASP (Version 0.18.3; JASP Team:https://jasp-stats.org/). To address RQ1, criterion-level frequency distributions were analysed to examine the actual use of the five-point scale and identify unused or rarely used scores. To address RQ2, the original five-point scale was recoded into a four-point ordinal scale aligned with the operational traffic-light logic by merging conceptually adjacent categories. The four-point scale comprised: *Unsatisfactory* (Red–1), *Acceptable* (Amber–2), *Very Good* (Light Green–3), and *Exceptional* (Dark Green–4). The description of the four-point scale can be found in the Appendix.

Descriptive statistics (standard deviation (SD), skewness, and kurtosis) were calculated for all 72 criteria to assess variability, distributional shape, and central-tendency effects. Measurement equivalence between the original and reduced scales was evaluated using non-parametric techniques appropriate for ordinal data. To address RQ3, the five criteria with the highest SD were selected within each subsection. Internal consistency was examined using Cronbach's α for both the original and reduced scales, and Spearman rank-order correlations were calculated between subsection mean scores derived from the full (nine-criterion) and shortened (five-criterion) versions.

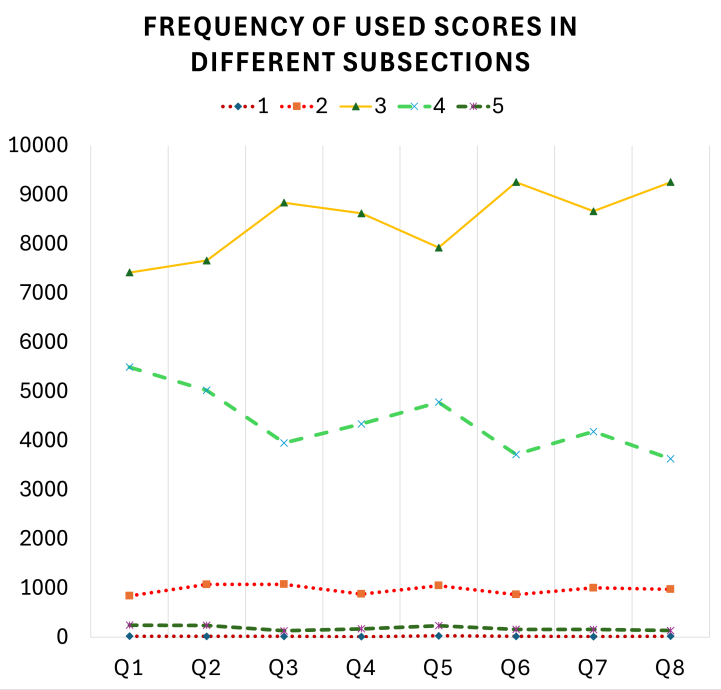

Figure 3. Five-point scale use of scores in different subsections for I and II command level (n=1558)

RESULTS

RQ1: How is the five-point scale actually used across assessment criteria and command levels?

A comparative overview of command level-specific differences in descriptive analyses is provided in Appendix Tables 5; 6; 7; 8; 9; 10; 11 and 12 for each subsection each criterion.

We illustrate how many times each score has been used in Fig. 3. Across the dataset, all 72 assessment criteria had an overall mean score of 3.27 and a SD = 0.6. Of these five-point scale criteria, 16 (22.2%) were never rated using the lowest score (1). Five of these criteria belonged to the decision-making behaviour subsection (Q4). Overall, the lowest score (1) was assigned only 157 times (0.14%) out of 112,176 individual ratings, whereas the central category (3; amber) was used 67,639 times (60.3%).

At the first command level, all 72 criteria were assigned the maximum score (5) at least once. Variability across behavioural subsections was primarily driven by ratings in the light green (4) and amber (3) scores. The mode was 3 for nearly all criteria in subsections Q2 to Q8 at both command levels (e.g., q5_1, q5_2, q6_1, q7_3, q8_2), indicating strong central-tendency.

However, scale use differed between command levels (see the Appendix). At Level II, certain criteria were never assessed using the lowest score (1). For example, within the Communication subsection (Q6), criterion q6_9 was never rated using red (scale points 1 or 2; Table 10). Similarly, the maximum score (5; dark green) was never assigned at Level II for selected criteria in Perception (Q1; q1_2; Table 5) and Prediction (Q3; q3_9; Table 7). Overall, dark red (1) was used only once at Level II (criterion q1_7), indicating that the effective rating range was largely restricted to 2–4. At Level I, the operational range was somewhat broader, with both minimum (1 or 2) and maximum (5) scores observed. Nevertheless, dark red (1) was not used in 16 criteria (see Tables 5–12), indicating similar compression tendencies.

The average SD across all 72 criteria for level I was 0.59 and for level II it was 0.60. Criterion q1_3 showed the highest variability (SD = 0.72), whereas q8_9 showed the lowest (SD = 0.44). Mean skewness was 0.13, with most values between –0.30 and +0.30, indicating near-symmetric distributions. Kurtosis averaged 0.32; only q8_9 showed elevated kurtosis (3.22), suggesting occasional concentration at extreme values. Across all criteria, mode clustered at 3, and variability remained limited (SD typically 0.50–0.70), providing evidence of central-tendency bias. Assessors consistently gravitated toward the midpoint of the five-point scale, resulting in compressed, symmetric distributions rather than the dispersion expected under full scale utilisation. The overall internal consistency of the 72-criterion instrument was high (α = 0.971; mean = 3.27, SD = 0.34). Subsection-level reliability estimates are presented in Table 1.

Table 1. Cronbach's α coefficients for the eight behavioural subsections using the five-point rating scale.

Subsection	α	Mean	SD
Q1 Perception (SA1)	.803	3.364	.383
Q2 Comprehension (SA2)	.818	3.313	.403
Q3 Prediction (SA3)	.854	3.221	.400
Q4 Decision-making	.863	3.270	.407
Q5 Plan	.818	3.295	.399
Q6 Communication	.874	3.224	.398
Q7 Command	.862	3.247	.411
Q8 Review	.869	3.206	.402

While Cronbach's α reflects internal consistency within each subsection, Spearman correlations between subsections reflect overlap between different behavioural domains. Therefore, we calculated the subsections Spearman's ρ for all subsections mean scores. All pairwise associations were positive and of moderate-to-strong magnitude ($\rho \approx 0.58$–0.795), indicating that higher scores on one behavioural domain tended to align with higher scores on the others. The strongest relations were observed between Q2 Comprehension (SA2) and Q3 Prediction (SA3) ($\rho = 0.795$), Q4 Decision-making and Q5 Plan ($\rho = 0.755$), and Q3 Prediction (SA3) and – Q5 Plan ($\rho = 0.735$), suggesting these subsections capture closely related behaviours. In contrast, the weakest (yet still moderate) correlations occurred for Q6 Communication and Q3 Prediction (SA3) ($\rho = 0.582$) and Q8 Review and and Q1 Perception (SA1) ($\rho = 0.596$), implying relatively more distinct content for those pairs. Overall, the pattern supports good coherence across the original subsections while preserving differentiation among them, consistent with a multi-faceted behavioural construct assessed. All correlations were smaller than the lowest Cronbach α in Table 1. Because all inter-subsection correlations were lower than the reliability coefficients of the individual subsections, the subsections can be considered related yet non-redundant, indicating that they capture distinct aspects of the underlying construct rather than duplicating one another.

RQ2: Does collapsing five-point rating scale to a four-point scale maintain reliability and conceptual coherence?

Descriptive analyses (see RQ 1 section) showed that not all criteria have been utilised across the full five-point range: 16 criteria (22.2%) were never assigned the lowest category (dark red; 1), and the dark red score was used only 157 times (0.14%) across the entire dataset (see Fig. 3). The findings from RQ1 indicated limited use of the lower end of the scale. Authors do not see the practical value to rate how much under the threshold the performance was by selecting between light and dark red. However, a 4-point, colour-coded scale could better support formative feedback by enabling the interpretation of progression between performance levels by making category distinctions more transparent. Prior research suggests that a four-point scale provides optimal discrimination and reliability without the drawbacks of excessive scale complexity (Jensen and Tøndering 2005).

To examine whether a shorter rating scale could maintain reliability and conceptual coherence, the original five-point scale (1–5) was collapsed into a four-point scale (1–4) by merging the two red categories (1 and 2) into a single category (1), while retaining amber as 2, light green as 3, and dark green as 4 (see Appendix for more details). Descriptive statistics (SD, skewness, and kurtosis) were recalculated for all 72 criteria. This reduction decreased the number of required ratings in the assessment matrix from 360 to 288, lowering assessors rater burden by 72 ratings.

As expected, the resulting distributions on the four-point scale closely mirrored those of the original scale. SD remained comparable (overall SD = 0.591 vs. 0.596 in the original data), and modes continued to cluster in the central categories (predominantly 2 and 3), indicating persistent central-tendency scoring. Skewness remained low (mean = 0.19), only slightly higher than in the original data (0.13), while average kurtosis decreased (0.16 vs. 0.32), suggesting continued concentration of scores around the centre. As expected, collapsing scale categories did not meaningfully alter central-tendency patterns, given that the same underlying data were used. Notably, the current system already defaults to the amber colour (3), and both amber and N/A responses are treated equivalently in the scoring logic (see Fig. 1).

To assess whether scale reduction affected measurement quality, internal consistency was recalculated for the mean scores of each subsection. As shown in Table 2, all eight behavioural dimensions retained high reliability, with Cronbachs α values ranging from .803 to .874. These values are comparable to those obtained with the original five-point scale (see Table 1), indicating that reducing the number of response categories did not compromise reliability.

Table 2. Cronbach's α coefficients for the eight behavioural subsections using the four-point rating scale.

Subsection	New α	New Mean	New SD
Q1 Perception (SA1)	.803	2.365	.380
Q2 Comprehansion (SA2)	.818	2.315	.400
Q3 Prediction (SA3)	.855	2.223	.398
Q4 Decision-making	.863	2.271	.406
Q5 Plan	.819	2.298	.395
Q6 Communication	.874	2.225	.395
Q7 Command	.862	2.248	.409
Q8 Review	.868	2.207	.398

Table 3. Cronbach's α, means, and standard deviations for the shortened four-point behavioural subsections.

Subsection	Included criteria	α	Short M	Short SD
SNQ1	Nq1_1; Nq1_3; Nq1_4; Nq1_5; Nq1_6	.683	2.377	.424
SNQ2	Nq2_1; Nq2_2; Nq2_6; Nq2_7; Nq2_8	.750	2.351	.461
SNQ3	Nq3_1; Nq3_2; Nq3_6; Nq3_7; Nq3_8	.784	2.249	.447
SNQ4	Nq4_1; Nq4_2; Nq4_4; Nq4_5; Nq4_6	.811	2.275	.472
SNQ5	Nq5_1; Nq5_3; Nq5_4; Nq5_5; Nq5_6	.766	2.365	.471
SNQ6	Nq6_1; Nq6_2; Nq6_3; Nq6_4; Nq6_9	.813	2.248	.445
SNQ7	Nq7_1; Nq7_2; Nq7_3; Nq7_4; Nq7_6	.817	2.284	.478
SNQ8	Nq8_1; Nq8_2; Nq8_3; Nq8_4; Nq8_8	.815	2.215	.458

Taken together, these results show that the four-point scale preserves the reliability and structure of the original instrument while reducing measurement complexity.

RQ3: Which criteria show the highest discriminative potential, and what is lost or gained in a short-form version of the instrument?

To reduce the complexity of the assessment tool, we identified the criteria that showed the greatest variation across assessments, as indicated by higher standard deviations. Appendix Table 13 presents the 34 criteria with a standard deviation above the selected cut-point ($SD > 0.60$). The remaining 38 criteria ($SD \leq 0.60$) are not shown in the table, as their low variability indicates pronounced central-tendency scoring.

Variability differed across behavioural subsections. In Q6 (Communication), only two criteria exceeded the SD threshold. In Q3 (Prediction; SA3) and Q4 (Decision-making), three criteria each showed SD > 0.60. Four criteria exceeded the threshold in Q1 (Perception; SA1) and Q8 (Review), while Q2 (Comprehension; SA2) and Q7 (Command) each contained seven such criteria. On average, four criteria per subsection demonstrated substantial variability, suggesting more consistent use of the full five-point scale. To retain the most discriminative indicators while reducing instrument length, five criteria per subsection with the highest variability were selected. Internal consistency (Cronbach's α) was recalculated for the shortened version using the four-point scale. The results are reported in Table 3.

As shown in Table 3, Cronbach's α values for the shortened subsections were lower in all cases compared to those obtained earlier (see Table 2). While the original nine criteria in subsections demonstrated strong internal consistency across all behavioural dimensions (α = .80–.87) (see Table 1), criteria reduction affected early cognitive processes (particularly Q1 Perception α = 0.68) more strongly than action-oriented behaviours (Q4 to Q8 α between 0.77 to 0.82). This suggests that perceptual dimensions rely on a wider range of contextual indicators, whereas decision-making and command behaviours remain internally coherent even with fewer criteria. We propose that wider use of assessment criteria in Q1 subsection reflects assessors behaviour rather than conceptual weakness of shortened assessment tool.

Spearman correlations between the original and shortened subsection means were very high (ρ = .93–.96) (see Table 4), indicating strong correspondence between the two versions. For each behaviour, the shortened scale correlated most strongly with its original counterpart, providing evidence that criterion reduction preserved construct meaning and relative ordering across participants. The original instrument required assessors to evaluate 360 individual criteria (eight subsections with nine criteria each, using a five-point scale). The shortened version reduces this

Table 4. Spearman's correlations (ρ) between mean scores of the original subsections (Q1–Q8) and the shortened subsections (SNQ1–SNQ8).

Variable	ρ	Q1M	Q2M	Q3M	Q4M	Q5M	Q6M	Q7M	Q8M
SNQ1M	ρ	**.932**	.616	.587	.565	.549	.534	.534	.536
SNQ2M	ρ	.635	**.932**	.728	.631	.667	.510	.587	.586
SNQ3M	ρ	.602	.752	**.948**	.682	.700	.527	.604	.603
SNQ4M	ρ	.593	.667	.706	**.956**	.729	.590	.690	.661
SNQ5M	ρ	.560	.679	.690	.692	**.936**	.498	.645	.608
SNQ6M	ρ	.570	.561	.542	.595	.567	**.955**	.626	.605
SNQ7M	ρ	.567	.620	.621	.685	.663	.629	**.960**	.648
SNQ8M	ρ	.576	.611	.616	.658	.638	.608	.664	**.960**

requirement to 160 criteria (eight subsections with five criteria each, using a four-point scale), resulting in 200 fewer judgments per assessment. This reduction is expected to lower rater burden and fatigue, particularly in time-constrained or repeated assessment contexts.

DISCUSSION

The findings suggest a clear misalignment between the intended granularity of the five-point scale and its use in practice. Although the EC is designed to support fine-grained behavioural differentiation (Lamb et al. 2021), assessors consistently avoided extreme ratings, resulting in pronounced central-tendency bias, particularly at the second command level. Instead, the results imply that scale complexity may exceed what is required for reliable and consistent evaluation in dynamic incident command contexts. From this perspective, reducing the number of scale points may better reflect actual assessment behaviour while supporting more manageable and interpretable evaluation practices, without compromising the core behavioural distinctions intended by the framework.

Within the analysed training scenarios, nearly half of the 72 criteria exhibited limited variability, suggesting restricted discriminative value in this specific assessment context. In contrast, a smaller subset of criteria demonstrated greater differentiation in the assessors' judgement. Retaining the five most discriminating criteria within each subsection allowed the construction of a shortened instrument that preserved the eight-subsection conceptual structure while substantially reducing the number of required ratings. As expected, internal consistency decreased following criteria reduction, particularly within the Q1 subsection. This difference should be interpreted in the light of both the original reliability estimates and the criterion-selection strategy applied during scale reduction. In the original dataset, the Q1 Perception subsection already demonstrated the lowest Cronbach's α among the eight behaviours (α = .803), compared to higher values for more consolidated behaviours such as Q4 Decision-making, Q6 Communication, and Q8 Review. This pattern suggests that perceptual processes were inherently less consistent, even prior to shortening the scale, possibly because they rely on multiple cues rather than clearly defined behaviours (Endsley 2000).

Moreover, the shortened instrument deliberately retained only those criteria that exhibited the highest variability across assessments. While this strategy supports discriminative capacity and interpretability, it also constrains internal consistency estimates, as Cronbach's α is sensitive to item intercorrelation and variance structure (Tavakol and Dennick 2011). While internal consistency typically decreases when items are removed, high reliability in longer instruments may partly reflect overlapping items rather than more precise measurement of the construct (Smith et al. 2000). From this perspective, the reduced α for SNQ1 represents a predictable statistical outcome of variability-driven criterion selection rather than a weakening of construct representation.

Importantly, evidence from correlations between the means of the original and shortened subsections further support this interpretation. The very high correlations between the original and shortened scales ($\rho \approx .93$–$.96$) indicate strong construct continuity despite both scale and criterion reduction. Such levels of correspondence suggest that the removed criteria were largely psychometrically redundant rather than conceptually essential, as the shortened scales preserve the relative ordering and substantive meaning of the original behavioural constructs. Taken together with the original reliability estimates, these findings indicate continuity in the relative ordering of behavioural subsections, with perception remaining structurally distinct from later, action-oriented behaviours. This pattern supports a process-based interpretation of the behavioural model, with lower internal consistency observed in perception-related ratings compared to later stages. This may reflect either greater context sensitivity in early processes (Blömeke et al. 2015) or hint to limitations in how these aspects are captured by the assessment instrument. Importantly, the shortened subsections remained strongly correlated with their original versions (see

Table 4), showing that the essential behavioural meaning was retained. These results suggest that a more compact assessment tool could reduce rater burden without substantially compromising measurement integrity, provided future studies validate the shortened scale in real assessment conditions.

Reducing the number of required ratings from 360 to 160 might lower assessor rater burden by decreasing the number of elements that must be processed during a single assessment. From a cognitive load theory perspective, task demands increase as more interacting elements compete for limited working memory resources (Sweller 2024). In complex evaluative tasks involving repeated judgments, high element interactivity can lead to increased fatigue and reduced attentional resources. By eliminating redundant criteria while preserving core behavioural structure, the shortened instrument reduces unnecessary rater burden without compromising the interpretability or developmental value of the assessment.

Shortening the scale and reducing the number of assessment criteria has implications not only for assessors but also for model transparency and learner-facing interpretability. When commanders can understand how they are assessed, the criteria that are used and what is expected of them, the assessment process itself can become a source of learning. In such cases, assessment moves beyond a one-off evaluative requirement and instead supports reflection, sense-making, and behavioural development. Model-based learning analytics emphasises explicit, theory-driven representations of competence that can be inspected and understood by users (Pishtari et al. 2023). A simpler model is easier to integrate into dashboards and simulation-based assessments, because too much complexity can make feedback harder to understand and less useful for learning. In this sense, the shortened instrument aligns with the dual aim of model-based learning analytics: supporting valid assessment while enabling sense-making and self-regulation for trainees.

To change the underlying behavioural tendency to choose mid-scale categories, the default setting should be N/A and if the assessor makes up his/her mind what the evaluation of the specific criterion is, he/she clicks on the coloured four-point scale to mark it. While it is possible to shorten a scale, it is crucial to carefully evaluate the impact on reliability and validity in future studies. The specific context in which the scale is used is VR-based assessment, and reducing the assessors' rater burden could also possibly be achieved with a better design of the VR scenarios (Polikarpus and Kasepalu 2025). Empirical testing and methodological rigour are essential to ensure that the shortened scale remains effective to assess command behaviours using EC. A natural next step would be to examine the proposed criteria using confirmatory factor analysis (CFA). Given the theory-driven definition of the eight behavioural dimensions, CFA could be used to test the extent to which the shortened criteria align with the hypothesised structure and to explore relationships between SA phases and later action-oriented behaviours.

Limitation

This work-in-progress study has several limitations. First, the analyses rely on historical assessment data, meaning that both the original five-point scale and the reduced four-point scale were applied to the same dataset; therefore, differences between scales reflect recoding rather than genuine assessor behaviour. Second, the study does not yet include empirical testing of the shortened and redesigned default settings in the assessment tool and it must be validated in real assessment conditions. Third, assessor behaviour may be influenced by the existing system design, such as the default colour settings or the structure of the scenario, which was not experimentally controlled in this study. Finally, because the entire dataset originates from one national system, generalizability to other countries or training contexts remains limited and requires further research.

CONCLUSION

This work-in-progress study analysed nine years of Estonian EC assessment data from VR-based incident command assessments to examine how the current 72-criterion, five-point scale functions in practice. The results indicate limited use of extreme scale values suggesting that the instrument could be more complex than necessary for reliable assessment. Despite the current use of a five-point scale, assessors primarily rely on a narrow rating range, particularly at the second command level, indicating that the theoretical granularity of the scale is not realised in practice. Recoding the scale to four points preserved score distributions and showed high internal consistency across eight behavioural subsections. Furthermore, criterion reduction showed that a smaller subset of criteria could account for most meaningful differentiation in assessor judgements. Overall, the findings demonstrate that criterion and scale reduction can potentially support a more efficient and interpretable assessment without compromising measurement integrity, while also enhancing learner-facing interpretability. Future research should validate the reduced assessment tool in live assessments, examine the influence of interface design features such as default colour settings, refine behavioural anchors, and explore scenario, training, and cross-national factors to support the development of a robust, evidence-based EC.

REFERENCES

Blömeke, S., Gustafsson, J.-E., and Shavelson, R. J. (Jan. 2015). "Beyond Dichotomies". In: *Zeitschrift für Psychologie* 223.1, pp. 3–13.

Endsley, M. R. (Jan. 2000). "Theoretical Underpinnings Of Situation Awareness: A Critical Review". In: *Situation Awareness Analysis and Measurement*. Ed. by M. R. Endsley and D. J. Garland. Lawrence Erlbaum Associates Publishers.

Jensen, C. and Tøndering, J. (Sept. 2005). "Choosing a scale for measuring perceived prominence". In: *Interspeech 2005*. ISCA: ISCA, pp. 2385–2388.

Lamb, K., Farrow, M., Olymbios, C., Launder, D., and Greatbatch, I. (June 2021). "Systematic incident command training and organisational competence". In: *International Journal of Emergency Services* 10.2, pp. 222–234.

Launder, D. and Perry, C. (Oct. 2014). "A study identifying factors influencing decision making in dynamic emergencies like urban fire and rescue settings". In: *International Journal of Emergency Services* 3.2, pp. 144–161.

Paravattil, B. and Wilby, K. J. (Nov. 2019). "Optimizing assessors' mental workload in rater-based assessment: a critical narrative review". In: *Perspectives on Medical Education* 8.6, pp. 339–345.

Pishtari, G., Ley, T., Khalil, M., Kasepalu, R., and Tuvi, I. (May 2023). "Model-Based Learning Analytics for a Partnership of Teachers and Intelligent Systems: A Bibliometric Systematic Review". In: *Education Sciences* 13.5, p. 498.

Põder, S.-F., Savimaa, R., and Link, M. (2015). "A framework for training internal security officers to manage joint response events in a virtual learning environment." In: *Proceedings Estonian Academy of Security Sciences: Sustained Security*, pp. 151–180.

Polikarpus, S. (2024). "The Role of Trainers in Designing and Implementing Virtual Simulation-Based Training in Rescue Organisations". PhD thesis. Tallinn University, p. 207.

Polikarpus, S. and Kasepalu, R. (May 2025). "Complexity Level of Virtual Simulation Scenarios for Command and Control Behaviors Assessment". In: *Proceedings of the International ISCRAM Conference* May.

Polikarpus, S., Kasepalu, R., and Sarmiento-Márquez, E. M. (2026). "From Dynamic Decision-Making Assessments Using Virtual Simulation-Based Training to Targeted Training of Incident Commanders". In: *Information Technology in Disaster Risk Reduction (ITDRR2024)*. Ed. by W. Seböck, T. J. Lampoltshammer, J. Dugdale, and I. Zeller. Springer Nature SwitzerlandAG, Cham (currently in press), pp. 82–97.

Polikarpus, S., Ley, T., and Poom-Valickis, K. (2020). "Developing the Situational Awareness of Incident Commanders: Evaluating a Training Programme using a Virtual Simulation". In: *Proceedings Estonian Academy of Security Sciences* 19, pp. 195–226.

Polikarpus, S., Ley, T., and Poom-Valickis, K. (2021). "Collaborative Authoring of Virtual Simulation Scenarios for Assessing Situational Awareness". In: *Proceedings of the 18th ISCRAM Conference*. Ed. by A. Adrot, R. Grace, K. Moore, and C. Zobel. Blacksburg: Blacksburg, VA, USA, pp. 229–237.

Polikarpus, S., Sarmiento-Márquez, E. M., and Ley, T. (2023). "Creation and Use of Virtual Simulations for Measuring Situation Awareness of Incident Commanders". In: *ITDRR2022*. Ed. by T. Gjøsæter, J. Radianti, and Y. Murayama. Informatio. Springer, Cham, pp. 23–38.

Singh, S. (July 2006). "Impact of color on marketing". In: *Management Decision* 44.6, pp. 783–789.

Smith, G. T., McCarthy, D. M., and Anderson, K. G. (2000). "On the sins of short-form development." In: *Psychological Assessment* 12.1, pp. 102–111.

Sweller, J. (Feb. 2024). "Cognitive load theory and individual differences". In: *Learning and Individual Differences* 110.December 2023, p. 102423.

Tavakol, M. and Dennick, R. (June 2011). "Making sense of Cronbach's alpha". In: *International Journal of Medical Education* 2, pp. 53–55.

Tavares, W., Ginsburg, S., and Eva, K. W. (Jan. 2016). "Selecting and Simplifying: Rater Performance and Behavior When Considering Multiple Competencies". In: *Teaching and Learning in Medicine* 28.1, pp. 41–51.

Wheeler, S. G., Hoermann, S., Lukosch, S., and Lindeman, R. W. (Feb. 2024). "Design and assessment of a virtual reality learning environment for firefighters". In: *Frontiers in Computer Science* 6.

XVR Simulation (2023). *XVR:Virtual Reality training software for safety and security*.

APPENDIX

Table 5. Descriptive statistics of I and II command levels for Q1 perception (SA1)

Criterion	*Command*	*Mode*	*Mean*	*SD*	*Skewness*	*Kurtosis*	*Range*	*Min*	*Max*
q1_1	I (n = 1399)	4	3.65	0.599	-0.595	0.392	4	1	5
q1_1	II (n = 159)	4	3.516	0.583	-0.547	-0.445	3	2	5
q1_2	I (n = 1399)	4	3.57	0.582	-0.440	0.161	4	1	5
q1_2	II (n = 159)	3	3.283	0.657	-0.374	-0.736	2	2	4
q1_3	I (n = 1399)	4	3.403	0.724	-0.339	-0.165	4	1	5
q1_3	II (n = 159)	3	3.270	0.691	0.169	-0.063	3	2	5
q1_4	I (n = 1399)	3	3.387	0.600	-0.017	-0.240	4	1	5
q1_4	II (n = 159)	3	3.252	0.626	0.071	-0.102	3	2	5
q1_5	I (n = 1399)	3	3.287	0.635	-0.094	-0.162	4	1	5
q1_5	II (n = 159)	3	3.296	0.591	0.187	-0.044	3	2	5
q1_6	I (n = 1399)	3	3.156	0.650	0.008	0.363	4	1	5
q1_6	II (n = 159)	3	3.434	0.651	-0.030	-0.217	3	2	5
q1_7	I (n = 1399)	3	3.311	0.517	0.696	-0.071	3	2	5
q1_7	II (n = 159)	3	3.201	0.593	0.282	1.667	4	1	5
q1_8	I (n = 1399)	3	3.343	0.583	0.153	-0.254	3	2	5
q1_8	II (n = 159)	3	3.371	0.622	-0.137	-0.370	3	2	5
q1_9	I (n = 1399)	3	3.194	0.584	0.244	0.599	4	1	5
q1_9	II (n = 159)	3	3.384	0.614	0.040	-0.252	3	2	5

Table 6. Descriptive statistics of I and II command levels for Q2 comprehension (SA2)

Criterion	*Command*	*Mode*	*Mean*	*SD*	*Skewness*	*Kurtosis*	*Range*	*Min*	*Max*
q2_1	I (n = 1399)	4	3.515	0.648	-0.538	-0.084	4	1	5
q2_1	II (n = 159)	3	3.390	0.655	-0.337	-0.436	3	2	5
q2_2	I (n = 1399)	4	3.443	0.656	-0.353	-0.349	3	2	5
q2_2	II (n = 159)	3	3.214	0.669	-0.020	-0.333	3	2	5
q2_3	I (n = 1399)	3	3.227	0.589	0.150	0.316	4	1	5
q2_3	II (n = 159)	3	3.151	0.576	0.193	0.370	3	2	5
q2_4	I (n = 1399)	3	3.162	0.599	0.122	0.885	4	1	5
q2_4	II (n = 159)	3	3.157	0.680	0.039	-0.326	3	2	5
q2_5	I (n = 1399)	3	3.259	0.594	0.215	0.487	4	1	5
q2_5	II (n = 159)	3	3.233	0.565	0.421	0.507	3	2	5
q2_6	I (n = 1399)	3	3.236	0.649	0.277	0.475	4	1	5
q2_6	II (n = 159)	3	3.094	0.571	0.215	0.634	3	2	5
q2_7	I (n = 1399)	3	3.224	0.652	-0.061	-0.096	4	1	5
q2_7	II (n = 159)	3	3.044	0.609	0.148	0.248	3	2	5
q2_8	I (n = 1399)	3	3.407	0.671	-0.226	-0.269	4	1	5
q2_8	II (n = 159)	3	3.346	0.684	0.033	-0.199	3	2	5
q2_9	I (n = 1399)	3	3.444	0.608	-0.071	-0.370	3	2	5
q2_9	II (n = 159)	3	3.314	0.586	-0.005	-0.360	3	2	5

Table 7. Descriptive statistics of I and II command levels for Q3 prediction (SA3)

Criterion	*Command*	*Mode*	*Mean*	*SD*	*Skewness*	*Kurtosis*	*Range*	*Min*	*Max*
q3_1	I (n = 1399)	3	3.392	0.621	-0.293	-0.256	4	1	5
q3_1	II (n = 159)	3	3.233	0.553	0.273	0.173	3	2	5
q3_2	I (n = 1399)	3	3.351	0.653	-0.184	-0.122	4	1	5
q3_2	II (n = 159)	3	3.245	0.672	-0.209	-0.602	3	2	5
q3_3	I (n = 1399)	3	3.181	0.569	0.275	0.420	3	2	5
q3_3	II (n = 159)	3	3.164	0.605	0.260	0.424	3	2	5
q3_4	I (n = 1399)	3	3.244	0.580	0.200	0.058	3	2	5
q3_4	II (n = 159)	3	3.157	0.601	0.102	0.079	3	2	5
q3_5	I (n = 1399)	3	3.174	0.546	0.504	1.186	4	1	5
q3_5	II (n = 159)	3	3.208	0.596	0.079	-0.038	3	2	5
q3_6	I (n = 1399)	3	3.173	0.619	0.050	0.340	4	1	5
q3_6	II (n = 159)	3	3.044	0.620	0.134	0.124	3	2	5
q3_7	I (n = 1399)	3	3.137	0.590	0.021	0.610	4	1	5
q3_7	II (n = 159)	3	3.031	0.578	0.199	0.700	3	2	5
q3_8	I (n = 1399)	3	3.232	0.597	0.050	0.114	4	1	5
q3_8	II (n = 159)	3	3.220	0.592	0.086	-0.047	3	2	5
q3_9	I (n = 1399)	3	3.157	0.497	0.619	1.924	4	1	5
q3_9	II (n = 159)	3	3.258	0.565	-0.026	-0.416	2	2	4

Table 8. Descriptive statistics of I and II command levels for Q4 decision-making

Criterion	*Command*	*Mode*	*Mean*	*SD*	*Skewness*	*Kurtosis*	*Range*	*Min*	*Max*
q4_1	I (n = 1399)	3	3.365	0.582	0.093	-0.359	3	2	5
q4_1	II (n = 159)	3	3.239	0.556	0.251	0.115	3	2	5
q4_2	I (n = 1399)	3	3.212	0.600	0.117	-0.002	3	2	5
q4_2	II (n = 159)	3	3.189	0.542	0.351	0.560	3	2	5
q4_3	I (n = 1399)	3	3.235	0.556	0.352	0.658	4	1	5
q4_3	II (n = 159)	3	3.277	0.502	0.687	0.021	3	2	5
q4_4	I (n = 1399)	3	3.261	0.717	-0.140	-0.194	4	1	5
q4_4	II (n = 159)	3	3.151	0.722	-0.031	-0.645	3	2	5
q4_5	I (n = 1399)	3	3.206	0.641	0.160	0.228	4	1	5
q4_5	II (n = 159)	3	3.346	0.584	0.141	-0.248	3	2	5
q4_6	I (n = 1399)	3	3.342	0.603	0.068	-0.119	4	1	5
q4_6	II (n = 159)	3	3.321	0.588	0.163	-0.147	3	2	5
q4_7	I (n = 1399)	3	3.373	0.563	0.221	-0.417	3	2	5
q4_7	II (n = 159)	3	3.365	0.641	-0.212	-0.401	3	2	5
q4_8	I (n = 1399)	3	3.260	0.510	0.659	0.284	3	2	5
q4_8	II (n = 159)	3	3.333	0.536	0.825	0.172	3	2	5
q4_9	I (n = 1399)	3	3.170	0.507	0.525	0.984	3	2	5
q4_9	II (n = 159)	3	3.245	0.559	0.230	0.060	3	2	5

Table 9. Descriptive statistics of I and II command levels for Q5 plan

Criterion	*Command*	*Mode*	*Mean*	*SD*	*Skewness*	*Kurtosis*	*Range*	*Min*	*Max*
q5_1	I (n = 1399)	3	3.438	0.626	-0.273	0.064	4	1	5
q5_1	II (n = 159)	3	3.377	0.672	-0.366	-0.481	3	2	5
q5_2	I (n = 1399)	3	3.352	0.623	-0.064	0.014	4	1	5
q5_2	II (n = 159)	3	3.283	0.628	0.014	-0.202	3	2	5
q5_3	I (n = 1399)	3	3.226	0.618	0.065	0.232	4	1	5
q5_3	II (n = 159)	3	3.138	0.707	0.015	-0.519	3	2	5
q5_4	I (n = 1399)	3	3.301	0.687	-0.113	-0.191	4	1	5
q5_4	II (n = 159)	3	3.226	0.665	-0.159	-0.545	3	2	5
q5_5	I (n = 1399)	4	3.442	0.687	-0.261	-0.241	4	1	5
q5_5	II (n = 159)	3	3.321	0.687	-0.041	-0.292	3	2	5
q5_6	I (n = 1399)	4	3.447	0.674	-0.387	-0.095	4	1	5
q5_6	II (n = 159)	4	3.440	0.632	-0.381	-0.386	3	2	5
q5_7	I (n = 1399)	3	3.104	0.513	0.472	2.325	4	1	5
q5_7	II (n = 159)	3	3.145	0.549	0.767	1.959	3	2	5
q5_8	I (n = 1399)	3	3.096	0.552	0.172	1.849	4	1	5
q5_8	II (n = 159)	3	3.101	0.506	0.776	2.700	3	2	5
q5_9	I (n = 1399)	3	3.281	0.609	0.028	0.276	4	1	5
q5_9	II (n = 159)	3	3.377	0.559	0.278	-0.420	3	2	5

Table 10. Descriptive statistics of I and II command levels for Q6 communication

Criterion	*Command*	*Mode*	*Mean*	*SD*	*Skewness*	*Kurtosis*	*Range*	*Min*	*Max*
q6_1	I (n = 1399)	3	3.306	0.577	0.124	0.227	4	1	5
q6_1	II (n = 159)	3	3.308	0.573	0.285	-0.031	3	2	5
q6_2	I (n = 1399)	3	3.320	0.570	0.193	0.263	4	1	5
q6_2	II (n = 159)	3	3.239	0.556	0.251	0.115	3	2	5
q6_3	I (n = 1399)	3	3.156	0.656	0.054	0.132	4	1	5
q6_3	II (n = 159)	3	3.239	0.641	0.469	0.556	3	2	5
q6_4	I (n = 1399)	3	3.080	0.605	0.313	1.194	4	1	5
q6_4	II (n = 159)	3	3.107	0.612	0.275	0.537	3	2	5
q6_5	I (n = 1399)	3	3.046	0.548	0.185	1.414	4	1	5
q6_5	II (n = 159)	3	3.044	0.640	0.109	-0.098	3	2	5
q6_6	I (n = 1399)	3	3.214	0.512	0.681	1.141	4	1	5
q6_6	II (n = 159)	3	3.201	0.461	1.077	1.364	3	2	5
q6_7	I (n = 1399)	3	3.284	0.519	0.534	-0.047	3	2	5
q6_7	II (n = 159)	3	3.270	0.581	0.475	0.456	3	2	5
q6_8	I (n = 1399)	3	3.242	0.508	0.740	0.632	3	2	5
q6_8	II (n = 159)	3	3.220	0.499	0.690	0.709	3	2	5
q6_9	I (n = 1399)	3	3.364	0.562	0.368	-0.082	4	1	5
q6_9	II (n = 159)	3	3.396	0.540	0.912	-0.252	2	3	5

Table 11. Descriptive statistics of I and II command levels for Q7 command

Criterion	*Command*	*Mode*	*Mean*	*SD*	*Skewness*	*Kurtosis*	*Range*	*Min*	*Max*
q7_1	I (n = 1399)	3	3.218	0.707	-0.155	-0.402	4	1	5
q7_1	II (n = 159)	3	3.176	0.742	-0.202	-0.964	3	2	5
q7_2	I (n = 1399)	3	3.342	0.606	0.054	-0.127	4	1	5
q7_2	II (n = 159)	3	3.283	0.657	0.032	-0.186	3	2	5
q7_3	I (n = 1399)	3	3.315	0.606	-0.052	-0.357	3	2	5
q7_3	II (n = 159)	3	3.220	0.726	-0.162	-0.717	3	2	5
q7_4	I (n = 1399)	3	3.250	0.617	0.056	0.203	4	1	5
q7_4	II (n = 159)	3	3.044	0.520	0.335	1.733	3	2	5
q7_5	I (n = 1399)	3	3.174	0.475	0.759	1.282	3	2	5
q7_5	II (n = 159)	3	3.390	0.562	0.228	-0.475	3	2	5
q7_6	I (n = 1399)	3	3.330	0.594	0.249	0.075	4	1	5
q7_6	II (n = 159)	3	3.314	0.648	0.154	-0.035	3	2	5
q7_7	I (n = 1399)	3	3.197	0.583	0.182	0.313	4	1	5
q7_7	II (n = 159)	3	3.094	0.525	0.377	1.386	3	2	5
q7_8	I (n = 1399)	3	3.190	0.537	0.399	1.023	4	1	5
q7_8	II (n = 159)	3	3.252	0.551	0.280	0.072	3	2	5
q7_9	I (n = 1399)	3	3.249	0.581	0.168	0.293	4	1	5
q7_9	II (n = 159)	3	3.088	0.532	0.343	1.288	3	2	5

Table 12. Descriptive statistics of I and II command levels for Q8 review

Criterion	*Command*	*Mode*	*Mean*	*SD*	*Skewness*	*Kurtosis*	*Range*	*Min*	*Max*
q8_1	I (n = 1399)	3	3.252	0.606	-0.053	-0.085	4	1	5
q8_1	II (n = 159)	3	3.220	0.592	0.086	-0.047	3	2	5
q8_2	I (n = 1399)	3	3.194	0.629	-0.070	0.275	4	1	5
q8_2	II (n = 159)	3	3.226	0.655	-0.135	-0.488	3	2	5
q8_3	I (n = 1399)	3	3.202	0.609	0.093	0.209	4	1	5
q8_3	II (n = 159)	3	3.170	0.731	-0.078	-0.720	3	2	5
q8_4	I (n = 1399)	3	3.175	0.608	0.046	0.518	4	1	5
q8_4	II (n = 159)	3	3.151	0.553	0.512	1.229	3	2	5
q8_5	I (n = 1399)	3	3.269	0.583	0.416	0.337	3	2	5
q8_5	II (n = 159)	3	3.164	0.625	0.027	-0.160	3	2	5
q8_6	I (n = 1399)	3	3.182	0.513	0.556	1.454	4	1	5
q8_6	II (n = 159)	3	3.195	0.521	0.484	0.734	3	2	5
q8_7	I (n = 1399)	3	3.197	0.545	0.300	0.757	4	1	5
q8_7	II (n = 159)	3	3.333	0.581	-0.007	-0.417	3	2	5
q8_8	I (n = 1399)	3	3.237	0.595	0.256	0.477	4	1	5
q8_8	II (n = 159)	3	3.314	0.575	0.059	-0.341	3	2	5
q8_9	I (n = 1399)	3	3.134	0.448	1.050	3.231	4	1	5
q8_9	II (n = 159)	3	3.170	0.409	1.759	2.952	3	2	5

Table 13. Criteria with standard deviations (SD) greater than 0.60 in each subsection

No.	*Subsection*	*Criterion*	*SD*	*SD Rank No*
1	Q1	q1_3	0.721	1
2	Q1	q1_6	0.655	10
3	Q1	q1_5	0.631	18
4	Q1	q1_4	0.604	32
5	Q2	q2_8	0.672	6
6	Q2	q2_2	0.661	8
7	Q2	q2_1	0.650	12
8	Q2	q2_7	0.650	13
9	Q2	q2_6	0.642	14
10	Q2	q2_4	0.608	27
11	Q2	q2_9	0.607	28
12	Q3	q3_2	0.656	9
13	Q3	q3_6	0.620	22
14	Q3	q3_1	0.616	24
15	Q4	q4_4	0.718	2
16	Q4	q4_5	0.637	15
17	Q4	q4_6	0.602	34
18	Q5	q5_5	0.687	4
19	Q5	q5_4	0.685	5
20	Q5	q5_6	0.670	7
21	Q5	q5_1	0.631	17
22	Q5	q5_3	0.628	19
23	Q5	q5_2	0.624	20
24	Q5	q5_9	0.605	31
25	Q6	q6_3	0.655	11
26	Q6	q6_4	0.605	29
27	Q7	q7_1	0.710	3
28	Q7	q7_3	0.620	23
29	Q7	q7_2	0.611	25
30	Q7	q7_4	0.611	26
31	Q8	q8_2	0.631	16
32	Q8	q8_3	0.622	21
33	Q8	q8_1	0.605	30
34	Q8	q8_4	0.602	33

We suggest that the four-point coloured scale should be defined as follows:

Unsatisfactory (Red-1): Behaviour demonstrates significant deficiencies, including safety-critical shortcomings, and fails to meet minimum performance expectations at occupational qualification standard. This category includes both explicitly unsafe behaviour and consistently inadequate performance requiring substantial development.

Acceptable (Amber-2): Behaviour meets the acceptable level at occupational qualification standard but might show clear limitations in consistency, quality, or situational adaptation. Performance is functional but requires targeted development.

Very Good (Light Green-3): Behaviour meets expected standards well with generally consistent and effective performance.

Exceptional (Dark Green-4): Behaviour consistently exceeds expected standards in quality, safety, and situational appropriateness. Performance demonstrates a high level of professional competence and robustness.