Advances in Human Factors of Transportation, Vol. 224, 2026, 408–416
https://doi.org/10.54941/ahfe1007869

Development of Human Factors Toolkit to Inform Behavioural and Cognitive Research in the Railway Domain

Sarah A. Kusumastuti[1], Tom H. J. Kolkman[1], Julia C. Lo[2], and Simone Borsci[1]

[1]Human Factors and Cognition Team, CODE Section, Learning and Data Analytics and Technology, University of Twente, Enschede 7521 PL, The Netherlands
[2]Innovation & Development, ProRail, Utrecht 3511 EP, The Netherlands

ABSTRACT

The application of automated systems in rail operations should account for its impact on train (traffic) driver. The introduction of automation is expected to transform the duties and responsibilities of operators, or more generally, the job description itself. The role of behavioural and cognitive research is essential in supporting the shift toward more advanced interactive technologies, helping to maximize their benefits while mitigating potential human factors issues in its adoption. We synthesised information from academic literature on human factors research in railways and developed a toolkit that identifies a set of human factors constructs that were commonly measured by practitioners in the railway domain, particularly in human-in-the-loop (HITL) simulations. We then conducted 3 rounds of feedback consisting of a workshop and 2 survey rounds with a total of 23 responses from human factors experts We identified 8 main constructs that are commonly measured in behavioural rail simulation studies: task performance, workload, communication, situation awareness, attention (including vigilance and attention allocation), user experience and usability, fatigue/sleepiness, and trust in automation. Additionally, we also compiled 63 objective and subjective methods for measuring the constructs, We provide descriptions and examples of how each method was utilised in research in the toolkit.

Keywords: Research methods, Cognitive systems, Performance measures, Psychological measurement, Human-in-the-loop simulation, Experimental research, Train (traffic) driver

INTRODUCTION

Behavioural and cognitive research is at the core of human factors research. It is an essential tool in testing hypothesis and assumptions about how we interact with interfaces and systems. In a complex system such as transport, the margin of error of a design that could not properly account for human factors is the difference between life and death (Read et al., 2017). For every technological shift arriving in the transport system, there is potential for changes in the way the system is interacting with operators, passengers, and its surroundings. One example is the transition of metro systems from Grade of Automation (GoA) level 1 to 2, which transforms the work of a train driver into a more supervisory role instead of an active operator, which has

Received February 6, 2026; Revised April 3, 2026; Accepted April 21, 2026; Available online July 20, 2026

consequences in the cognitive processes involved in the job such as decreased workload which may lead to errors induced by boredom or inattention (Hunter & McDermid, 2022).

There have been some examples of compiling human factors methods and research practices in the transport domain, for example in road transport, maritime transport (Wu et al., 2022) and aviation (Skybrary, n.d.). We aim to extend this body of knowledge by providing an inventory for human factors research methods and practices used in the railway domain that can help human factors practitioners to share a common methodological vision. While this toolkit is initially developed with a focus on human-in-the-loop simulation studies, the content of the toolkit are applicable to a wide variety of human factors studies that involve collecting behavioural or physiological data.

Our main goal is to map the key human factors measurements in the railway domain, specifically when rail operators are involved as subjects. This includes compiling the methods/tools utilized to empirically measure aspects of human factors. This paper describes the synthesis of the toolkit (Kusumastuti et al., 2025b) from data extraction through multiple iterations of expert review.

METHODS

The human factors toolkit is essentially a compilation of essential constructs measured in human factors studies in rail and automation as well as methods, instruments, and measurement techniques that were utilised in those studies.

The formation of the toolkit involves a literature review for the initial synthesis of the toolkit followed by multiple phases of expert reviews based on the Delphi method (Linstone & Turoff, 1975). There are three main phases in the production of the toolkit and each phase produces a version of the human factors toolkit that becomes a basis for review for the next phase. The review phases are (1) literature review phase (2) workshop and follow up survey phase (3) dissemination survey phase

Phase 1: Literature Review

As a part of a literature review on human factors research in rail and automation (Kusumastuti et al., 2025a) we identified human factors constructs that were measured in behavioural studies involving rail human factors. These constructs were identified and compiled by two independent reviewers with expertise in human factors and psychological measurement. This list of constructs (also referred to as aspects) were then reviewed by two additional senior experts in human factors to further adjust the clarity and accuracy of the content of the toolkit.

Phase 2: Workshop and Follow Up Survey

Human factors experts from the European research program for the rail sector, Europe's Rail, were involved in the initial expert review phase by

participating in a workshop discussing the validity of constructs compiled from the literature review and further input on how the toolkit could be developed. Following the workshop, a survey asking for further evaluation and feedback, as well as suggestions on how to enrich the toolkit, was sent to a group of experts consisting of both workshop participants and additional human factors experts. The responses from the feedback survey were anonymized.

We utilized a likert scale for the respondents to evaluate the relevance, usefulness, and sufficiency of information presented in the toolkit. The questions are as follows: (1) *Relevance*: To what extent do you agree that evaluating [aspect] is relevant in HITL simulations within the railway context? (1–7) (2) *Usefulness*: How would you rate the usefulness of the table and description we provided on [aspect]? (1–5) (3) *Sufficiency:* To what extent do you agree that we have listed all main methods for measuring [aspect]? (1–7) (4) *Terminology:* Do you have any suggestions for alternative or additional terminology or concepts related to [aspect]? (5) *Suggestions:* (a) Do you have suggestions on other methods that can be used to measure [aspect]? (If yes, Please explain below, and please ensure that you provide sufficient details and references for the measurement you are proposing, allowing us to incorporate it into the toolkit) (b) What changes would you like to suggest in order to improve the presentation and/or content of the information we presented on [aspect]?

Phase 3: Dissemination Survey

In the second review phase, we solicited evaluation and feedback from participants of a European rail human factors conference in February 2025. As part of a scheduled talk, we presented an overview of the 2nd version of our toolkit and provided a link to the full toolkit document, as well as a survey asking to evaluate the toolkit. Participants of the conference consist of human factors experts in transportation from academia and industry.

The questions presented on the survey are as follows: (1) *Suggestions:* Do you have any suggestions for additional constructs? If yes, please explain below and Do you have any suggestions for additional measurements for existing constructs? If yes, please explain below (2) *Usefulness:* How useful do you think is this human factors toolkit for HITL rail research? (Not at all useful [1] – extremely useful [5]) (3*) Utility*: Please indicate your agreement to the following statements (strongly disagree [1]) - strongly agree [5]): (a) the HITL toolkit can help harmonize the practice in rail research domain (b) the HITL toolkit can improve the replicability of research results (c) the HITL toolkit can improve the comparability of research results

RESULTS

A summary of the result of each phase is outlined in Table 1. The table oulines the number of participants involved in the development of the toolkit during the phase as well as the content of the toolkit

Table 1: Toolkit review participants and content at each phase.

Phase	Participants	Toolkit Content
Literature review	4 human factors experts (2 extractors and 2 reviewers)	10 constructs, 39 methods/ measurements
Workshop and follow up survey	8 workshop participants and 7 survey respondents from 12 survey recipients	8 constructs+ 2 sub constructs, 59 methods/measurements
Dissemination survey	4 complete responses from around 40 attendants of the presentation	8 constructs+2 sub constructs, 63 methods/measurements

Phase 1: Literature Review

An initial list of constructs and measure were extracted from a list of behavioural studies on human factors in railways. The list consists of 10 human factors related constructs and 39 methods used to measure them were extracted from the literature. The list of constructs and their definitions based on this initial phase is as follows:

1. *Task Performance*
2. *Workload*
3. *Communication*
4. *Situation Awareness*
5. *Attention Allocation*
6. *User experience and usability*
7. *Fatigue*
8. *Responsiveness*
9. *Vigilance*
10. *Trust in Automation*

The first version of the toolkit document that contains the construct definition and various objective and subjective methods to measure them.

Phase 2: Workshop and Follow Up Survey

Workshop

There were 8 experts who attended the workshop both in person and on-line as part as Europe's Rail project meeting. In the workshop, the list of constructs was validated, and it was agreed that the workshop participants will present more thorough feedback through a questionnaire after reviewing the full document.

The first version of the toolkit document and feedback questionnaire was sent to all participants of the workshop and an addition of 4 human factors and/or railway experts that did not attend to workshop but was part of the project. There were 7 full responses received from 12 surveys sent.

Survey Results

A summary of the results of the rating scale questions of the survey is as follows:

- *Relevance:* Performance is noted to be the most relevant with a mean of 6.9 out of 7 and responsiveness is evaluated to be the least relevant (5.3/7)
- *Usefulness:* The information on fatigue is judged to be most useful (4.2/5) and communication to be least useful (3/5).
- *Sufficiency:* The information on fatigue and attention allocation is judged to be most sufficient (both 5.1/7) and performance to be the least sufficient (4.3/7).

As a result of feedback from this phase, we made 3 major changes to the main construct list. First, the construct responsiveness is absorbed into task performance as it was agreed to be a measure of performance. Secondly, the constructs attention allocation and vigilance were merged into a single construct called attention where the two become subconstructs. Finally, the construct fatigue was renamed to fatigue & sleepiness. The list of constructs and their definitions based on this initial phase is as follows:

1. *Task Performance*
2. *User experience and usability*
3. *Situation Awareness*
4. *Workload*
5. *Fatigue and sleepiness*
6. *Attention (vigilance and attention allocation)*
7. *Communication*
8. *Trust in Automation*

Additionally, 20 more methods or measurements were added to the toolkit across all 8 constructs. A version of the toolkit consisting of 8 constructs and 59 measures was then published to be evaluated at the next review phase.

Phase 3: Dissemination Survey

The second version of the toolkit was presented as part of a talk at the 6th German Conference on Rail Human Factors in Berlin, Germany. During the talk, a link to the toolkit document produced from phase 2 and feedback survey was presented to all attendants of the talk and was available up until a week after the conference. In total 4 full responses was collected from the talk attendants.

A summary of the results of the rating scale questions of the survey are as follows:

- *Usefulness:* The average usefulness rating of the toolkit is judged to be 4 out of a scale of 5
- *Utility:* For statement (a) the mean response is 3/5, statement (b): 2.75/5, and statement (c): 4.25/5

Two measures were added to the toolkit based on the feedback: safety measures under task performance and stress management measures under workload. Additionally, workload and situation awareness now make distinction between individual and team measures. After the changes were made, a third and final version of the toolkit was published (Kusumastuti et al., 2025b). Table 2 outlines the list of constructs and measures compiled in the final version. The toolkit document provides a more detailed explanation on the measurement validity and how it has been used by other studies.

Table 2: List of constructs and measures included in the final version of toolkit.

Construct		**Methods of Measurements**	
		Objective	**Subjective**
Task Performance	Train Driver	• Takeover time • Speed maintenance • Acceleration variability • Braking errors • Driver to interface response time • Driver to emergency response time • Rate of intervention by Train Protection System • Signal failure detection	
	Traffic Controller	• Response Latency • Punctuality, Arrival/Depart Delay • Maintenance of free track order • Platform consistency	• Observational scoring system
	All	*Additional validated and used methods from other domains and applications:* Safety performance	
		Performance measures are usually incorporated in usability and UX assessments with a focus on efficiency and effectiveness	• General experience by interviews • User Preference by A/B Testing Scales • Questionnaire Measuring Subjective Consequences of Intuitive Use • Subjective Acceptance *Additional validated and used scales from other domains and applications* o System Usability Scale (SUS) o Software Usability Measurement Inventory (SUMI) o Expectation rating

(Continued)

Table 2: Continued.

Construct		Methods of Measurements	
		Objective	**Subjective**
			o Post-Study System Usability Questionnaire (PSSUQ) o Usability Metric for User Experience (UMUX)
Situation Awareness	Individual	• Eye movements Querying techniques • Situation Awareness Global Assessment Technique (SAGAT) *Additional validated and used methods from other domains and applications:* o Situation Presence Assessment Method (SPAM)	• Situation Awareness Rating Technique (SART) • Low-Event Task Subjective Situation Awareness (LETSSA) • Mission Awareness Rating Scale (MARS)
		Expert evaluation techniques and mixed methods	
	Team/ network		• Mission Awareness Rating Scale (MARS) for teams
		• Communication content through network analysis • Perceived coordination and communication	
Workload	Individual	• Heart rate variability • Eye movements	• Instantaneous Self Assessment (ISA) • Rating Scale Mental Effort (RSME) • NASA – Task Load Index (NASA-TLX) • DLR – Workload Assessment Tool (DLR-WAT) • Integrated Workload Scale (IWS)
		Additional validated and used methods from other domains and applications: Stress management measures	
	Team/ network		Communication flow
Fatigue and sleepiness		• Heart rate variability • Brain activity	• Visual analogue scale to evaluate fatigue severity (VAS-F) • Observation (Physical Characteristics) • Sleepiness scales - Karolinska Sleepiness Scale (KSS) - Stanford Sleepiness Scale (SSS)

(Continued)

Table 2: Continued.

Construct		**Methods of Measurements**	
		Objective	**Subjective**
Attention	Attention allocation	• Time in areas of interest by eye tracking	• Task related time allocation by observation
	Vigilance	• Psychomotor vigilance task (PVT) • Safety Critical Event Detection Task	• Mind-Wandering Scale (MWS)
Communication		• Observation, including: o Frequency of communication o Length of communication o Direction of communication o Voice tone	
Trust in automation		*Validated methods utilised in other domains and applications:* Eye movements (glance)	• Direct self-report *Validated methods utilised in other domains and applications:* o Dynamic reporting of trust o Trust in automation scale o Checklist for trust
		Other methods of measuring trust, including: • Level of monitoring automated system • Reaction to automated warning • Reaction to conflicting information	

CONCLUSION

This paper illustrates the development of a toolkit for human factors research in railways, consisting of relevant constructs that can be measured in behavioural studies involving rail workers, particularly those utilising HITL simulations. The toolkit compiles measures, methods, and techniques for human factors research that has gone through multiple validation stages by human factors and/or rail experts.

Compared to other domains like aviation and road transport, the railway sector still has a gap in terms of the availability of literature examining human factors constructs (Papadimitriou, 2020). We hope that this toolkit can function as a valuable inventory of knowledge and resources, aiding HF researchers and railway practitioners in understanding and designing applied cognitive and behavioural experiments in rail human factors.

ACKNOWLEDGMENT

This project has received funding from the Europe's Rail Joint Undertaking under the European Union's Horizon 2022 research and innovation program under grant agreement No. 101101973 (FP1 - MOTIONAL).

REFERENCES

Hunter, J., & McDermid, J. (2022). Investigating Human Error Within GoA-2 Metro Lines. *Reliability, Safety, and Security of Railway Systems. Modelling, Analysis, Verification, and Certification: 4th International Conference, RSSRail 2022, Paris, France, June 1–2, 2022, Proceedings*, 179–191. https://doi.org/10.1007/978-3-031-05814-1_13

Kusumastuti, S. A., Kolkman, T. H. J., Lo, J. C., & Borsci, S. (2025a). Charting the landscape of rail human factors and automation: A systematic scoping review. *Transportation Research Interdisciplinary Perspectives*, *30*, 101350. https://doi.org/10.1016/j.trip.2025.101350

Kusumastuti, S. A., Kolkman, T. H. J., Lo, J. C., & Borsci, S. (2025b). Human factors toolkit in railways V.3 (September 2025). *OSF*. https://osf.io/5rw24

Linstone, H. A., & Turoff, M. (Eds.). (1975). *The delphi method* (Vol. 1975, pp. 3–12). Reading, MA: Addison-Wesley.

Papadimitriou, E., Schneider, C., Tello, J. A., Damen, W., Vrouenraets, M. L., & Ten Broeke, A. (2020). Transport safety and human factors in the era of automation: What can transport modes learn from each other?. *Accident analysis & prevention*, *144*, 105656.

Read, G. J. M., Beanland, V., Lenné, M. G., Stanton, N. A., & Salmon, P. M. (2017). *Integrating Human Factors Methods and Systems Thinking for Transport Analysis and Design*. CRC Press.

Skybrary. (n.d.) *Human Factors Methods*. Retrieved February 2, 2026, from https://skybrary.aero/human-factors-methods

Wu, B., Yip, T. L., Yan, X., & Guedes Soares, C. (2022). Review of techniques and challenges of human and organizational factors analysis in maritime transportation. *Reliability Engineering & System Safety*, *219*, 108249. https://doi.org/10.1016/j.ress.2021.108249