Cognition, Technology & Work (2022) 24:483–508
https://doi.org/10.1007/s10111-022-00701-7

ORIGINAL ARTICLE

Assessing the reliability and validity of an FRAM model: the case of driving in an overtaking scenario

Niklas Grabbe[1] · Almin Arifagic[1] · Klaus Bengler[1]

Received: 21 February 2022 / Accepted: 2 May 2022 / Published online: 26 May 2022

Abstract

Over the past two decades, systemic-based risk assessment methods have garnered more attention, and their use and popularity are growing. In particular, the functional resonance analysis method (FRAM) is one of the most widely used systemic methods for risk assessment and accident analysis. FRAM has been progressively evolved since its starting point and is considered to be the most recent and promising step in understanding socio-technical systems. However, there is currently a lack of any formal testing of the reliability and validity of FRAM, something which applies to Human Factors and Ergonomics research as a whole, where validation is both a particularly challenging issue and an ongoing concern. Therefore, this paper aims to define a more formal approach to achieving and demonstrating the reliability and validity of an FRAM model, as well as to apply this formal approach partly to an existing FRAM model so as to prove its validity. At the same time, it hopes to evaluate the general applicability of this approach to potentially improve the performance and value of the FRAM method. Thus, a formal approach was derived by transferring both the general understanding and definitions of reliability and validity as well as concrete methods and techniques to the concept of FRAM. Consequently, predictive validity, which is the highest maxim of validation, was assessed for a specific FRAM model in a driving simulator study using the signal detection theory. The results showed that the predictive validity of the FRAM model is limited and a generalisation with changing system conditions is impossible without some adaptations of the model. The applicability of the approach is diminished because of several methodological limitations. Therefore, the reliability and validity framework can be utilised to calibrate rather than validate an FRAM model.

Keywords FRAM · Validation · Driving · Overtaking manoeuvre

1 Introduction

Risk assessment is a crucial aspect of Human Factors and Ergonomics (HFE) research. Instead of the reactive approach taken in accident analyses, which looks at a particular erroneous scenario, risk assessment adopts a proactive approach, trying to identify hazards or looking for what could happen in the future to prevent or mitigate adverse events or to facilitate desirable outcomes. Over the past 20 years, systemic based risk assessment methods have garnered more attention and their use and popularity are growing (e.g., Dallat et al. 2017; Hollnagel 2012; Hughes et al. 2015; Hulme et al. 2019; Larsson et al. 2010; Leveson 2004; Salmon et al. 2012). These methods try to describe performance at the level of the overall system and see the accident process as a complex and interwoven event that cannot be broken down into its individual parts. Emerging events caused by complex and non-linear interactions between the various system parts can affect the performance of the system and cause an accident (Laaraj and Jawab 2018; Qureshi 2007; Wienen et al. 2017). In general, systemic models acknowledge the complexity and socio-technical nature of systems, and further emphasise the need for an understanding of the functional abstraction of the system, rather than structural decomposition (Rasmussen 1997).

In particular, the functional resonance analysis method (FRAM) (Hollnagel 2012) is one of the most widely used systemic methods for risk assessment and accident analysis. It allows the modelling of mechanisms within complex socio-technical systems (STS), including their interfaces between humans and technology, coupling and dependency

✉ Niklas Grabbe
n.grabbe@tum.de

[1] Chair of Ergonomics, Technical University of Munich, Boltzmannstr. 15, 85748 Garching, Germany

effects, nonlinear interactions between elements, and functional variability (Woltjer and Hollnagel 2008). In general, the results of an FRAM analysis contribute to an understanding of real work and reveal unsafe functional interactions within one agent and between different agents; these are needed to assist risk management as regards the proactive assessment of technological changes and their impacts (Ferreira and Cañas 2019; Patriarca and Bergström 2017). In addition, FRAM should form the basis for systemic risk assessments in complex STS, for example for contemporary applications, such as automated driving in road traffic (Grabbe et al. 2020, 2022). These authors do so by providing a useful understanding of the actual system mechanisms and interactions that are needed to assist the system design, enhanced by considering non-linear, complex, and emergent system behaviour (Grabbe et al. 2022). In the past, FRAM has been widely used and enhanced methodologically in a variety of domains for retrospective as well as prospective analyses, as detailed in a comprehensive review by Patriarca et al. (2020). Hence, FRAM has been progressively evolved since its starting point in 2004 (Hollnagel 2004) and is considered to be the most recent and promising step in understanding STS (Nemeth 2013).

However, there is currently a lack of any formal testing of the reliability and validity of FRAM. This applies to the HFE research as a whole, where validation is both a particularly challenging issue and an ongoing concern (Stanton and Young 1999a, 2003; Stanton 2016). In fact, Stanton and Young (1999a) stated that practitioners often assume validity, but seldom test and prove it empirically. Furthermore, methods are often chosen by practitioners that are based on familiarity and ease of use rather than on reliability and validity evidence (Stanton et al. 2013). Thus, findings from the application of HFE methods suffer from an objective evaluation, making the research findings questionable. However, HFE methods must prove that these methods can intentionally work in their applied domains (Stanton 2014) and to promote the credibility of HFE methods and their whole community (Stanton 2016). In this context, and since FRAM should form the basis for systemic risk assessments in complex STS (Grabbe et al. 2020, 2022), validation is an absolute priority and a compulsory aspect in engineering disciplines (where HFE is part of it), especially in the aforementioned field of automated driving, due to the enormous societal impact which benefits FRAM by providing a clear evaluation of its performance and value.

Thus, this paper aims to first define a more formal approach to achieving and demonstrating the reliability and validity of an FRAM model that forms the basis for risk identification and design recommendations within the FRAM method, and second, to apply this formal approach partly to an existing FRAM model so as to prove its validity, and to evaluate the general applicability of this approach.

The remainder of this paper is structured as follows. Section 1.1 summarises the theoretical foundations and individual analytical steps of FRAM, as well as previous validation approaches. Following on from this, Sect. 1.2 summarises approaches for testing the reliability and validity of HFE methods. Section 2 outlines the understanding and definitions of reliability and validity in literature and transfers these to the context of FRAM to define a framework that addresses the reliability and validity of FRAM models. In Sect. 3, we describe the methodology for the evaluation of predictive validity in a driving simulator experiment. Section 4 presents the results, including the evaluation of the predictive validity of the analysed FRAM model according to the three different research questions of the study. Section 5 then discusses the results with respect to the research goals of this paper and also outlines methodological limitations. Finally, a brief conclusion and outlook for future research are provided in Sect. 6.

1.1 Basics of FRAM and previous validation approaches

The purpose of the model produced by the FRAM method is to describe and understand what is happening in an STS in terms of functions rather than components. An FRAM model focuses on adjustments to everyday performance, which usually contribute to things going right. In rare cases, these performance adjustments aggregate in unexpected ways, leading to functional resonance, with accidents being the most extreme result.

FRAM relies on four principles (the equivalence of success and failures, approximate adjustments, emergence, and functional resonance), and follows four steps (modelling the system by identifying its functions, identifying the function's performance variability, aggregating the variability, and managing the variability), as detailed in Hollnagel (2012). The steps are briefly described in the following. In the first step, the essential functions of a system are identified to build a model. Basically, each function is characterised by six aspects (i.e., input, output, precondition, resource, control, and time), which couple each function with several other functions representing a specific instantiation of the model that traditionally is represented graphically by hexagons. Furthermore, the functions can be divided into two classes: foreground and background functions. Foreground functions are the core of the analysis and may vary significantly during an instantiation of the model. In contrast, background functions are stable and represent common conditions as a system boundary that are used by foreground functions. The second step is to specify the performance variability of each function that can be characterised in its simple form using two phenotypes, namely, timing and precision. Here, the function's output in terms of timing can occur too early, on time, too late, or not at all, whereas for precision, the output can be precise, acceptable, or imprecise (Hollnagel 2012). In the third step, the variability is aggregated to understand how the variability can propagate through the system and where functional resonance emerges leading to adverse outcomes. This is done by defining upstream–downstream couplings, where variability can be caused through couplings of upstream functions, when the output used as input or resource, for example, is variable and thus affects the variability of downstream functions. The fourth and final step consists of the monitoring and management of the previously identified performance variability to ensure the safety and performance of the system.

In the past, some attempts were made to formally verify an FRAM model. The first attempt at formal verification was the FRAM model-based safety assessment that used model checking and theorem proving to verify the FRAM model so as to determine whether pre-set safety requirements can be observed (Yang and Tian 2015). The same authors enhanced this approach using the Simple Promela Interpreter (SPIN) tool and applied it to develop an air traffic management system. The analysis demonstrated that FRAM can benefit from a formal verification with the aid of model checking through more rigorous computation that improves its efficiency and accuracy (Yang et al. 2017). In addition, the software tool FRAM Model Interpreter (FMI) (Hollnagel 2020) has recently become available, which is a stepwise automatic interpretation of the syntactical and logical correctness of an FRAM model to formally check and adjust its consistency and completeness. With regard to validation, subjective evaluation through interviews with experts, workshops, and discussions was mainly used to improve the face validity of developed FRAM models, as pointed out by Bridges et al. (2018), Kaya et al. (2019), and Ross et al. (2018). The reason may be associated with an experts' deep knowledge of the work system and daily operations, which can help to enrich developed FRAM models and to provide more reliable models (Salehi et al. 2021). However, a more formal approach for validation is still lacking.

1.2 Previous approaches to testing the reliability and validity of HFE methods

On the whole, studies are rarely conducted that report the reliability or validity of HFE methods. However, some examples can be found and are summarised in the following. The reliability of ergonomics methods is often assessed using a test–retest paradigm (Baysari et al. 2011). Examples of the measures used here include percentage agreement (Baber and Stanton 1996; Baysari et al. 2011; O'Connor 2008), Pearson's correlation (Harris et al. 2005; Stanton and Young 2003), the index of concordance (e.g., Olsen and Shorrock 2010), and Cohen's kappa (e.g., Makeham et al. 2008).

Studies assessing the validity of ergonomics methods can also be found in literature (Baber and Stanton 1996; Stanton et al. 2009; Stanton and Young 2003). Many of these have focussed on human reliability and error prediction methods in general (Baysari et al. 2011; Kirwan et al. 1997; Stanton and Young 2003) or more specifically on the systematic human error reduction and prediction approach (SHERPA) (Stanton and Stevenage 1998) and task analysis for error identification (TAFEI) (Stanton and Baber 2005). In these studies, the validity of methods was assessed by comparing a method's results (e.g., errors predicted) against actual observations (e.g., errors observed). More recently, system analysis methods, such as the cognitive work analysis (Cornelissen et al. 2014), a factor classification scheme for Rasmussen's Accimap (Goode et al. 2017), the networked hazard analysis and risk management system (Net-HARMS) (Hulme et al. 2021a), and the operator event sequence diagrams (Stanton et al. 2021a, b, c, d) have also been empirically validated. Furthermore, there has been a thorough comparison of intra-rater reliability and criterion-referenced concurrent validity between three systems-based risk assessment approaches: the systems-theoretic process analysis (STPA) method, the event analysis of systemic teamwork broken links (EAST-BL) method, and the Net-HARMS method (Hulme et al. 2021b; see also Hulme et al. 2021c). In general, quantitative methods to compare expert results versus novice results (or predicted versus actual outcomes) are often based on the use of signal detection theory (SDT) to calculate the sensitivity of the method under analysis (Baber and Stanton 1994; Stanton et al. 2009; Stanton and Young 2003). The SDT and its metrics are commonly used to assess the reliability and validity of ergonomics methods, such as human error prediction (Stanton et al. 2009). This was pioneered in particular by Stanton and Young (1999a, b) as a means of establishing empirical validity of methods.

A comparison of the reliability and validity of a range of HFE methods has been undertaken by Stanton and Young (1999a, b, 2003), which showed that the methods vary quite considerably in their performance. This demonstrates the urgent need for more reliability and validation studies of other HFE methods, and in particular FRAM. Moreover, FRAM follows a safety-II perspective (Hollnagel 2014) for which validity is seldomly addressed instead of safety-I (Hollnagel 2014) based methods as, e.g., human error analysis methods as mentioned previously.

2 Proposed reliability and validity framework

2.1 Understanding and definitions of reliability and validity

According to Stanton and Young (1999a), reliability and validity are interrelated, where a method can only be valid if it is reliable but may be reliable and not valid. Thus, these two criteria have to be evaluated mutually.

Reliability is a measure of the stability of the method over time and across analysts, ideally demonstrating that the application of an ergonomics method will result in the same results if it is used by different people (inter-rater) or at different points in time by the same people (intra-rater) (Stanton et al. 2016). This is often assessed using a test–retest paradigm between experts and novices, including measures, such as percentage agreement and Cohen's Kappa (e.g., Baysari et al. 2011; Hulme et al. 2021a; Makeham et al. 2008).

When considering validity, we have to distinguish between the following two main terms: verification and validation. According to Balci (1998), verification determines whether the formal implementation of a model is correct, which deals with building the model correctly. On the other hand, validation determines whether a model can be substituted for the real system for the intended purposes and objectives in the applied domain, which deals with building the right model. Overall, a model must be useful with regard to its objective, which means providing a reasonably accurate answer to the question to be answered (Liebl 2018, p. 203). Consequently, the concept of validity has to be guided by this requirement and should not be regarded as absolute (Schrank and Holt 1967). This has various implications for the nature of validation (Liebl 2018, pp. 203–205):

- *model-individual,* meaning that it is impossible to postulate a standardised validation procedure due to various forms and applications of models. Rather, the required validity criteria and their weighting change depending on the problem (Banks et al. 1987).
- *gradual,* showing how good or bad a model is in fulfilling its purpose and describing the validation process as a trade-off between additional costs/effort and the added information value of increased validity (Van Horn 1971).
- *result of a negotiation process,* according to which the validity of a model largely equates to the question of credibility and acceptance. Within this process, it is negotiated when the model is considered sufficiently valid and which validity criteria and methods should be applied (cf. Sargent 1984).
- *continuous and iterative,* meaning that validation takes place during the entire development process and "confidence is built into the model as the study proceeds" (Bulgren 1982, p. 126) rather than depicting a separate section at the end as an end state.

Furthermore, different categories of validation can be found in literature. For instance, Liebl (2018) distinguishes between outcome-based, function-based, and theory-based validation. Outcome-based validation aims to compare results, checking the extent to which the model produces results that match those of the real system. Function-based validation comes into play when the real system is not fully observable, so one has to validate exclusively on the model itself. Here, the reaction mode of the model is checked for plausibility, hence validity ultimately presents itself as a failed falsification of the model (Hanssmann 2018, p. 93). Theory-based validation compares the model results with theoretically expected results, which usually come from analytical models or literature.

As for HFE methods, Stanton and Young (1999b) proposed four types of validity for ergonomics methods: construct, content, concurrent and predictive. Construct validity concerns the underlying theoretical basis of a method. Content validity relates to the credibility that a method can achieve with its users, which can also be referred to as face validity. Finally, concurrent and predictive validity address the extent to which an analysed performance is representative of the performance that might have been analysed, where concurrent validity describes the current performance sampled, and predictive validity (i.e., criterion-referenced empirical validity) concerns the performance in the future. Furthermore, HFE methods should possess a certain level of concurrent or predictive validity suitable for their application (Stanton 2016). However, it is debatable as to whether all ergonomics methods have to fulfil all four types of validation, as shown by a distinction between analytic and evaluative methods, assuming that construct and content validity might be sufficient for analytic methods, whereas predictive validity might be required for evaluative methods (Annett 2002).

Finally, various concrete techniques can be used to test the aforementioned validation and verification types. Balci (1998, p. 355) presented an overview of more than 75 techniques, placing them into four categories: informal, static, dynamic, and formal. The use of mathematical and logic formalism by the techniques increases from informal to formal. Informal techniques are the most commonly used and rely heavily on subjectivity. Examples include audits, face validation, turing tests, and walkthroughs. Static techniques assess the model's accuracy based on the characteristics of the static model design, including, for example, control analysis, semantic and syntax analysis as well as structural analysis. Dynamic models, on the other hand, evaluate the model based on its execution behaviour, including, among others, predictive validation, sensitivity analysis, and statistical techniques. Last but not least, the formal techniques are quite objective and are based on a mathematical proof of correctness, for instance, induction and logical deduction.

2.2 Transfer and applicability to FRAM

As we have seen before, validity is not an absolute concept, but rather a relative one. Thus, there is no standard approach to validity. Instead, an approach to prove validity and reliability has to be developed for each method itself according to the features and context of the application. Therefore, the aforementioned knowledge will be transferred to the concept of FRAM to define one potential approach to demonstrate reliability and validity for an FRAM model in the following (see Fig. 1). It should be pointed out that we have to distinguish between an FRAM model and a particular instantiation of the model when trying to define a validation approach for the FRAM method. According to Hollnagel (2012), the functions are potentially coupled in an FRAM model, meaning that there is no predetermined a priori order or fixed sequence of the functions, whereby the functions actually become coupled in an instantiation for a specific set of conditions, resulting in temporal and causal relations. Against this background, validation is only possible for a particular instantiation of an FRAM model, but not for an FRAM model in general. For the sake of simplicity, we use the term “FRAM model” as meaning an “instantiation of an FRAM model” in this paper.

Basically, FRAM is a qualitative modelling method that offers great flexibility in terms of how it is applied and used, since it is a method-sine-model which means that FRAM is used as a method to produce a model and not vice versa (Hollnagel 2012, pp. 127–133). In addition, an experienced team of experts is required to analyse and model the system (Accou and Reniers 2019; Jensen and Aven 2018; Pereira 2013), where the quality of the output in FRAM directly depends on the team of experts and the information they provide as input for the functions and their variability (Salehi et al. 2021). Although some practical guidance material exists in Hollnagel et al. (2014), there is no explicit standard for determining how much information should be included in the analytical process to define the objective, scope, and granularity of the model, as highlighted by Anvarifar et al. (2017), Grabbe et al. (2020), Li et al. (2019), and Patriarca et al. (2017). Due to these low limitations or regulations regarding modelling, as well as the strong dependency between model outputs and the competence of the modeller team, an FRAM model is ultimately subject to a very strong subjective component. This means that when applied to the same work context and using the method traditionally, an FRAM model and its risk derivation are unlikely to be congruent between different users and even with the same user on a different occasion. For this reason, the classic test–retest

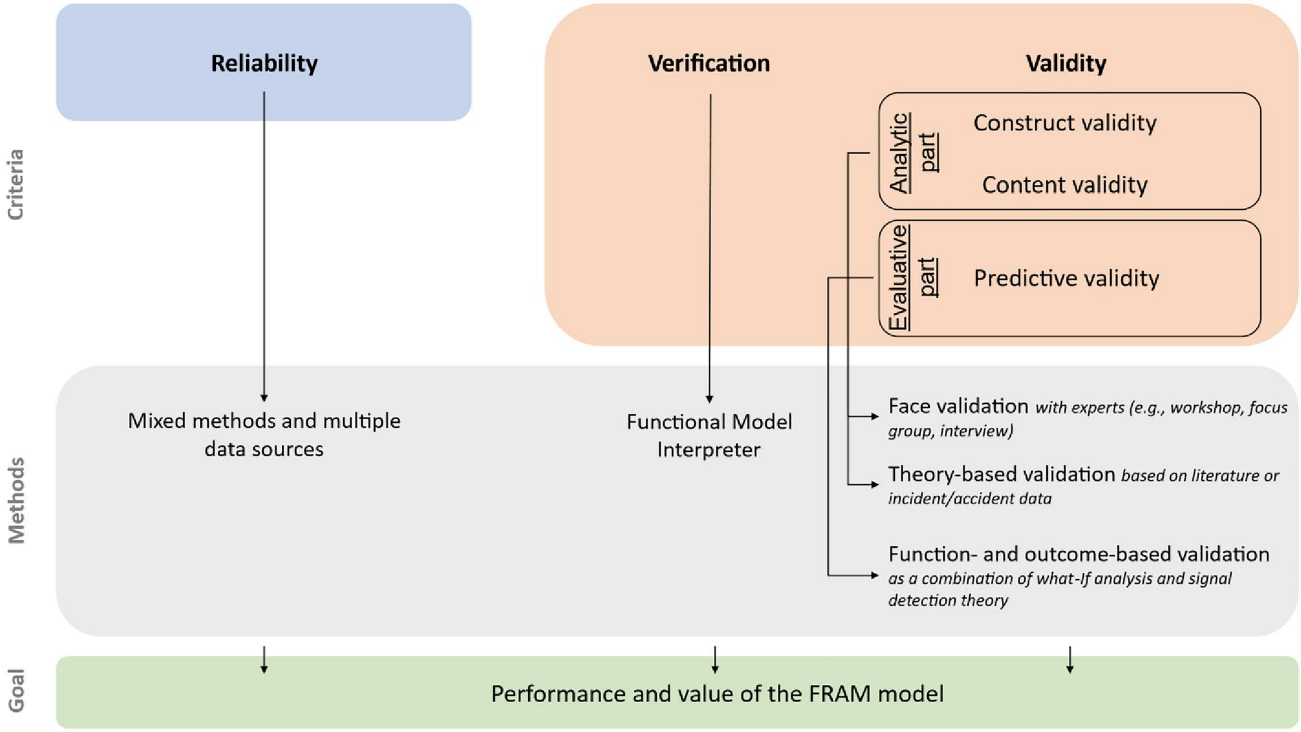

Criteria
Methods
Goal
Reliability
Verification
Validity
Analytic part
Construct validity
Content validity
Evaluative part
Predictive validity
Mixed methods and multiple data sources
Functional Model Interpreter
Face validation with experts (e.g., workshop, focus group, interview)
Theory-based validation based on literature or incident/accident data
Function- and outcome-based validation as a combination of what-If analysis and signal detection theory
Performance and value of the FRAM model

Fig. 1 Validation approach for an FRAM model

paradigm, which is often used to assess the reliability of HFE methods as mentioned in Sect. 2.1, seems inappropriate in the context of FRAM, particularly for largely complex FRAM models. Therefore, the reliability for an FRAM model cannot be proven but it can be achieved and increased using mixed methods and multiple data sources, such as document reviews, interviews, observations and simulations, as well as workshops and focus groups (see Fig. 1). These help to integrate multiple limited perspectives and dimensions that adhere to the verification strategies of Creswell and Miller (2000), complying with the four qualitative terms of credibility, transferability, dependability, and confirmability (Anfara et al. 2002) to improve the quality, scientific rigour, and trustworthiness of the model. One application of this can be found in Adriaensen et al. (2019) and Grabbe et al. (2022) in the context of aircraft cockpits and automated driving, respectively.

In addition, an FRAM model can be simply verified through the already established FMI (Hollnagel 2020) software that automatically interprets and parses the syntactical and logical correctness of an FRAM model step-by-step to formally check and adjust its structure with regard to consistency and completeness while obeying the FRAM “rules” (see Fig. 1). An important part of this is the identification of orphans or potential auto-loops as well as the question of whether relations between functions are mutually consistent, thus allowing an event to develop as intended. The use of FMI can be enhanced by other tools, such as model checking and theorem proving, as described in Sect. 1.1.

Validity should be divided into construct, content, and predictive validity according to Stanton and Young (1999b), since FRAM is an ergonomics method. In this case, concurrent validity is omitted, because an FRAM model does not generate absolute outputs; the outputs can only be evaluated relatively if something is changed in the model. This means that only future performance and not current performance can be validated. However, this is not a problem, since predictive validity is the higher maxim of the two anyway. Furthermore, FRAM is both an analytic and evaluative method. The analytic part is used through the qualitative and traditional application to gain an understanding of the mechanisms that underlie the functional interactions between system elements by modelling to comprehend what is happening, for example, to facilitate design decisions or to identify sources of performance failures and successes. In contrast to this, the evaluative part is used more in a semi-quantitative approach to measure and predict a certain parameter, such as performance variability, which is the fundamental factor explaining system behaviour in the FRAM method with its core principles of performance adjustments, emergence, and functional resonance. The analytic and evaluative parts are covered by construct and content validity, and predictive validity, respectively (cf. Annett 2002) (see Fig. 1). Construct validity should be ensured through the strong and sound systems theory basis of FRAM, as well as the tremendous credibility that the method gained amongst users over the last decade (cf. Patriarca et al 2020), which is also an argument for the content validity. Thus, construct validity can be generally assumed for an FRAM model as long as the method and its principles were correctly and comprehensively used, once again emphasising the strong dependency between an FRAM model’s output quality and the experience and training of the user and modeller as mentioned above. Content validity can mainly be proved by face validity using subjective evaluation through interviews, workshops, and discussions with experts who have a deep knowledge of normal work systems and daily operations, as already applied by Bridges et al. (2018), Kaya et al. (2019), and Ross et al. (2018). In addition, a theory-based validation could be used to further increase the content validity by comparing the FRAM model’s outputs with both other models or indicators in literature or incident and accident reports (including contributory factors and reasons) regarding the same application context. For instance, Bridges et al. (2018) modelled real accidents as "Mini FRAMs" based on accident reports that served as a comparison for the logic of the overall FRAM model.

Finally, predictive validity could be demonstrated by a mixture of function- and outcome-based validation. The reason for the combination is that an outcome-based validation alone is not possible, because an FRAM model does not generate absolute, observable outputs as a final product of the entire model. Instead, it must be linked to a function-based validation to produce relative, observable outputs through controlled variations in the model. The function-based validation can be realised by a sensitivity analysis with deliberate and controlled variations in the model to evaluate responses in the model for plausibility, which can also be called a “structured what-if analysis” (SWI-FRAM) (cf. Hill et al. 2020; MacKinnon et al. 2021). Here, one upstream function will be manipulated to vary its output to understand its impact on the system as well as how this variability can propagate through the system. In terms of predictive validation, this can be used to check whether the variation in the output of the upstream function actually influences the output of the coupled downstream functions while keeping all other functions constant at the same time. This process must be carried out for all direct upstream–downstream couplings of foreground functions in an FRAM model to fully test its predictive validity. This is exemplified in Fig. 2 and Table 1, which will be described in the following. Function A, highlighted in green, is initially manipulated to test the couplings AB, AC, and AD and to see if these couplings actually lead to a change in the output of functions B, C, and D. In the next steps, this procedure is also carried out for the other upstream–downstream couplings of the remaining

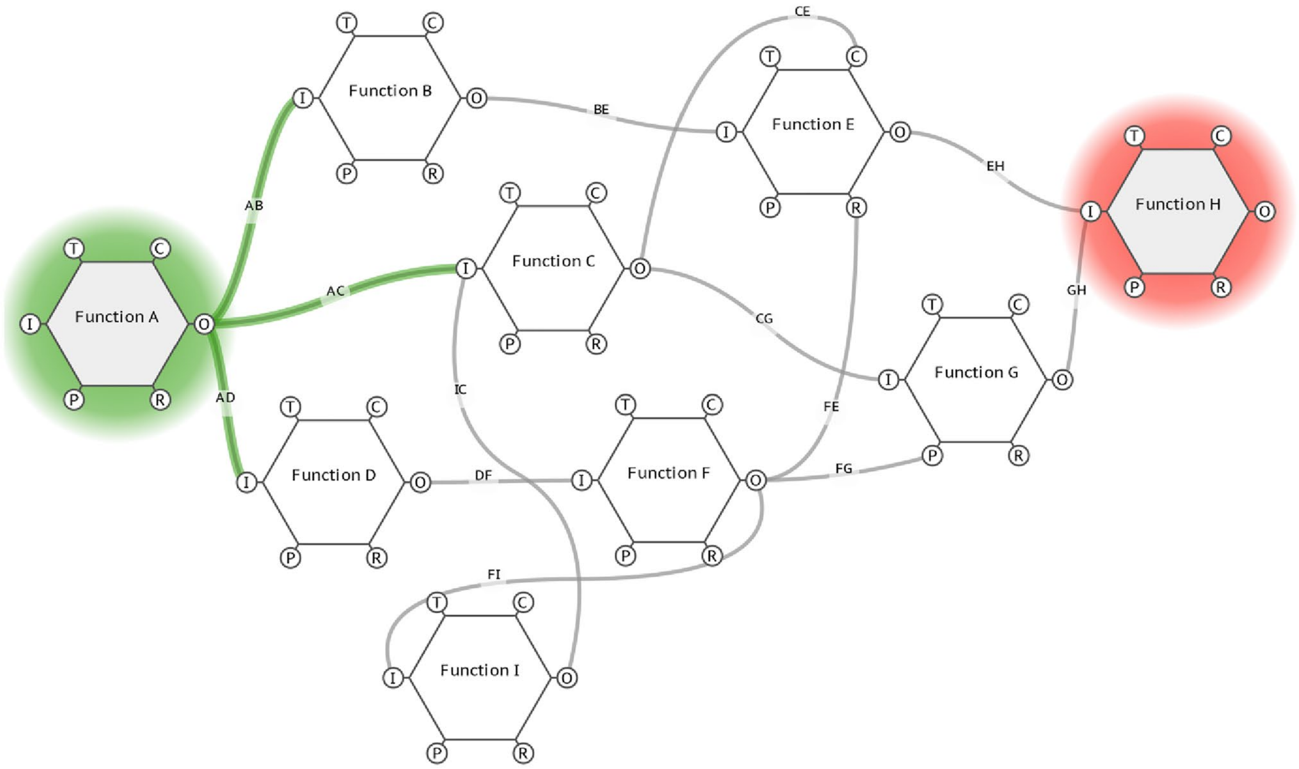


Fig. 2 Fictitious instantiation of an FRAM model with nine functions and thirteen couplings marked through letters. Function A is the start function and function H is the end function, as highlighted in green and red, respectively

Table 1 Assignment of upstream functions, downstream function, and their related couplings with regard to the fictive FRAM model in Fig. 2

Upstream function	Couplings	Downstream functions
A->	AB	-> B
	AC	-> C
	AD	-> D
B->	BE	-> E
C->	CE	->E
	CG	-> G
D->	DF	-> F
E->	EH	-> H
F->	FE	-> E
	FG	-> G
	FI	-> I
G->	GH	-> H
I->	IC	-> C

functions (see Table 1) up to the end function H, highlighted in red. The proof of all direct couplings is sufficient as this automatically explains the indirect effects too, for example, if function A has a direct impact on function D and D in turn on function F, then A also has an indirect effect on F. If the expected effect for one coupling can actually be confirmed, it is valid and if not, the coupling is not relevant and, therefore, invalid. However, this only validates the predictive performance for one instantiation and thus for one specific scenario, which does not mean that the model will be generally valid or invalid for other situations.

The final comparison between expected and actual effect then corresponds to an outcome-based validation, where the predictions of the FRAM model are matched with actual observations in reality. This is where the SDT comes into play, which was pioneered by Stanton and Young (1999a, b) to establish the empirical validity of ergonomics methods as mentioned in Sect. 1.2. This technique divides the method's outputs up into hits (H), misses (M), false alarms (FA), and correct rejections (CR). In the context of FRAM, it provides a method to compare the predicted variability effect of an upstream function to its coupled downstream functions, illustrated through the FRAM model, with the actual observed variability effect in simulator or field tests. In this work, the four events in Fig. 3 are defined as follows:

- *Hits*: predicted variability effect in a downstream function's output through the manipulation of its upstream

Observed (e.g., simulator or field test)

Predicted (FRAM model)

	Yes	No
Yes	H	FA
No	M	CR

Fig. 3 Signal detection theory (SDT) matrix

function's output by the FRAM model and observed variability effect in a simulator or field test.

- *Misses*: no predicted variability effect in a downstream function's output through the manipulation of its upstream function's output by the FRAM model, but observed variability effect in a simulator or field test.
- *False alarms*: predicted variability effect in a downstream function's output through the manipulation of its upstream function's output by the FRAM model, but no observed variability effect in a simulator or field test.
- *Correct rejections*: no predicted variability effect in a downstream function's output through the manipulation of its upstream function's output by the FRAM model and no observed variability effect in a simulator or field test.

In the following, the four events mentioned above will be explained using examples with the fictive FRAM model in Fig. 2. For instance, we will manipulate the output of function C to test the predictive validity. Potential hits or false alarms could be the couplings CE and CG to its direct downstream functions E and G, with one potential result being that coupling CE is a hit and CG a false alarm. The couplings EH and GH are indirect downstream effects of function C to function H and, therefore, out of the scope as we only measure direct downstream effects. Potential misses or correct rejections could be all the remaining functions that are not indirectly influenced by function C, and where no direct downstream couplings currently exist with function C and thus no variability effects are expected. It has to be proven whether the manipulation of function C has a variability effect on the outputs of the functions A, B, D, F, and I. Potential results could be that the "potential" coupling to function B is a miss and the potential couplings to the functions A, D, F, and I are correct rejections. Several metrics comprising the four events can now be used for the subsequent and concrete evaluation of predictive validity, which will be explained in more detail in Sect. 3.6.

All of the methods described above to demonstrate or increase the reliability, verification, and validity either influence or improve the performance and value of an FRAM model to increase the objective evaluation of research findings by FRAM as depicted in Fig. 1. In the next step, the process of predictive validity will be exemplified through an FRAM model for human and automated driving by Grabbe et al. (2022) to show its credibility as well as the applicability of the previously described predictive validation approach. This is because first, predictive validity represents the highest maxim of validation, and second, reliability, verification, and content validity for the analytical part of the validation have already been implemented by Grabbe et al. (2022) for the model to be examined. Therefore, the evaluative part of the validation is still open and thus addressed in the methods section.

3 Methods

3.1 FRAM model

The FRAM model to be validated in this paper is the FRAM model for overtaking in road traffic created by Grabbe et al. (2022). This model is very large and detailed, comprising 285 functions and including 210 foreground functions, all of which theoretically have to be analysed individually to test the predictive validity of the entire model. This is practically impossible and would go beyond the scope of this work. We, therefore, selected the two functions '*driving free*' (lead vehicle, LV) and '*driving free*' (oncoming vehicle, OV) to demonstrate the predictive validity. Both functions have a major impact on the system or rather the model and basically represent the longitudinal and lateral driving behaviour of LV and OV. The two functions and their couplings as well as their context will be described in more detail in Sects. 3.5 and 3.7.1, and 3.4.2, respectively.

3.2 Research questions

The analysis of the predictive validity of the FRAM model by Grabbe et al. (2022) pursues three research questions:

1) Is the model predictively valid for the basic scenario?

2) Is the model predictively valid for changing environmental conditions?
3) Is the model predictively valid for changing human factors conditions?

3.3 Sample

Forty German participants with valid driving licences took part in this experiment. This sample was divided into two subgroups with twenty participants each for the between-subjects factor levels of time pressure or no time pressure. The mean (*M*) age of the time pressure group was 29.4 years (SD = 14.5 years) with a range from 19 to 75 years, and that of the no time pressure group was 31 years (SD = 14.2 years) ranging from 21 to 72 years. The time pressure group consisted of twelve (60%) men and eight (40%) women, while the no time pressure group consisted of eleven (55%) men and nine (45%) women. In addition, Table 2 gives an overview of a comparison between the no time pressure and time pressure group as regards driving experience and driving style. Based on this data, the two samples can be considered as comparable.

3.4 Apparatus

3.4.1 Driving simulator

The experiment was carried out in the static driving simulator of the Chair of Ergonomics at the Technical University of Munich (see Fig. 4). The simulator consisted of a BMW E64 vehicle mock-up. A high-quality, 6-channel projection system provided a realistic driving environment. Three projectors were used for the front and back view each. The front field of view is approx. 180°. The back view through the mirrors is realised through three separate canvases. SILAB 6.5 of the Würzburg Institute for Traffic Sciences GmbH, with a refresh rate of 60 Hz, was used as the driving simulation software. An additional sound system provided vehicle and environmental sounds.

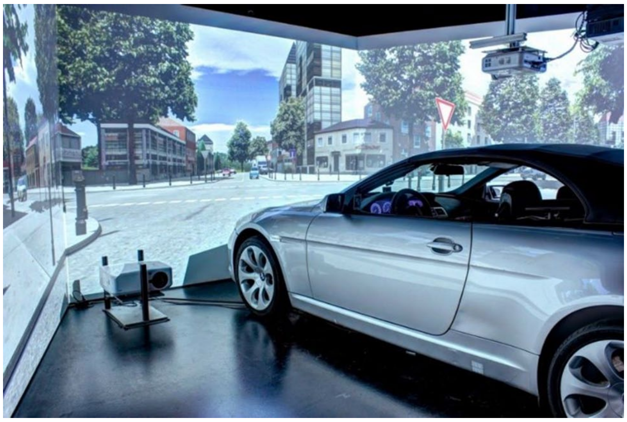


Fig. 4 Static driving simulator

3.4.2 Scenario and experimental track

The scenario of the analysed FRAM model was an overtaking manoeuvre on a rural road as detailed in Grabbe et al. (2022). An ego vehicle (EV) driven by the participant wants to overtake an LV travelling at a speed of 80 km/h on a straight rural road for a distance of 2500 m with no vertical elevation. The maximum speed limit is 100 km/h, overtaking is permitted and no obstructions exist. A rear vehicle (RV) is following the EV, and a line of cars are approaching on the oncoming lane at 100 km/h with different fixed time gaps. There were a total of ten gaps on the straight, with the first four time gaps being 10 s and for the last six gaps 12 s, corresponding to critical and uncritical time gaps according to the mean of 11.5 s (Crawford 1963) and median of 9.9 s (Tapio 2003) found in literature regarding accepted gaps when overtaking passenger cars. The road is 6 m wide, with one lane in each direction and a dotted line in the middle. The road is well constructed and all necessary road markings are in place. There is light vegetation on the side of the road. The weather conditions are sunny and dry. All simulation-controlled vehicles, which are passenger cars, always keep the necessary safety distance to their vehicle in front and comply with the traffic regulations. Before the actual test scenario,

Table 2 Comparison between the no time pressure and time pressure group regarding driving experience and driving style

Measurement	No time pressure group	Time pressure group
Participation in driving simulator studies	$M = 7.7$ (SD = 8.5)	$M = 10.3$ (SD = 24.3)
Mileage [km/year]	$M = 12{,}272$ (SD = 5,054)	$M = 12{,}777$ (SD = 5,995)
Driving regularity [daily, weekly, monthly, annually]	Daily 40% Weekly 45% Monthly 15%	Daily 50% Weekly 40% Monthly 10%
Driving style *[5-Likert scale: from (1) very safe to very risky (5)]*	$M = 2.5$ (SD = 0.8)	$M = 2.5$ (SD = 0.9)
Driving pace *[5-Likert scale: from (1) very leisurely to very rapid (5)]*	$M = 3.3$ (SD = 0.6)	$M = 3.4$ (SD = 0.9)
Driving capability *[5-Likert scale: from (1) very inexperienced to very experienced (5)]*	$M = 4.0$ (SD = 0.8)	$M = 4.1$ (SD = 0.9)

the overtaking manoeuvre on the straight rural road, each of the test subjects drove a small winding course for a distance of 2,000 m through a wooded area so that the entire scenario would appear as natural as possible. To get a better overview, the scenario can be divided into five temporal and spatial stages from the EV's point of view (see Fig. 5): following a vehicle in front, swerving into the oncoming lane, passing the leading vehicle, merging back into the starting lane, and getting in the lane again.

3.5 Experimental design

We used a 2×3×3 mixed factorial design for this experiment. The human factors condition (no time pressure or time pressure) was the between-subject factor, while the environmental condition (basic, truck, fog and rain) and the function manipulation (no manipulation, manipulation of driving free LV, manipulation of driving free OV) were within-subject factors (see Fig. 6). Half of the participants experienced time pressure as realised by an expiring time counter in the head-up display, forcing them to overtake as early as possible. The timer was set to expire as soon as the fourth gap had passed, forcing the test persons to overtake in the gaps with the smaller and critical time gaps described in Sect. 3.4.2. The reason for this is that impatient drivers under time pressure tend to reduce the accepted gaps during passing manoeuvres (Pollatschek and Polus 2005). Each test subject drove all nine scenarios, comprising the three different environmental conditions as well as function manipulations, where the scenarios were permuted to mitigate potential sequence and learning effects. The basic condition corresponded to the standard scope of the examined FRAM model, whereas the LV, which was basically a passenger car, was substituted through a truck in the truck condition, and the weather conditions, that were basically sunny and dry, were changed to fog and rain in the third condition. The first three scenarios,

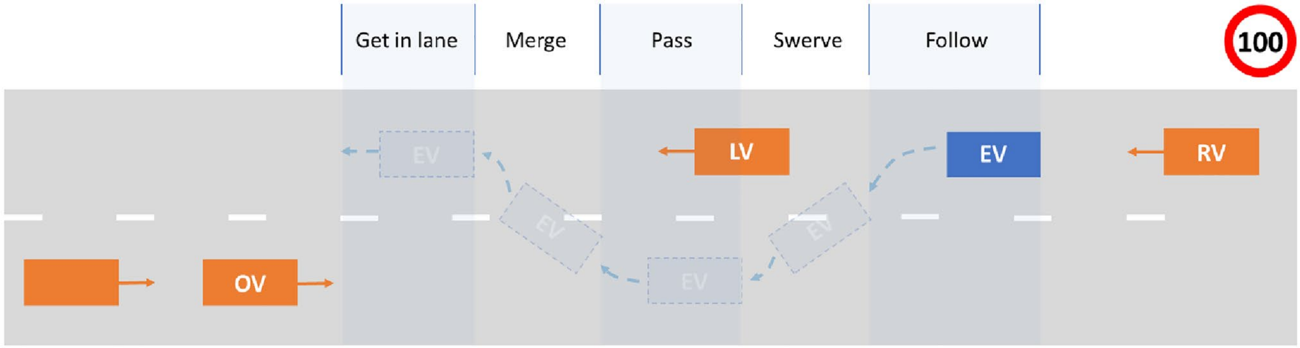
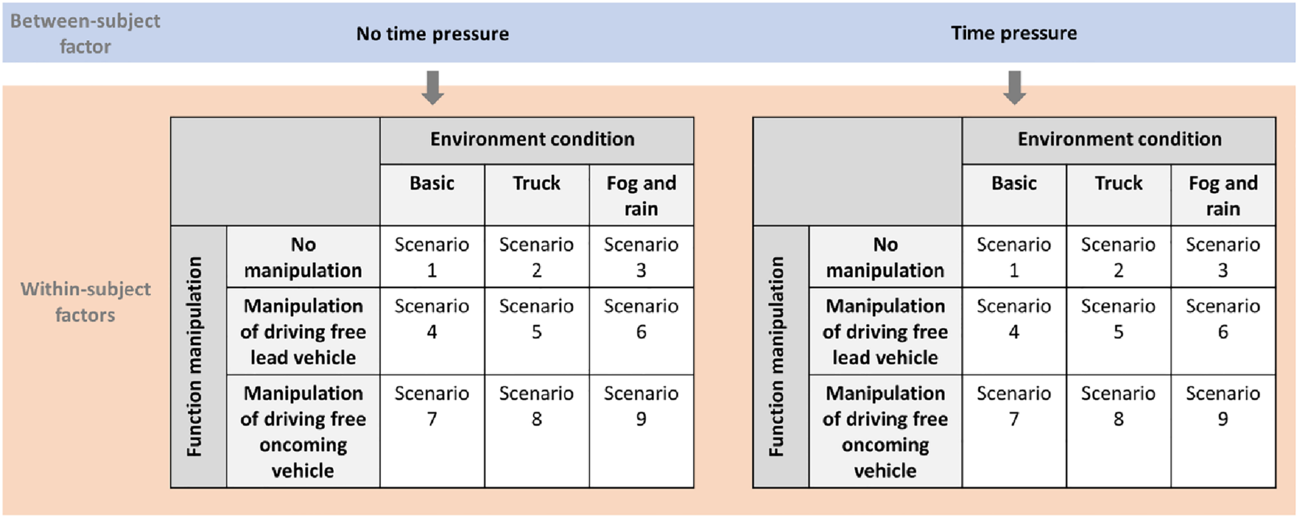


Fig. 5 Schematic illustration of the overtaking scenario comprising different road users/agents and divided into five temporal and spatial stages. *EV* ego vehicle, *LV* lead vehicle, *RV* rear vehicle, *OV* oncoming vehicle, according to Grabbe et al. (2022)



Between-subject factor: No time pressure

Within-subject factors

		Environment condition		
		Basic	Truck	Fog and rain
Function manipulation	No manipulation	Scenario 1	Scenario 2	Scenario 3
	Manipulation of driving free lead vehicle	Scenario 4	Scenario 5	Scenario 6
	Manipulation of driving free oncoming vehicle	Scenario 7	Scenario 8	Scenario 9

Between-subject factor: Time pressure

		Environment condition		
		Basic	Truck	Fog and rain
Function manipulation	No manipulation	Scenario 1	Scenario 2	Scenario 3
	Manipulation of driving free lead vehicle	Scenario 4	Scenario 5	Scenario 6
	Manipulation of driving free oncoming vehicle	Scenario 7	Scenario 8	Scenario 9

Fig. 6 Illustration of the mixed factorial design

in which no manipulation was implemented, served as references for the three environmental conditions to analyse the predictive validity for the function manipulation of driving free for both the LV and the OV. The manipulation of driving free was realised for the LV by multiple abrupt braking and acceleration as well as repeated lateral offsets, such as "weaving around ", and for the OV by increasing the speed from 100 to 120 km/h, which reduced the time gaps of the first four gaps to 8.33 s and for the last six gaps to 9.99 s, resulting in even more critical time gaps. Finally, scenarios 4 and 7 must be compared with scenario 1, scenarios 5 and 8 with scenario 2, and scenarios 6 and 9 with scenario 3 (see Fig. 6).

3.6 Procedure

Participants were welcomed and informed about the study goals and the procedure. After risks such as nausea and the option of withdrawing from the study without needing to cite reasons were outlined, written consent was obtained. Participants filled out a demographic questionnaire, which also asked for details of their driving experience and driving style. They then drove in the driving simulator for about 10 min to familiarise themselves with the steering, braking, and the driving simulator system. Afterwards, the participants drove a modified basic scenario with no oncoming vehicles and just the LV, with the goal of overtaking this. They were then asked to fill out a questionnaire to rate the timing and precision variability performance of some subjective functions on a 7-Likert scale, as will be more detailed in Sect. 3.7.1. This initial trial run served as a familiarisation for the participants with the basic procedure of the subsequent nine test drives as well as the non-trivial subjective rating of the functions. The actual nine test runs then began, each followed by completing the questionnaire on the subjective functions. Finally, the participants filled out five follow-up questions to rate the perception of the simulated drive. In general, the test subjects were instructed to overtake the LV before the end of the straight by obeying the traffic regulations but also showing her or his most natural and everyday driving behaviour. No restrictions were given regarding overtaking behaviour to ensure idiosyncratic and diverse driving styles. However, an exception exists for the subjects in the group with time pressure who were intentionally instructed to overtake the LV before the timer expires.

3.7 Measures and analysis

3.7.1 Independent and dependent variables

The overall study consisted of three independent variables comprising the function manipulation, environmental condition, and the human factors condition. Moreover, the dependent variables were the performance variability values of several subjective and objective functions in which the performance variability of their outputs, if driving free LV or OV are manipulated, should either change (expected direct downstream effects) or should not change (no expected direct downstream effects) according to the FRAM model by Grabbe et al. (2022) (see Table 3). It should be emphasised that the variability of the outcome or output from a function was measured and not the variability of the function itself. This work only investigated the expected and unexpected downstream couplings of the two manipulated functions to the functions of the agent EV and not to the other agents to test for predictive validity. In the case of expected direct downstream effects, the corresponding functions were assigned to the H/FA category, and in case of no expected direct downstream effects, the corresponding functions were assigned to the M/CR category, according to the application of SDT to FRAM as described basically at the end of Sect. 2.2. Furthermore, in the case of the M/CR category, these functions do not represent all of the potential functions that have to be tested, but only a selection of functions, as otherwise there would be far too many functions for any practical test. Theoretically, these would be all of the remaining functions of the entire model that are not expected to be directly or indirectly influenced by the manipulated functions.

The performance variability of the subjective functions was based on the rating of the timing and precision variability performance in the questionnaire on a 7-Likert scale. Here, the timing was coded as 1 for too early, 3 for on time, 5 for too late, and 7 for not at all, whereas precision was coded as 1 for precise, 4 for acceptable, and 7 for imprecise. The subjects were asked when (timing) or how (precision) they, for example, estimated the distance to OV until they swerved. Finally, the two values for timing and precision were multiplied into one representative value for the performance variability of the subjective functions. By contrast, the performance variability of the objective functions was based on driving data (e.g., speed, lane deviation, and distance between cars) gathered in the driving simulator. For the sake of simplicity, we gathered the performance variability of the objective functions either in terms of timing or precision but not both. An overview of the measurement definitions of each objective function is given in Table 4.

3.7.2 Statistical analysis

To evaluate the predictive validity, the performance variability had to be reduced into the four events of SDT, namely, hits, false alarms, misses, and correct rejections, by comparing the predictions of the model with the observations in the simulator.

Table 3 Assignment of manipulated functions and analysed functions of EV, and their allocation to the type of rating and SDT event category

Manipulated function	Analysed functions of EV	Type of rating	SDT event category
Driving free LV	Check vehicles in front of LV	Subjective	H/FA
	Check LV is not about to change speed		H/FA
	Gauge future driving actions of LV		H/FA
	Check LV is not indicating or about to turn		H/FA
	Maintain an adequate view of the road ahead		H/FA
	Evaluate reasonableness for overtaking		H/FA
	Assess the situation to enter safely		H/FA
	Judge LV's relative speed to OV		H/FA
	Judge LV's speed		H/FA
	Judge available passing time		H/FA
	Determine pass can be completed		H/FA
	Observe road behind		M/CR
	Check for safe distance to merge		M/CR
	Judge first OV's speed		M/CR
	Judge distance from first OV		M/CR
	Maintain headway separation	Objective	H/FA
	Keep in lane		H/FA
	Position car to the right		H/FA
	Position car to the left		H/FA
	Reduce headway from normal following		H/FA
	Avoid tailgating and intimidating LV		H/FA
	Adjust speed to that of LV		H/FA
	Adopt overtaking position		H/FA
	Swerve completely to the oncoming lane		H/FA
	Accelerate LV decisively		H/FA
	Merge back into starting lane		H/FA
	Merge progressively into starting lane		H/FA
	Comply with the speed limit		M/CR
Driving free OV	Judge first OV's speed	Subjective	H/FA
	Judge LV's relative speed to OV		H/FA
	Judge available passing time		H/FA
	Determine pass can be completed		H/FA
	Assess the situation to enter safely		H/FA
	Judge distance from first OV		M/CR
	Judge LV's speed		M/CR
	Observe road behind		M/CR
	Check for safe distance to merge		M/CR
	Accelerate LV decisively	Objective	H/FA
	Merge back into starting lane		H/FA
	Merge progressively into starting lane		H/FA
	Comply with the speed limit		M/CR
	Maintain headway separation		M/CR
	Keep in lane		M/CR

First, the mean and standard deviation of the performance values were calculated for the scenarios 1–3 (as these form the respective reference for testing for differences in performance variability as described in Sect. 3.5) for each analysed function per between-subject factor group, from which the 95% confidence interval was calculated to define a "normal" everyday variability range. In medicine, one also speaks of normal ranges, which are defined for blood pressure or blood

Table 4 Overview of the measurement definitions of each objective function

Objective function	Stage	Phenotype	Definition
Maintain headway separation	Follow	Precision	The average distance between EV and LV in the period, where the straight begins and the driver of EV starts to swerve, indicated by the left activated indicator or the steering angle
Keep in lane	Follow	Precision	The average absolute lane deviation between of EV in the period, where the straight begins and the driver of EV starts to swerve
Position car to right/left	Follow	Precision	The average gap to the left/right lane edge of in the period, where the straight begins and the driver of EV starts to swerve
Reduce headway from normal following	Swerve	Precision	The average distance between EV and LV in the period, where the driver of EV starts to swerve and driving completely in the oncoming lane, indicated by the left activated indicator or the steering angle, and the lane index showing in which lane EV is driving, respectively
Avoid tailgating and intimidating LV	Swerve	Precision	The distance between EV and LV at the last point, where the driver of EV is driving in the starting lane and already has started to swerve
Adjust speed to that of LV	Swerve	Precision	The average speed difference between EV and LV in the period, where the straight begins and the driver of EV starts to swerve
Adopt overtaking position	Swerve	Precision	The sum of the speed of EV, absolute lane deviation of EV, and the distance between EV and LV at the point, where the driver of EV starts to swerve
Swerve completely to oncoming lane	Swerve	Timing	The time difference between starting to swerve and driving completely in the oncoming lane
Accelerate LV decisively	Pass	Precision	The average speed of EV in the period, where the driver of EV starts to drive completely in the oncoming lane and starts to merge, indicated by the lane index showing in which lane EV is driving, and the right activated indicator or the steering angle, respectively
Merge back into starting lane	Pass	Precision	The number of times the driver of EV merged back into the starting lane even though the driver has already swerved into the oncoming lane to overtake
Merge progressively into starting lane	Merge	Timing	The time difference between starting to merge and driving completely in the starting lane
Comply with the speed limit	All	Precision	The average speed difference between EV's speed and the speed limit in the period, where the straight begins and the driver of EV is driving completely in the starting lane again after passing LV

sugar, for example, to distinguish healthy patients from sick patients. Afterwards, the difference between the upper/lower limit of the confidence interval and the mean was calculated, which reflects a maximum positive or negative everyday fluctuation in performance that is normal and thus should not be regarded as a significant performance variability.

We then calculated the absolute differences between the intraindividual performance values of scenario 4 and 7 to 1, scenario 5 and 8 to 2, and scenario 6 and 9 to 3 for each analysed function as we were interested in both the positive and negative direction of the performance variability. In the next step, one-sided one-sample *t* tests with a p-value of 5% were used to determine whether the sample mean of the absolute differences in performance of, for example, scenario 4 to 1 was statistically greater than the respective maximum value of everyday fluctuation in performance. The Wilcoxon signed-rank test was used as an alternative when the statistical requirements for the one-sample t-test were not met. If the p-value was lower than 5%, then a significant performance variability in the analysed function in the respective scenario was assumed, otherwise not.

From this, it was possible to finally assess which of the four events according to SDT applies per analysed function, group and scenario. Subsequently, the number of the four events per manipulated function (driving free LV and OV), human factors condition (time pressure, no time pressure) and environmental condition (basic, truck, rain and fog) were calculated. Based on this, the accuracy, H-rate (HR), and CR-rate (CRR) were calculated to be able to prove the predictive validity. We decided to use the accuracy and not the Matthews (1975) correlation coefficient (MCC), which is generally recommended by Stanton and Young (1999a, 2003) and successfully applied, for example, by Stanton et al. (2021a; b, c, d) and Hulme et al. (2021a; b, c), as an appropriate statistical metric to validate human factors methods in binary classification problems. The reasons are twofold. First, the analysed FRAM model is clearly complex with a wide scope, and according to Stanton and Young (2003), the wider the scope of the method or model, the more difficult it is to obtain favourable data on validity performance, so it would be detrimental to use a harsh metric like the MCC. Second, the true positive results should be favoured over the true negative results as considerably more

H than CR can be identified in an FRAM model validation due to the practical limitations mentioned in Sect. 3.7.1. The accuracy score tends to favour positive cases (Baber and Young 2022). With this in mind, accuracy seemed to be more appropriate than MCC to obtain a high validity score, because it is quite difficult to obtain a high score through good prediction results in only all four of the confusion matrix categories. Nevertheless, as using the accuracy alone as a single value to prove predictive validity could be misleading in the case of imbalanced classification data sets (cf. Chicco and Jurman 2020), we also considered the HR, and especially CRR, to achieve a broader and more detailed analysis.

The numerical value of accuracy represents the proportion of true or expected results (both true positive (H) and true negative (CR)) and was calculated as follows (1):

$$\text{Accuracy} = \frac{\text{H} + \text{CR}}{\text{H} + \text{FA} + \text{M} + \text{CR}} \tag{1}$$

HR or sensitivity represents the proportion of true positives or expected and observed results and was calculated as follows (2):

$$\text{HR} = \frac{\text{H}}{\text{H} + \text{M}} \tag{2}$$

CRR or specificity represents the proportion of true negatives or not expected and not observed results and was calculated as follows (3):

$$\text{CRR} = \frac{\text{CR}}{\text{FA} + \text{CR}} \tag{3}$$

All three metrics are expressed along a percentage scale ranging from 0 to 100. Ultimately, a criterion for acceptable levels of predictive validity has to be considered, as there is no universally accepted measure. A review of reliability and validity levels found that, across 25 studies, the average value used to indicate acceptable percentage agreement was 76%, with a range of 70–88% (Olsen 2013). As described in Sect. 2.1, validation is gradual rather than binary. Thus, a single value indicating that an FRAM model is predictively valid or not seems to be inappropriate. Rather, a more differentiated approach was used in this work, defining different levels for predictive validity from *poor* to *almost perfect* according to the reliability result levels applied to SDT by Olsen (2013) (see Table 5). However, to answer the research questions in Sect. 3.2, we additionally defined a value for sufficient predictive validity, which was set at 70%. We have chosen this value, because it defines first, the minimum of acceptable percentage agreement (Olsen 2013), and second, the median of the category of substantial predictive validity, which should be the minimum category to aim for (see Table 5).

Table 5 Levels for predictive validity associated with percentages of selected metrics according to Olsen (2013)

Predictive validity level	Percentage of accuracy, HR, and CRR
Poor	0
Slight	> 0–20
Fair	21–40
Moderate	41–60
Substantial	61–80
Almost perfect	81–100

4 Results

This section presents the results according to the three different research questions defined in Sect. 3.2. An overview of the results of the SDT event category for every analysed function per manipulated function, with a differentiation between human factors and environmental conditions, is shown in Table 6.

4.1 Predictive validity for the basic scenario

Figure 7 shows the comparison of the accuracy, HR and CRR associated with the predictive validity levels (see Table 5) between the environmental and human factor conditions with the manipulated function of driving free LV. Furthermore, the 70% threshold as the value for sufficient predictive validity is indicated by a horizontal dashed red line. For the basic scenario in the no time pressure group, the accuracy, HR, and CRR account for 79%, 81% and 0%, respectively. The accuracy and HR lie above the sufficient predictive validity, reaching a substantial and almost perfect predictive validity level, respectively. However, the predictive validity level of the CRR is poor. In total, there are six (21%) functions that do not meet expectations: '*observe road behind*', '*check for safe distance to merge*', '*judge first OV's speed*', '*judge distance from first OV*' (all are M instead of CR) as subjective functions and '*merge back into starting lane'* (FA instead of H) and '*comply with the speed limit'* as objective functions (M instead of CR). It is noticeable that the false predictions are mainly based on misses.

Figure 8 is the same as Fig. 7, but for the manipulated function of driving free OV. Here, the accuracy, HR, and CRR account for 53%, 56% and 50%, respectively, for the basic scenario in the no time pressure group. Therefore, all three metrics lie below the sufficient predictive validity and achieve a moderate predictive validity level. In total, there are seven (47%) functions that do not meet expectations: '*judge LV's relative speed to OV'* (FA instead of H), '*judge distance from first OV'*, '*judge LV's speed'*, and '*observe road behind'* (all are M instead of CR) as subjective

Table 6 Assignment of manipulated functions and analysed functions of EV, and their respective results of SDT event category, with a differentiation between human factors and environmental conditions

Manipulated function	Analysed functions of EV	No time pressure			Time pressure		
		Basic	Truck	Rain/fog	Basic	Truck	Rain/fog
Driving free LV	Check vehicles in front of LV	H	H	FA	H	H	H
	Check LV is not about to change speed	H	H	H	H	H	H
	Gauge future driving actions of LV	H	H	H	H	H	H
	Check LV is not indicating or about to turn	H	H	H	H	H	H
	Maintain an adequate view of the road ahead	H	H	H	H	H	H
	Evaluate reasonableness for overtaking	H	FA	H	H	H	H
	Assess the situation to enter safely	H	H	H	FA	FA	FA
	Judge LV's relative speed to OV	H	H	H	H	H	H
	Judge LV's speed	H	H	H	H	H	H
	Judge available passing time	H	H	H	H	H	H
	Determine pass can be completed	H	H	H	H	H	H
	Observe road behind	M	CR	M	M	M	M
	Check for safe distance to merge	M	CR	M	M	M	CR
	Judge first OV's speed	M	M	M	M	M	M
	Judge distance from first OV	M	M	M	CR	CR	CR
	Maintain headway separation	H	H	FA	H	H	FA
	Keep in lane	H	H	H	H	H	FA
	Position car to the right	H	FA	FA	H	H	H
	Position car to the left	H	FA	FA	H	H	H
	Reduce headway from normal following	H	H	H	H	H	H
	Avoid tailgating and intimidating LV	H	H	H	H	H	H
	Adjust speed to that of LV	H	H	H	H	H	FA
	Adopt overtaking position	H	H	H	H	H	H
	Swerve completely to the oncoming lane	H	H	H	H	H	FA
	Accelerate LV decisively	H	H	H	H	H	H
	Merge back into starting lane	FA	FA	FA	FA	FA	FA
	Merge progressively into starting lane	H	H	H	H	H	H
	Comply with the speed limit	M	M	M	M	M	M
Driving free OV	Judge first OV's speed	H	FA	H	H	H	FA
	Judge LV's relative speed to OV	FA	H	H	H	H	FA
	Judge available passing time	H	FA	H	H	H	H
	Determine pass can be completed	H	H	H	H	H	H
	Assess the situation to enter safely	H	FA	H	FA	H	FA
	Judge distance from first OV	M	M	M	CR	CR	CR
	Judge LV's speed	M	M	M	CR	CR	M
	Observe road behind	M	M	M	M	M	CR
	Check for safe distance to merge	CR	M	M	M	M	M
	Accelerate LV decisively	FA	FA	H	FA	H	FA
	Merge back into starting lane	FA	FA	FA	FA	FA	FA
	Merge progressively into starting lane	H	FA	H	FA	H	H
	Comply with the speed limit	CR	M	M	M	CR	CR
	Maintain headway separation	CR	CR	CR	CR	M	CR
	Keep in lane	M	CR	M	M	M	CR

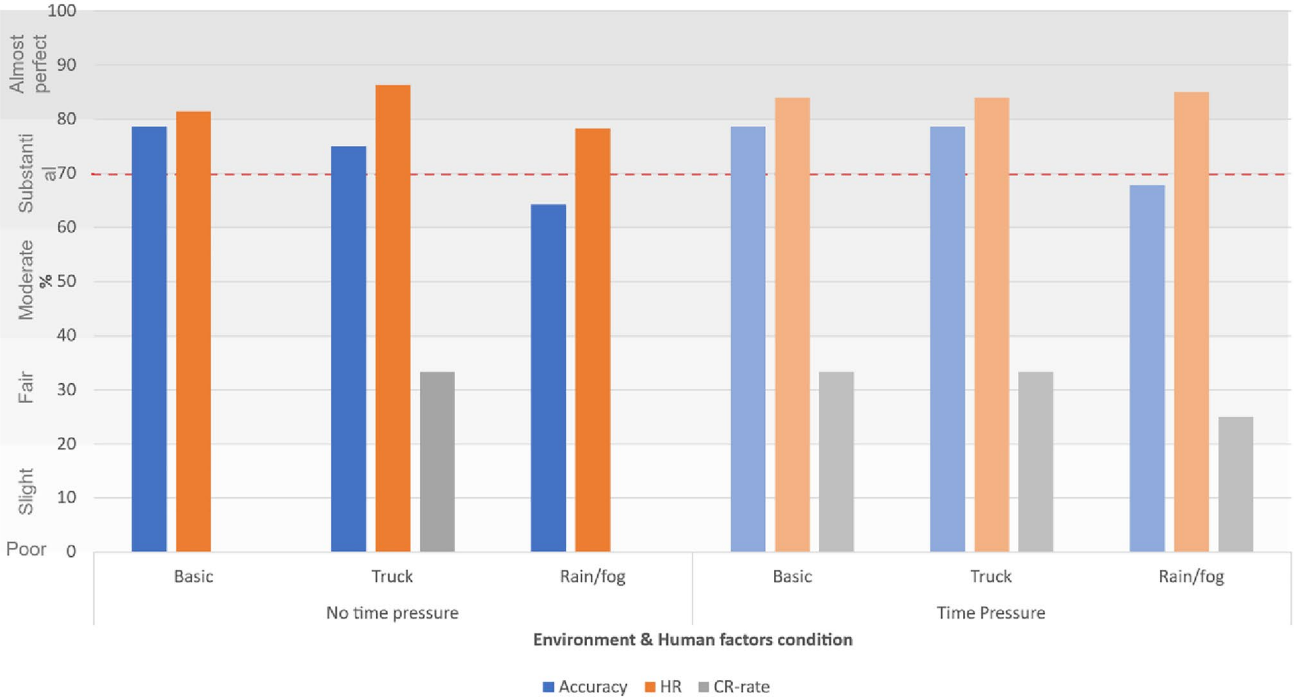


Fig. 7 Comparison of the accuracy, HR, and CRR associated with the predictive validity levels between the environmental and human factors conditions for the manipulated function of driving free LV

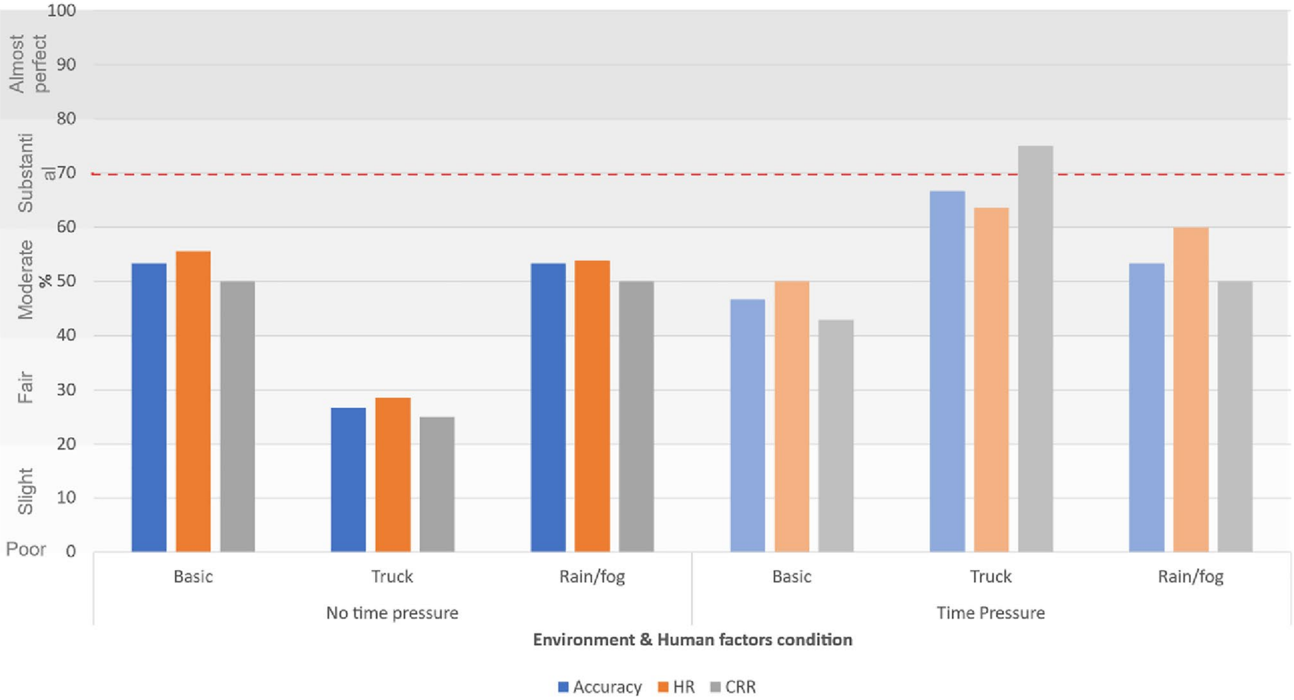


Fig. 8 Comparison of the accuracy, HR, and CRR associated with the predictive validity levels between the environmental and human factors conditions for the manipulated function of driving free OV

functions and '*accelerate LV decisively*', '*merge back into starting lane*' (both are FA instead of H) and '*keep in lane*' (M instead of CR) as objective functions. There is a roughly equal distribution of false alarms and misses here.

Besides, a comparison of the accuracy, HR, and CRR between objective and subjective functions shows no clear differences in terms of the type of rating (see Fig. 9). In most cases, the differences amount to a maximum of 10% and alternate, so that sometimes the objective functions achieve a higher value than the subjective functions and vice versa.

4.2 Predictive validity for other environmental conditions

The accuracy, HR, and CRR account for 75%, 86% and 33%, respectively, for the truck scenario in the no time pressure group with the manipulated function of driving free LV (see Fig. 7). Thus, the accuracy and HR lie above the sufficient predictive validity, reaching a substantial and almost perfect predictive validity level, respectively. However, the predictive validity level of the CRR is fair. These results are similar to the ones of the basic scenario. For the rain/fog scenario in the no time pressure group with the manipulated function of driving free LV, the accuracy, HR, and CRR account for 64%, 78% and 0%, respectively (see Fig. 7). Therefore, the accuracy lies below and the HR lies above the sufficient predictive validity, both reaching a substantial predictive validity level, respectively. However, the predictive validity level of the CRR is poor. Slight differences can thus be determined compared to the basic scenario.

For the truck scenario in the no time pressure group with the manipulated function of driving free OV, the accuracy, HR, and CRR account for 27%, 29% and 25%, respectively (see Fig. 8). This means that all three metrics lie below the sufficient predictive validity, reaching a fair predictive validity level. Compared to the basic scenario, this is one level lower. In contrast, the accuracy, HR, and CRR account for 53%, 54% and 50%, respectively, for the rain/fog scenario in the no time pressure group with the manipulated function of driving free OV (see Fig. 8). Thus, all three metrics lie below the sufficient predictive validity, reaching a moderate predictive validity level. These results are similar to those for the basic scenario.

If we consider the functional level of each analysed function and respective changes to the SDT event category between the environmental conditions in relation to the basic scenario for the no time pressure group in Fig. 10, we see that in the truck scenario, 18% of the analysed functions for the manipulated function of driving free LV and 53% of the analysed functions for the manipulated function of driving free OV deviate relative to the SDT event category. In the rain/fog scenario, 14% of the analysed functions for the manipulated function of driving free LV and 33% of the analysed functions for the manipulated function of driving

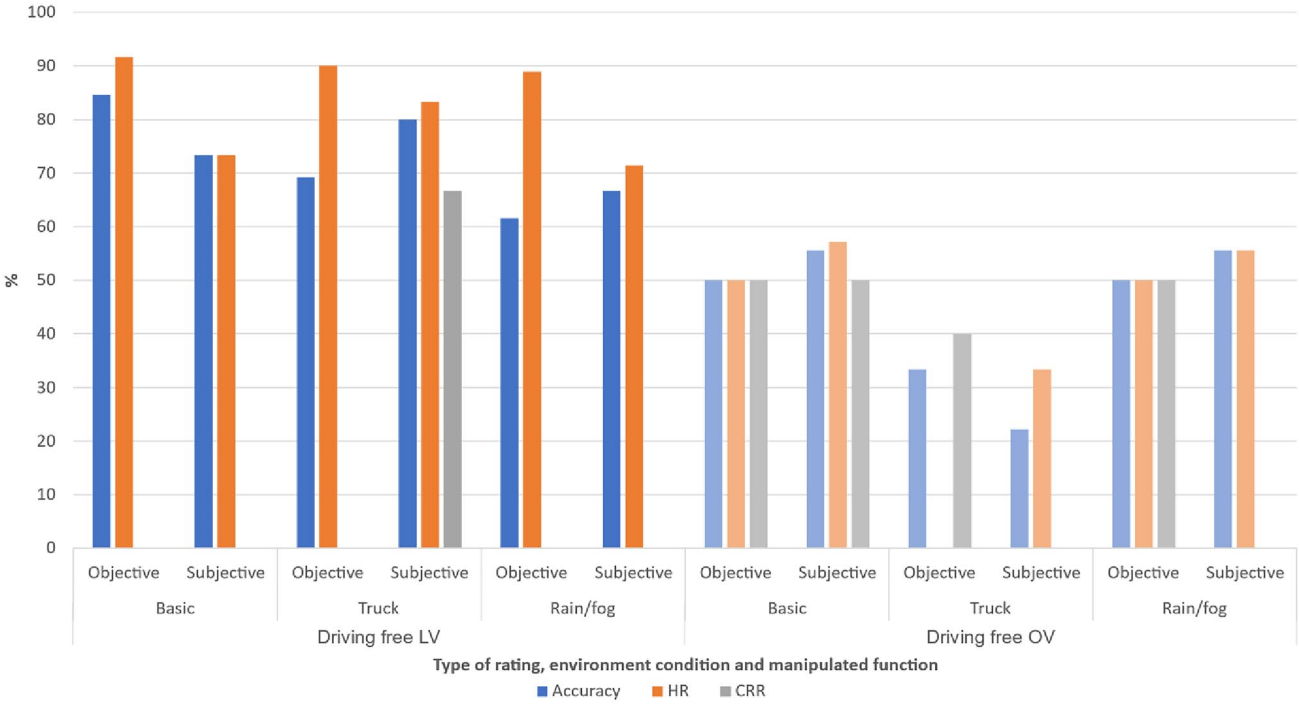


Fig. 9 Comparison of the accuracy, HR, and CRR between the subjective and objective type of rating, with a differentiation between the environmental conditions and manipulated function

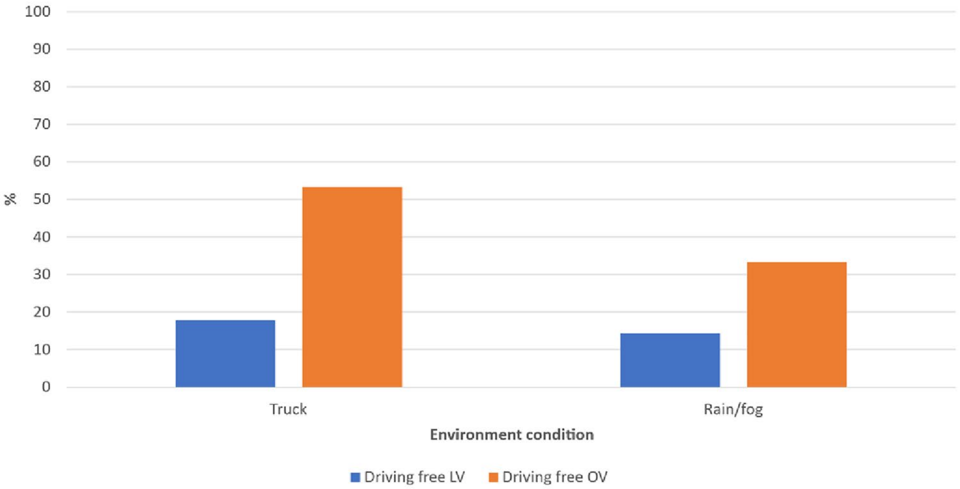


Fig. 10 Relative frequencies of deviations in the SDT event categories within the analysed functions between the manipulated function of driving free LV and OV for the no time pressure group in the truck, and fog and rain scenario, each compared to the basic scenario

free OV deviate relative to the SDT event category. Hence, we can see far greater differences in the predictive validity on the functional level when the environmental condition changes compared to the results of the three metrics shown above, especially with the manipulated function of driving free OV. The number of deviations between the two environmental conditions are similar in the case of the manipulation of driving free LV and different in the case of the manipulation of driving free OV, where the truck scenario shows considerably more deviations than the rain/fog scenario.

4.3 Predictive validity for other human factors conditions

First, we present the results for the manipulated function of driving free LV. For the basic scenario in the time pressure group, the accuracy, HR, and CRR account for 79%, 84% and 33%, respectively (see Fig. 7). Thus, the accuracy and HR lie above the sufficient predictive validity, reaching a substantial and almost perfect predictive validity level, respectively. However, the predictive validity level of the CRR is fair. These results are similar to those for the basic scenario in the no time pressure group, except for the CRR, which is two levels higher. The accuracy, HR, and CRR account for 79%, 84% and 33%, respectively, for the truck scenario in the time pressure group (see Fig. 7). Therefore, the accuracy and HR lie above the sufficient predictive validity, reaching a substantial and almost perfect predictive validity level, respectively. However, the predictive validity level of the CRR is fair. These results are similar to those for the truck scenario in the no time pressure group. The accuracy, HR, and CRR account for 68%, 85% and 25%, respectively, for the rain/fog scenario in the time pressure group (see Fig. 7). Hence, only the HR lies above the sufficient predictive validity, reaching an almost perfect predictive validity level. However, the predictive validity levels of the accuracy and CRR are substantial and fair, respectively. These results are similar to those for the rain/fog scenario in the no time pressure group, except for the CRR, which is two levels higher.

The results for the manipulated function of driving free OV are presented below. The accuracy, HR, and CRR account for 47%, 50% and 43%, respectively, for the basic scenario in the time pressure group (see Fig. 8). This means that all three metrics lie below the sufficient predictive validity level, reaching a moderate predictive validity level. These results are similar to those for the basic scenario in the no time pressure group. The accuracy, HR, and CRR account for 67%, 64% and 75%, respectively, for the truck scenario in the time pressure group (see Fig. 8). Therefore, all three metrics reach a substantial predictive validity level, but only the CRR achieves the sufficient predictive validity threshold. These results differ from those for the truck scenario in the no time pressure group, since all three metrics are one predictive validity level higher. The accuracy, HR, and CRR account for 53%, 60% and 50%, respectively, for the rain/fog scenario in the time pressure group (see Fig. 8). Thus, all three metrics lie below the sufficient predictive validity, reaching a moderate predictive validity level. These results are similar to those for the rain/fog scenario in the no time pressure group.

On the functional level of each analysed function and their possible respective changes to the SDT event category between the human factors conditions relative to each environmental condition in Fig. 11, we can see a trend of increasing deviations for the manipulated function of driving free LV, starting from the basic scenario (7%), via the truck scenario (25%) to the rain/fog scenario (32%). In contrast, the deviations for the basic scenario (47%), the truck scenario (67%) and the rain/fog scenario (53%) are similar in the basic and rain/fog scenario, whereas the truck scenario has a clearly greater deviation in the case of the manipulation of

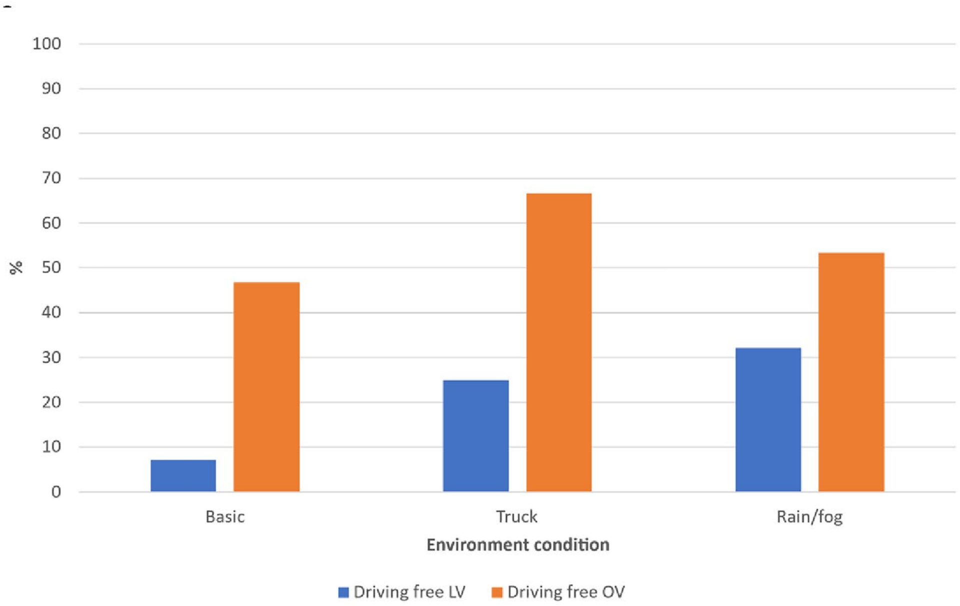


Fig. 11 Relative frequencies of deviations in the SDT event categories within the analysed functions between the manipulated function of driving free LV and OV, comparing each environmental condition between the no time/time pressure groups

driving free OV. Moreover, the number of deviations is considerably higher than for driving free LV. We can see much greater differences in predictive validity on the functional level when the human factors condition changes compared to the results of the three metrics shown above, the same as with the environmental conditions. However, the number of deviations is below 10% in the case of the manipulation of driving free LV for the basic scenario, which should be acceptable, whereas the remaining cases represent clearly higher deviations.

5 Discussion

The aim of this paper is first, to define a more formal approach to achieving and demonstrating the reliability and validity of an FRAM model, and second, to apply this formal approach partly to an existing FRAM model so as to prove its validity and to evaluate the applicability of this approach. In the first part of the paper, a formal approach was derived by transferring both the general understanding and definitions of reliability and validity along with concrete methods and techniques that have been applied in other research areas, or specifically to HFE methods, to the concept of FRAM. In the second part, the predictive validity, which is one part of the formal approach to demonstrate the evaluative part of the validity of an FRAM model, was assessed for a specific FRAM model by Grabbe et al. (2022) in a driving simulator study. Predictive validity represents the highest maxim of validation and the remaining parts of the formal approach had already been applied by Grabbe et al. (2022). Finally, the results of the study have to be discussed so as to prove the credibility of the analysed FRAM model, to cover methodological limitations and to evaluate the utility and applicability of the approach in general.

5.1 Predictive validity of the analysed FRAM model

The research questions from Sect. 3.2 have to be answered in the following to assess the predictive validity of the analysed FRAM model. The following rule applies here: if both the accuracy and HR are sufficient, then predictive validity can be assumed as the true positive results are favoured over the true negative results.

The FRAM model is predictively valid for the basic scenario in the case of the manipulation of driving free LV, because the accuracy and HR are sufficient and reach at least a substantial predictive validity level with high sensitivity. However, the CRR is poor due to several misses, indicating a low specificity. In contrast, the FRAM model is not enough predictively valid for the basic scenario in case of the manipulation of driving free OV, because all three evaluation criteria are insufficient and only reach a moderate level of predictive validity. Overall, the results show that the predictive validity of the FRAM model for the basic scenario is limited, in particular in its specificity, indicating deficiencies in the credibility of the examined FRAM model. In total, the couplings to 13 functions have to be updated. The validation performance of the FRAM model is comparable with the better performing HFE methods in terms of validation (Stanton and Young 1999a; Stanton et al. 2013) only in case of the manipulation of driving free LV, except for the low specificity. Some of the best methods in the field, for example, are associated with the prediction of human error (Baber and Stanton 1996; Harris et al. 2005; Stanton et al. 2009).

These better-performing methods typically achieve validity statistics above 0.8 (Stanton et al. 2021b).

When comparing the differences between the environmental conditions, the results show that the predictive validity is comparable between the three different conditions for both manipulation cases, apart from the truck condition for manipulation of driving free OV, though the deviations in the SDT event categories within each analysed function are clearly high. Therefore, the FRAM model is not predictively valid for other environmental conditions. When the human factors condition is changed, the results indicate that the predictive validity is similar for the two conditions with every environmental condition and both manipulation cases, except the truck condition for manipulation of driving free OV. However, the deviations in the SDT event categories within each analysed function are once again clearly high, except the basic condition for manipulation of driving free LV, which shows low deviations. Hence, the FRAM model is predictively valid for other human factors conditions in the case of the basic scenario with the manipulation of driving free LV, but not for the remaining cases. Consequently, it can be said that a generalisation of the predictive validity of an FRAM model is greatly limited so that an FRAM model has to be adapted to changes in both the environmental as well as human factors conditions, especially if conditions are combined. This is not surprising as an FRAM model can only be validated for specific instantiations, and if the conditions change, the instantiation will change and the model will then have to be adapted and no generalisation will be possible. Against this background, it can also be assumed that the effects of shared and traded control (Sheridan 1992) between the driver and an automation system by enhancing the scenario through an interaction of the driver with an advanced driver assistance system (ADAS), e.g., lane-keeping assist (LKA) and adaptive cruise control (ACC), cannot be validly predicted without adapting the FRAM model. Here, the effects and their prediction of conflict or confusion situations between the two agents would be of particular interest. For example, a dangerous situation can occur when the driver performs a lane change without activating the turn signal, which the LKA could then interpret as an unintentional drift and decide to return the car to the main lane. In addition, this could lead to a decrease in trust or an increased stress level which in turn degrade the driving performance or potentially result in a deactivation of the ADAS by the driver. Such conflicting decisions are called human–machine dissonance when contradictory information exists between humans' and autonomous systems' knowledge, from information processing to actions on a controlled process (Vanderhaegen 2021), and these discrepancies can affect human factors and produce, e.g., discomfort, overload, or stress (Vanderhaegen 2014, 2016). In the FRAM model examined, these conflicts are already present in the form of human–human dissonances, e.g., the manipulation of driving free for the LV by multiple abrupt braking and acceleration could be interpreted by the driver of EV in two main aspects: either that LV is reacting to an obstacle or leading vehicle or that the driver of LV is drunk. Here, the result of the interpretation probably leads to two different reactions of the driver of EV which can lead to dangerous situations. For instance, the EV's driver gauges future driving actions of LV which is significantly facilitated when LV is reacting to the traffic in front and the EV's driver has a clear sight compared to the situation when LV's driver is drunk as her/his driving behaviour is random. In addition, the strange driving behaviour of LV may affect human factors by causing anxiety or increased stress for the driver of EV. This could be a possible reason for false expectations in the FRAM model. Thus, various behavioural changes can be triggered in the system, which primarily affects human factors, which in turn cause behavioural adaptations in the system through interdependencies (cf. Wege et al. 2014). As previously described, changing human factors conditions and their effects cannot be fully predicted with the FRAM model. It would, therefore, be relevant in the future to adapt the FRAM model in this direction and to prove whether the FRAM model is valid in the context of interaction between drivers and ADAS. This appears to be especially important given the increasing introduction of such automation systems into the road system and their risk assessment. In principle, possible conflicts in the sense of dissonance can be represented and identified in an FRAM model via the couplings between the functions when analysing them in the form of "what-if analyses" (Hill et al. 2020; MacKinnon et al. 2021) to understand how a potential conflicting coupling affects several downstream functions and how this propagates through the system.

5.2 Limitations

Some methodological limitations are discussed in the following, including the sample, the driving simulator validity as well as the test setup, the statistical analysis, and the theoretical concept of the predictive validation approach.

The participant characteristics play a role in a driving simulator study (Blana 1996). The narrower sample here might not represent the entire driver population, which is why the evaluation of predictive validity based on performance variability is only valid to a limited extent. Nevertheless, the sample size can be considered as sufficient for the narrower population, since a sample size of 20 test drivers, for example, is sufficient to test the controllability of driver assistance systems according to ISO 26262 (2018).

If we take a closer look at the perceptions in the simulator and compare these between the two different groups (see Table 7), we see that the feeling of time pressure cannot be

Table 7 Comparison of the no time pressure and time pressure group with regard to perception in the driving simulator

Measurement	No time pressure group	Time pressure group
Realistic simulation behavior *[5-Likert scale: from (1) very realistic to very unrealistic (5)]*	*M* = 2.5 (SD = 1.0)	*M* = 2.6 (SD = 1.1)
Realistic driving behaviour of other road users *[5-Likert scale: from (1) very realistic to very unrealistic (5)]*	*M* = 2.8 (SD = 1.1)	*M* = 2.6 (SD = 0.9)
Equivalent overtaking manoeuvers in real life *[5-Likert scale: from (1) very equal to very unequal (5)]*	*M* = 3.1 (SD = 1.4)	*M* = 3.0 (SD = 1.2)
The feeling of time pressure *[5-Likert scale: from (1) very strong to very weak (5)]*	*M* = 3.2 (SD = 1.0)	*M* = 3.5 (SD = 1.0)
Efficiency/safety trade-off of overtaking manoeuver *[5-Likert scale: from (1) efficient to safe (5)]*	*M* = 2.3 (SD = 1.0)	*M* = 2.2 (SD = 1.1)

assumed for the time pressure group as the value is even higher than in the no time pressure group. Furthermore, there are no clear differences in the efficiency/safety trade-off between both groups, which is in contrast to the expectation that the no time pressure group should drive as safely as possible, and the pressure group more efficiently. Thus, it is questionable whether the measures to generate time pressure actually worked. According to Rastegary and Landy (1993), time constraints such as those used in this study may be insufficient for eliciting time pressure per se. These authors attested that not having enough time creates a feeling of time pressure only if the time limit is compulsory and if violating the time limit leads to a sanction. Although the time limit was compulsory, it did not lead to any sanctions. Nevertheless, almost all the drivers in the group with time pressure tried to overtake seriously before the time expired and actually overtook. Therefore, it can be argued that the main intention, to simulate impatient drivers under time pressure who tend to reduce the accepted gaps while performing passing manoeuvres, was accomplished.

According to Grabbe et al. (2022), a driving simulator is an appropriate tool for assessing performance variability in terms of action functions at the operational level, but not for perception and cognitive functions, where we have chosen a mix of objective and subjective measurement of performance variability. This leaves room for criticism, as the variables selected to measure performance and the data collection measures affect the driving simulator validity (Blana 1996; Kaptein et al. 1996). In particular, the variability measured subjectively could be limited in representing the real performance variability as the self-awareness of humans about their performance may be biased. However, this does not appear justified, since no great differences could be found between the type of rating and level of validity. Another issue is the definition of performance variability for the objective functions. For the sake of simplicity, their variability was based either on a timing or a precision metric, but not both. Furthermore, the variability measurement of each objective function was subjectively defined. Thus, it is uncertain whether the variability that is measured objectively completely fits the real performance of a respective function.

The driving simulator could, on the whole, have a great impact on the validity results as the validity of driving simulators is an ongoing concern. Typically, they are valuable tools in road safety and human factors research and have been used to assess a variety of driving performances (Mullen et al. 2011) by providing a safe and controllable environment to investigate driver behaviour ethically, effectively, and efficiently (Larue et al. 2018). However, simulators will never reproduce reality accurately and tend to compromise real-life situations (Espié et al. 2005). For instance, participants will probably not drive normally, because they perceive the driving task as a game, experience motion sickness, or find the driving task unrealistic (Larue et al. 2018). In particular, simulator validity depends on the simulator fidelity (Hoskins and El-Gindy 2006; Nilsson 1993), the specific driving task, and the realism of its implementation (Kaptein et al. 1996). Ultimately, literature shows that relative validity for driving simulators can be assumed, but absolute validity is limited (Mullen et al. 2011). This means that the validation results of the FRAM model are valid within the simulator environment but cannot be completely transferred to real on-road behaviour.

The calculation of the normal range of everyday variability per analysed function could be improved in the future by performing the reference scenarios 1–3 at least twice to discover which deviations in variability are normal, even if the participants are driving the same scenario again. However, this would increase the number of scenarios as well as the time needed, which was already high for the test subjects. This makes it a cost–benefit question, where we think that our simplified approach should be acceptable and sufficient.

In addition, the purely descriptive evaluation of the predictive validity can be criticised. It should be remembered that the focus of the predictive validity assessment was to analyse those functions, and how many functions, for which the predictions about performance variability through the FRAM model are valid or invalid rather than to know the number of test subjects for which the predictions are valid or not. The reason for this function focus is that potential invalid predictions could subsequently be refined to calibrate the model, which would otherwise be impossible. Therefore, it was not possible to calculate a distribution of the evaluation metrics per scenario, but only a single value in each case. This is why no inferential statistical analysis could be applied to evaluate the potential effects of changing environmental or human factors conditions.

Furthermore, scientific researchers can employ several statistical rates to evaluate binary classifications and their confusion matrices. In this work, the accuracy, HR, and CRR are used to evaluate the predictive validity in contrast to the MCC. This contradicts the general recommendation of Stanton and Young (1999a, 2003) to use the MCC as an appropriate statistic for the validation of human factors methods using the SDT, as well as the conclusion of Chicco and Jurman (2020) that the MCC is the most informative score for evaluating binary classification tasks and should be given preference over accuracy and F1 score by all scientific communities. However, the findings of Zhu (2020) challenge this general statement. Finally, there is no clear recommendation that just one specific metric should be used; this depends to a large extent on the context of the use and objective of the validation. Rather, a mix of different metrics, as applied in this paper, should be used to avoid misleading interpretations.

Last but not least, some methodological issues concerning the theoretical concept of the predictive validation approach can be identified. First, it is impossible to validate the whole FRAM model due to the overwhelming number of functions that have to be tested in a large and complex FRAM model. Only a few functions and their expected, as well as unexpected effects can be examined. Second, when manipulating one function, it is difficult to actually keep all of the remaining functions constant that were supposed to be constant, since the type of manipulation measure can potentially affect the performance of other functions. This problem is exacerbated by the fact that it is not even possible to check which functions this applies to, as it is impossible to analyse the performance variability for all functions. This results in interaction effects, whereby observed effects can no longer be fully attributed to the manipulated function. Furthermore, it might be difficult to find a targeted manipulation measure for each function in the model, e.g., for cognitive functions, since either no targeted manipulation is possible or several functions would be manipulated at the same time. Moreover, the extent and manner in which a manipulation has to be carried out to achieve the desired effect are generally unclear. Thus, following the method of constant stimuli from psychophysics (Fechner 1860), different stimulus intensities or types would have to be varied per manipulated function to see to which extent or manner a manipulation of an upstream function has to be carried out that results in a significant change in the performance variability of the individual downstream functions. Naturally, the extent and manner of the stimulus required vary between the individual downstream functions. Third, the performance variability of a downstream function may only change when several upstream inputs are varied instead of just the one manipulated function. Thus, an expected coupling could make sense and be valid even if no effect was observed in isolation. Consequently, all what-if combinations would have to be taken into account to be able to represent the complexity, which is simply impractical. Fourth, it is impossible to test whether there is also a direct influence for the functions that are indirectly influenced by the manipulated function. In addition, some functions are tested, where a direct influence by the manipulated function can be expected, and at the same time other functions that are also directly influenced by the manipulated function provide upstream inputs for the tested function. Hence, in these cases, there is always a degree of uncertainty as to whether the effect is direct or indirect.

5.3 Utility and applicability of the formal approach to assess predictive validity in FRAM

A research-practice gap of systemic models and methods (Underwood and Waterson 2012), especially FRAM, currently exists in literature, which means that researchers are presently applying systemic methods due to the current state-of-the-art and, in contrast, many practitioners press ahead with more traditional methods because of their ease of use or popularity despite known limitations (Grabbe et al 2022). Frequently mentioned reasons for this are a difficult and time-consuming application (Salmon et al. 2020), reduced model validation and usability, and a potential analyst bias (Underwood and Waterson 2012). Against this background, the results of the validation must be correlated to usability as a cost-effectiveness trade-off to be able to evaluate the utility benefit of the predictive validation approach in general (cf. Stanton and Young 2003). The effectiveness hereby represents the validity of the FRAM model to explain performance variability in an overtaking scenario, and the costs are related to the resources and time used by the method. As shown in Sect. 5.1, the validity is limited and can only be partly assumed. In contrast, the costs of using the method are high, since the model development by function identification and variability data collection was very time- and resource-consuming (Grabbe et al. 2022), something that also applies to the validation process. It should be noted that only two and not all of the functions of the model could be validated by this great effort. Therefore, the utility of the analysed FRAM model is questionable in terms of predictive validity if it is used as an evaluative method. On the other hand, the utility of the FRAM model as an analytical method is still an open question and difficult to demonstrate objectively.

In addition, and as shown in Sect. 5.2, there are several methodological issues related to the theoretical concept of the predictive validation approach for an FRAM model due to high complexity, leading to the conclusion that a complete validation of an FRAM model is impossible. Rather, the predictive validation approach developed in this paper should be applied to calibrate and not validate an FRAM model. This means that it can be used to select a few interesting functions in the model and to refine their modelling for a better understanding of their potential effects on the system behaviour with regard to specific system conditions, but not to prove that an FRAM model is valid or not. Consequently, the approach is appropriate to enhance any basic knowledge about system mechanisms gained by the FRAM model, but inappropriate to reach any final decisions concerning the approval of designs in safety–critical systems.

6 Conclusions and outlook

This paper developed a framework for evaluating the reliability and validity of an FRAM model, assessed the predictive validity of one specific FRAM model, and evaluated the applicability of this validation approach. The study shows that the validity and usefulness of the FRAM model by Grabbe et al. (2022) is limited and that the model results cannot be generalised to changing system conditions without any model adaptations. However, it is not clear whether this arises from the FRAM method itself or from the manner in which it was applied (cf. Stanton et al. 2013). Also, the applicability of the approach to demonstrate predictive validity is greatly reduced on account of several methodological limitations.

In future, the formal reliability and validity framework, and especially the predictive validation approach, should also be applied to other FRAM models in different application contexts so as to determine the reliability and validity generalisation of the FRAM method. Especially, human–machine dissonances and their predicted effects through an FRAM model should be validated. Moreover, the test–retest paradigm should be applied to rather small FRAM models to evaluate the reliability of the FRAM method and potential training effects in this context.

In conclusion, this paper contributes to making up for the lack of a formal validity approach for the FRAM method as well as to the research-practice gap of systemic HFE models and methods and their associated ongoing concerns of reliability and validity. In particular, this work helps analysts compare the cost-effectiveness of FRAM with other HFE methods. Overall, the developed framework provides a good foundation to evaluate the reliability and validity of an FRAM model. However, there is still potential for improvement and extension, especially against the background of the methodological advancement of FRAM and integration with other methods offering new opportunities for validation. Indeed, the reliability and validity framework can be used to calibrate rather than validate an FRAM model.

Acknowledgements The authors are grateful for all subjects who participated in this study. Furthermore, we would like to thank the editing service KERN for proofreading the paper. The authors are grateful to colleagues who provided valuable comments during the writing process of this paper.

Author contributions Conceptualisation, NG, AA; methodology, NG, AA; software, NG, AA; validation, NG, AA; formal analysis, NG, AA; investigation, NG, AA; resources, NG; data curation, NG, AA; writing—original draft preparation, NG; writing—review and editing, NG; visualisation, NG; supervision, KB; All authors have read and agreed to the published version of the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflicts of interest The authors have no relevant financial or non-financial interests to dis-close. In addition, no funds, grants or other support were received.

Institutional review board statement The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of the Ethics Committee of the Technical University of Munich (protocol code 509/21 S and date of approval 23 August 2021).

Informed consent Informed consent was obtained from all subjects involved in the study.

References

Accou B, Reniers G (2019) Developing a method to improve safety management systems based on accident investigations: the SAfety FRactal ANalysis. Saf Sci 115:285–293

Adriaensen A, Patriarca R, Smoker A, Bergström J (2019) A sociotechnical analysis of functional properties in a joint cognitive system: a case study in an aircraft cockpit. Ergonomics 62(12):1598–1616

Anfara VA Jr, Brown KM, Mangione TL (2002) Qualitative analysis on stage: making the research process more public. Educ Res 31(7):28–38

Annett J (2002) A note on the validity and reliability of ergonomics methods. Theor Issues Ergon Sci 3(2):228–232

Anvarifar F, Voorendt MZ, Zevenbergen C, Thissen W (2017) An application of the functional resonance analysis method (FRAM) to risk analysis of multifunctional flood defences in the Netherlands. Reliab Eng Syst Saf 158:130–141

Baber C, Stanton NA (1994) Task analysis for error identification: a methodology for designing error-tolerant consumer products. Ergonomics 37(11):1923–1941

Baber C, Stanton NA (1996) Human error identification techniques applied to public technology: predictions compared with observed use. Appl Ergon 27(2):119–131

Baber C, Young MS (2022) Making ergonomics accountable: reliability, validity and utility in ergonomics methods. Appl Ergon 98:103583

Balci O (1998) Verification, validation, and testing. Handb Simul 10(8):335–393

Banks J, Gerstein D, Searles SP (1987) Modeling processes, validation, and verification of complex simulations: a survey. In: 1987 SCS simulators conference, p 13–18

Baysari MT, Caponecchia C, McIntosh AS (2011) A reliability and usability study of TRACEr-RAV: the technique for the retrospective analysis of cognitive errors–for rail, Australian version. Appl Ergon 42(6):852–859

Blana E (1996) Driving simulator validation studies: a literature review. Working paper, Institute of Transport Studies, University of Leeds, Leeds, UK

Bridges KE, Corballis PM, Hollnagel E (2018) “Failure-to-Identify” hunting incidents: a resilience engineering approach. Hum Factors 60(2):141–159

Bulgren WG (1982) Discrete system simulation. Prentice Hall, Upper Saddle River

Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom 21(1):1–13

Cornelissen M, McClure R, Salmon PM, Stanton NA (2014) Validating the strategies analysis diagram: assessing the reliability and validity of a formative method. Appl Ergon 45(6):1484–1494

Crawford A (1963) The overtaking driver. Ergonomics 6(2):153–170

Creswell JW, Miller DL (2000) Determining validity in qualitative inquiry. Theory Pract 39(3):124–130

Dallat C, Salmon PM, Goode N (2017) Risky systems versus risky people: to what extent do risk assessment methods consider the systems approach to accident causation? A review of the literature. Saf Sci 119:266–279

Espié S, Gauriat P, Duraz M (2005) Driving simulators validation: the issue of transferability of results acquired on simulator. In: Driving simulation conference North-America (DSC-NA 2005), Orlondo, FL

Fechner GT (1860) Elemente der Psychophysik [elements of psychophysics]. Breitkopf und Härtel, Leipzig, pp 280–286

Ferreira PN, Cañas JJ (2019) Assessing operational impacts of automation using functional resonance analysis method. Cognit Technol Work 21:1–18

Goode N, Salmon PM, Taylor NZ, Lenné MG, Finch CF (2017) Developing a contributing factor classification scheme for Rasmussen’s AcciMap: reliability and validity evaluation. Appl Ergon 64:14–26

Grabbe N, Kellnberger A, Aydin B, Bengler K (2020) Safety of automated driving: the need for a systems approach and application of the functional resonance analysis method. Saf Sci 126:104665

Grabbe N, Gales A, Höcher M, Bengler K (2022) Functional resonance analysis in an overtaking situation in road traffic: comparing the performance variability mechanisms between human and automation. Safety 8(1):3

Hanssmann F (2018) Einführung in die Systemforschung. Oldenbourg Wissenschaftsverlag, Munich

Harris D, Stanton NA, Marshall A, Young MS, Demagalski J, Salmon P (2005) Using SHERPA to predict design-induced error on the flight deck. Aerosp Sci Technol 9(6):525–532

Hill R, Boult M, Sujan M, Hollnagel E, Slater D (2020) Predictive analysis of complex systems’ behaviour. https://www.researchgate.net/profile/David-Slater/publication/343944100_PREDICTIVE_ANALYSIS_OF_COMPLEX_SYSTEMS'_BEHAVIOUR_SWIFTFRAM/links/5f4907e0299bf13c5047f8d3/PREDICTIVE-ANALYSIS-OF-COMPLEX-SYSTEMS-BEHAVIOUR-SWIFTFRAM.pdf. Accessed 18 Nov 2021

Hollnagel E (2004) Barriers and accident prevention. Ashgate, Hampshire

Hollnagel E (2012) FRAM: the functional resonance analysis method: modelling complex socio-technical systems. CRC Press, Boca Raton

Hollnagel E (2014) Safety–I and safety–II: the past and future of safety management. CRC Press, Boca Raton

Hollnagel E (2020) FRAM model interpreter. https://functionalresonance.com/onewebmedia/FMI%20basicPlus%20V3.pdf. 09 Nov 2021

Hollnagel E, Hounsgaard J, Colligan L (2014) FRAM—the functional resonance analysis method—a handbook for the practical use of the method. https://functionalresonance.com/onewebmedia/FRAM_handbook_web-2.pdf. 17 Nov 2021

Hoskins AH, El-Gindy M (2006) Technical report: Literature survey on driving simulator validation studies. Int J Heavy Veh Syst 13(3):241–252

Hughes BP, Newstead S, Anund A, Shu CC, Falkmer T (2015) A review of models relevant to road safety. Accid Anal Prev 74:250–270

Hulme A, Stanton NA, Walker GH, Waterson P, Salmon PM (2019) What do applications of systems thinking accident analysis methods tell us about accident causation? A systematic review of applications between 1990 and 2018. Saf Sci 117:164–183

Hulme A, Stanton NA, Walker GH, Waterson P, Salmon PM (2021a) Testing the reliability and validity of Net-HARMS: a new systems-based risk assessment method in HFE. In: Congress of the International Ergonomics Association, Springer, Cham, p 354–362

Hulme A, Stanton NA, Walker GH, Waterson P, Salmon PM (2021b) Testing the reliability and validity of risk assessment methods in Human Factors and Ergonomics. Ergonomics 65:1–22

Hulme A, Stanton NA, Walker GH, Waterson P, Salmon PM (2021c) Are accident analysis methods fit for purpose? Testing the criterion-referenced concurrent validity of AcciMap, STAMP-CAST and AcciNet. Saf Sci 144:105454

ISO Standard 26262 (2018) Road vehicles—functional safety—part 3: concept phase. https://www.iso.org/standard/68385.html. Accessed 16 Dec 2021

Jensen A, Aven T (2018) A new definition of complexity in a risk analysis setting. Reliab Eng Syst Saf 171:169–173

Kaptein NA, Theeuwes J, Van Der Horst R (1996) Driving simulator validity: some considerations. Transp Res Rec 1550(1):30–36

Kaya GK, Ovali HF, Ozturk F (2019) Using the functional resonance analysis method on the drug administration process to assess performance variability. Saf Sci 118:835–840

Kirwan B, Kennedy R, Taylor-Adams S, Lambert B (1997) The validation of three human reliability quantification techniques—THERP, HEART and JHEDI: Part II—results of validation exercise. Appl Ergon 28(1):17–25

Laaraj N, Jawab F (2018) Road accident modeling approaches: literature review. In: 2018 International colloquium on Lo-1769 gistics and supply chain management (LOGISTIQUA). IEEE, p 188–193

Larsson P, Dekker SW, Tingvall C (2010) The need for a systems theory approach to road safety. Saf Sci 48(9):1167–1174

Larue GS, Wullems C, Sheldrake M, Rakotonirainy A (2018) Validation of a driving simulator study on driver behavior at passive rail level crossings. Hum Factors 60(6):743–754

Leveson N (2004) A new accident model for engineering safer systems. Saf Sci 42(4):237–270

Li W, He M, Sun Y, Cao Q (2019) A proactive operational risk identification and analysis framework based on the integration of ACAT and FRAM. Reliab Eng Syst Saf 186:101–109

Liebl F (2018) Simulation. Oldenbourg Wissenschaftsverlag, Munich

MacKinnon RJ, Pukk-Härenstam K, Kennedy C, Hollnagel E, Slater D (2021) A novel approach to explore safety-I and safety-II perspectives in in situ simulations—the structured what if functional resonance analysis methodology. Adv Simul 6(1):1–13

Makeham MA, Stromer S, Bridges-Webb C, Mira M, Saltman DC, Cooper C, Kidd MR (2008) Patient safety events reported in general practice: a taxonomy. BMJ Qual Saf 17(1):53–57

Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta (BBA) Protein Struct 405(2):442–451

Mullen N, Charlton J, Devlin A, Bedard M (2011) Simulator validity: behaviours observed on the simulator and on the road. In: Fisher DL, Rizzo M, Caird JK, Lee JD (eds) Handbook of driving simulation for engineering, medicine and psychology, 1st edn. CRC Press, Boca Raton, pp 1–18

Nemeth C (2013) Erik Hollnagel: FRAM: the functional resonance analysis method, modeling complex socio-technical systems. Cogn Technol Work 1(15):117–118

Nilsson L (1993) Behavioural research in an advanced driving simulator-experiences of the VTI system. In: Proceedings of the human factors and ergonomics society annual meeting, vol 37, no 9. Sage: Los Angeles, p 612–616

O'Connor P (2008) HFACS with an additional layer of granularity: validity and utility in accident analysis. Aviat Space Environ Med 79(6):599–606

Olsen NS (2013) Reliability studies of incident coding systems in high hazard industries: a narrative review of study methodology. Appl Ergon 44(2):175–184

Olsen NS, Shorrock ST (2010) Evaluation of the HFACS-ADF safety classification system: inter-coder consensus and intra-coder consistency. Accid Anal Prev 42(2):437–444

Patriarca R, Bergström J (2017) Modelling complexity in everyday operations: functional resonance in maritime mooring at quay. Cogn Technol Work 19(4):711–729

Patriarca R, Bergström J, Di Gravio G (2017) Defining the functional resonance analysis space: combining abstraction hierarchy and FRAM. Reliab Eng Syst Saf 165:34–46

Patriarca R, Di Gravio G, Woltjer R, Costantino F, Praetorius G, Ferreira P, Hollnagel E (2020) Framing the FRAM: a literature review on the functional resonance analysis method. Saf Sci 129:104827

Pereira AG (2013) Introduction to the Use of FRAM on the effectiveness assessment of a radiopharmaceutical dispatches process. In: International nuclear Atlantic conference

Pollatschek M, Polus A (2005) Modelling impatience of driver in passing manuevers. Transp Traffic Theory 16:267–279

Qureshi ZH (2007) A review of accident modelling approaches for complex socio-technical systems. In: Proceedings of the 1757 twelfth Australian workshop on Safety critical systems and software and safety-related programmable systems, vol 86. Australian Computer Society, Inc., p 1758 47–59

Rasmussen J (1997) Risk management in a dynamic society: a modelling problem. Saf Sci 27(2–3):183–213

Rastegary H, Landy FJ (1993) The interactions among time urgency, uncertainty, and time pressure. In: Time pressure and stress in human judgment and decision making. Springer, Boston, p 217–239

Ross A, Sherriff A, Kidd J, Gnich W, Anderson J, Deas L, Macpherson L (2018) A systems approach using the functional resonance analysis method to support fluoride varnish application for children attending general dental practice. Appl Ergon 68:294–303

Salehi V, Veitch B, Smith D (2021) Modeling complex socio-technical systems using the FRAM: a literature review. Hum Factors Ergon Manuf Serv Ind 31(1):118–142

Salmon PM, McClure R, Stanton NA (2012) Road transport in drift? Applying contemporary systems thinking to road safety. Saf Sci 50(9):1829–1838

Salmon PM, Read GJ, Walker GH, Stevens NJ, Hulme A, McLean S, Stanton NA (2020) Methodological issues in systems Human Factors and Ergonomics: perspectives on the research–practice gap, reliability and validity, and prediction. Hum Factors Ergon Manuf Serv Ind. https://doi.org/10.1002/hfm.20873

Sargent RG (1984) A tutorial on verification and validation of simulation models. In: Proceedings of the 16th conference on Winter simulation, pp 115–121. https://repository.lib.ncsu.edu/bitstream/handle/1840.4/4929/1984_0017.pdf?sequence=1

Schrank WE, Holt CC (1967) Critique of: "Verification of computer simulation models." Manag Sci 14(2):B-104

Sheridan TB (1992) Telerobotics, automation, and human supervisory control. MIT Press, Cambridge

Stanton NA (2014) Commentary on the paper by Heimrich Kanis entitled 'Reliability and validity of findings in ergonomics research': where is the methodology in ergonomics methods? Theor Issues Ergon Sci 15(1):55–61

Stanton NA (2016) On the reliability and validity of, and training in, ergonomics methods: a challenge revisited. Theor Issues Ergon Sci 17(4):345–353

Stanton NA, Baber C (2005) Validating task analysis for error identification: reliability and validity of a human error prediction technique. Ergonomics 48(9):1097–1113

Stanton NA, Stevenage SV (1998) Learning to predict human error: issues of acceptability, reliability and validity. Ergonomics 41(11):1737–1756

Stanton NA, Young MS (1999a) What price ergonomics? Nature 399(6733):197–198

Stanton NA, Young MS (1999b) A guide to methodology in ergonomics: designing for human use. Taylor & Francis, London

Stanton NA, Young MS (2003) Giving ergonomics away? The application of ergonomics methods by novices. Appl Ergon 34(5):479–490

Stanton NA, Salmon P, Harris D, Marshall A, Demagalski J, Young MS, Dekker S et al (2009) Predicting pilot error: testing a new methodology and a multi-methods and analysts approach. Appl Ergon 40(3):464–471

Stanton NA, Salmon PM, Rafferty LA, Walker GH, Baber C, Jenkins DP (2013) Human factors methods: a practical guide for engineering and design. CRC Press, Boca Raton

Stanton NA, Brown JW, Revell KM, Clark JR, Richardson J, Langdon P et al (2021a) Modelling automation-human driver interactions in vehicle takeovers using OESDs. Designing interaction and interfaces for automated vehicles. CRS Press, Boca Raton, pp 299–320

Stanton NA, Brown JW, Revell KM, Kim J, Richardson J, Langdon P et al (2021b) OESDs in an on-road study of semi-automated vehicle to human driver handovers. Cognit Technol Work 24:1–16

Stanton NA, Brown JW, Revell KM, Kim J, Richardson J, Langdon P et al (2021c) Validating OESDs in an on-road study of semi-automated vehicle-to-human driver takeovers. In: Designing interaction and interfaces for automated vehicles. CRC Press, Boca Raton, p 443–464b

Stanton NA, Brown JW, Revell KM, Langdon P, Bradley M, Politis I et al (2021d) Validating operator event sequence diagrams: the case of automated vehicle-to-human driver takeovers. In: Designing interaction and interfaces for automated vehicles. CRC Press, Boca Raton, p 137–157

Tapio J (2003) Ohitukset kaksikaistaisilla teilla (Summary in English). Finnish Road Administration, Helsinki

Underwood P, Waterson P (2012) A critical review of the STAMP, FRAM and Accimap systemic accident analysis models. In: Advances in human aspects of road and rail transportation. CRC Press, Boca Raton, pp 385–394

Van Horn RL (1971) Validation of simulation results. Manag Sci 17(5):247–258

Vanderhaegen F (2014) Dissonance engineering: a new challenge to analyse risky knowledge when using a system. Int J Comput Commun Control 9(6):776–785

Vanderhaegen F (2016) A rule-based support system for dissonance discovery and control applied to car driving. Expert Syst Appl 65:361–371

Vanderhaegen F (2021) Heuristic-based method for conflict discovery of shared control between humans and autonomous systems—a driving automation case study. Robot Auton Syst 146:103867

Wege CA, Pereira M, Victor TW, Krems JF, Stevens A, Brusque C (2014) Behavioural adaptation in response to driving assistance technologies: a literature review. Driver adaptation to information and assistance systems. The Institution of Engineering and Technology, London, p S.3-34

Wienen HCA, Bukhsh FA, Vriezekolk E, Wieringa RJ (2017) Accident analysis methods and models—a systematic literature 1767 review. In: Centre for Telematics and Information Technology (CTIT), p 1768

Woltjer R, Hollnagel E (2008) Functional modeling for risk assessment of automation in a changing air traffic management environment. In: Proceedings of the 4th international conference working on safety, vol 30

Yang Q, Tian J, Zhao T (2017) Safety is an emergent property: illustrating functional resonance in air traffic management with formal verification. Saf Sci 93:162–177

Yang Q, Tian J (2015) Model-based safety assessment using FRAM for complex systems. In: Proceedings of the 25th European safety and reliability conference

Zhu Q (2020) On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. Pattern Recognit Lett 136:71–80

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.