Are accident analysis methods fit for purpose? Testing the criterion-referenced concurrent validity of AcciMap, STAMP-CAST and AcciNet

1.0 Introduction

The field of accident analysis has evolved considerably over the past century. A sizable body of work has described the development of accident analysis theory, models and methods from a historical perspective (Hollnagel 2004; Salmon et al. 2011; Khanzode et al. 2012; Leveson 2016; Waterson et al. 2017; Stanton et al. 2019). Within this body of work, four accident analysis generations characterised by various techniques and approaches are outlined: (i) accident proneness theory and behaviourism (e.g., Greenwood and Woods 1919); (ii) linear cause-effect and sequential modelling techniques (e.g., Heinrich 1931); (iii) epidemiological methods representing the influence of organisational and individual factors (e.g., Reason 1990; Shappell and Wiegmann 2001); and, (iv) systems-based approaches that aim to model the relationships among factors across multiple sociotechnical system levels (e.g., Rasmussen 1997; Leveson 2004). Systems-based methods currently sit at the cutting-edge of human factors and ergonomics (HFE) and safety science efforts (Leveson 2016; Salmon et al. 2017; Hulme et al. 2019; Stanton et al. 2019). These approaches consider accidents to be emergent phenomena arising from the complexity of systems whose behaviour cannot be predicted from an analysis of their constituent parts in isolation (Lindberg et al. 2010). The goal of systems-based accident analysis methods is to understand the systemic and interacting causes of adverse incidents as a basis to effectively prevent their recurrence and facilitate organisational learning (Salmon et al. 2011).

A number of systems-based accident analysis methods dominate the HFE and safety science literature (Hulme et al. 2019). The most widely used include the Accident Mapping (AcciMap) (Rasmussen 1997; Rasmussen and Suedung 2000) method, the Systems-Theoretic Accident Model and Processes approach and associated Causal Analysis based on Systems Theory (STAMP-CAST) (Leveson 2004) method, the Functional Resonance and Analysis Model (FRAM) (Hollnagel 2012), and more recently, the Accident Network (AcciNet) (Salmon et al. 2020b) method. These methods were developed to keep pace not just with the increasing complexity and interconnectedness of modern-day sociotechnical systems (Salmon et al. 2011; Underwood and Waterson 2013; Hulme et al. 2019; Stanton et al. 2019), but also to cope with those problems left over from deterministic methods that

have reached their limits (Walker et al. 2010). Indeed, within the last two decades, the introduction of new technologies and advanced forms of automation to support human work has profoundly influenced how accidents are conceptualised, investigated and analysed (Leveson et al. 2009; Leveson 2011; Salmon et al. 2017). Despite the widespread application of state-of-the-art systems-based models and methods in various domains (Hulme et al. 2019), pressing questions remain around their capacity to accurately describe accident aetiology in all its forms (Katsakiori et al. 2009; Underwood and Waterson 2013; Ahmadi et al. 2019; Goncalves Filho et al. 2019; Hulme et al. 2020). A useful method should provide repeatable results and help even relatively novice analysts deliver outcomes that are comparable with experts (Stanton and Young 2003). This requires renewed attention to the critical issues of reliability and validity of HFE and safety science methods (Stanton and Young 1999; Kanis 2000; Annett 2002; Stanton and Young 2003; Olsen 2013; Kanis 2014; Stanton 2014; Waterson et al. 2015; Shorrock and Williams 2016; Stanton 2016; Goode et al. 2017; Salmon et al. 2020c; Thoroman et al. 2020; Hulme et al. 2021b; Hulme et al. 2021c).

In the context of accident analysis research, examining the reliability and validity of systems-based methods provides objective evidence around their practical utility and potential to enhance organisational safety (Shorrock and Williams 2016; Goncalves Filho et al. 2019; Goncalves Filho et al. 2021). Reliability testing measures the stability of a method over time, either within (intra-rater) or between (inter-rater) analysts (Olsen 2013). Reliability testing therefore points to whether the results following the application of a method are consistently obtained on repeat occasions. Validity is more nuanced, with the form of validity testing undertaken depending on the goal of the study and analysis. Construct and content validity are concerned with the theoretical basis and comprehensiveness or credibility of a method in the eyes of experts, respectively (Stanton 2016). Conversely, criterion-referenced validity assesses the extent to which the results produced by a method agree with (i.e., concurrent validity) or predict over time (i.e., predictive validity) an external criterion, such as an expert or gold standard analysis (Stanton 2016; Goode et al. 2017; Thoroman et al. 2020; Hulme et al. 2021c). Criterion-referenced concurrent validity is arguably the most interesting and applicable measure for evaluating the accuracy of systems-based accident analysis methods as it reflects the

extent to which analysis outputs represent what actually happened as determined by an expert or group of experts. This includes the degree to which contributory factors and relationships are successfully identified in relation to an accepted standard or reference.

Despite the importance of determining whether methods produce stable and accurate results, few studies have tested the reliability and validity of systems-based accident analysis methods. A limited number of studies have compared methods in terms of their application time, data requirements, prerequisite expertise, model complexity, supporting taxonomic features, and/or coverage of systems theory and concepts (Johnson and de Almeida 2008; Katsakiori et al. 2009; Salmon et al. 2012; Underwood and Waterson 2014; Stanton et al. 2019). However, only one study has attempted to formally examine the reliability and validity of systems-based accident analysis methods (Goncalves Filho et al. 2019). The study by Goncalves Filho (2019) compared AcciMap and STAMP-CAST using percentage overlap of the identified casual factors as the main measure of inter-rater stability and validity. This approach does not consider the potential for false positives, false negatives or true negatives when determining relative agreement between analyst results and methods. Further, the validity of AcciMap and STAMP-CAST was extrapolated based on the similarity between results obtained from different methods (Goncalves Filho et al. 2019), rather than against an expert or gold standard criterion which represents the ideal analytical option.

In consideration of the lack of reliability and validity work in the area of accident analysis research, including the need for robust scientific study designs and sophisticated statistical approaches, the purpose of this study is to test the criterion-referenced concurrent validity of AcciMap, STAMP-CAST and AcciNet. A comparison of the new AcciNet method with more established approaches echoes the need for early testing in a method's lifecycle which is seldom if ever undertaken (Shorrock and Williams 2016; Stanton 2016; Salmon et al. 2020c; Hulme et al. 2021c). Testing new methods in HFE and safety science highlights areas of potential methodological refinement to support future applications. Before describing the methods used in this study, the following two sections will outline the ongoing need for reliability and validity testing in HFE and safety science, as well as provide a short background to the three included systems-based methods.

2.0 Reliability, validity and systems-based accident analysis methods

2.1 The ongoing need for reliability and validity testing

Various authors have put forward many reasons that explain the lack of reliability and validity testing in HFE and safety science, including difficulties around identifying and recruiting large enough analyst samples; the challenge of incorporating the necessary subject matter expertise; the need for an expert or detailed gold standard reference; the resource intensive and time consuming nature of test-retest study designs and data analyses (e.g., the data in this study took over four months to clean, format and analyse); and, the limited knowledge around which analytical and statistical approach to select for computing the corresponding reliability and validity measures (Shorrock and Williams 2016; Stanton 2016; Goncalves Filho et al. 2019; Salmon et al. 2020c; Hulme et al. 2021c). Aside from these common logistical considerations, another and perhaps more discouraging reason is that reliability and validity testing in eyes of many may simply be considered a relatively dry and uninspired science; one that demands the highest degree of groundwork for the smallest perceived payoff when it is said that our methods can 'work' anyway (Stanton and Young 1999; Salmon et al. 2020c). For instance, in terms of systems-based accident analysis methods, it has been argued that reliability and validity is less important when analysing isolated incidents as the output produced is still considered informative and useful (Waterson et al. 2017). In other words, the application of systems-based methods can identify potentially effective safety enhancing interventions regardless of their stability and accuracy. If this is indeed the case, why is there even a need to invest excessive amounts of time testing and comparing methods? Whilst such claims may be persuasive on the surface, deeper issues of validity begin to emerge when outputs are aggregated and synthesised as a basis to formulate recommendations to improve system safety (Goode et al. 2017), as recent studies have shown (Salmon et al. 2020a). Likewise, it has been noted that poorly performing accident analysis methods can provide misleading and potentially dangerous outputs (Salmon 2016).

From a research-based perspective, the reliability and validity of models and methods directly affects all other aspects of the scientific process, including the internal validity of experiments and the generalizability of results across geographical locations, populations, domains and issues. Further, the

accumulation of evidence and knowledge around the reliability and validity of HFE and safety science methods can inform their selection for different problem types; it can increase confidence in their use, and, provide data as to whether their application will produce results that meaningfully impact practice (Stanton and Young 1999). Reliability and validity testing is also valuable for understanding discrete aspects of methods that could benefit from refinement, which in turn sharpens predictions, increases the accuracy of modelled outputs and ultimately informs the development of the most optimal interventions under the specific circumstances (Hulme et al. 2021c). Stanton and Young (2003) note that the reliability and validity of HFE and safety science methods is inextricably linked to their utility from the perspective of a cost-benefit analysis. This works by calculating the cost of applying the method (in terms of person-hours, resources, etc.) and subtracting this cost from the estimated savings generated as a result of the changes implemented in practice. As such, the potential economic savings and benefits made to systems may be constrained to a greater or lesser degree if methods have failed to demonstrate a minimally acceptable level of validity (Stanton and Young 2003). From this and in light of the many reasons outlined above, it is concluded that there is a need for reliability and validity testing of accident analysis methods in HFE and safety science.

2.2 Accident analysis methods overview

The following section contains a short background to the three systems-based methods and justifies their inclusion in this study. Further information and guidance around each method can be found in the source material (AcciMap (Rasmussen 1997; Rasmussen and Suedung 2000; Svedung and Rasmussen 2002); STAMP-CAST (Leveson 2004; Leveson et al. 2009; Leveson 2011); AcciNet (Salmon et al. 2020b)).

2.2.1 AcciMap

An overview of the AcciMap method requires a brief background to Rasmussen's (1997) Risk Management Framework (RMF). This is because AcciMap uses the RMF as its underpinning theoretical basis. The RMF explains that sociotechnical systems comprise various hierarchical levels (e.g., government, regulators, company, management, staff and work), each of which contain actors,

organisations and technologies that share responsibility for production and safety. Decisions and actions occurring across levels of the system interact to shape behaviour, demonstrating that performance and safety is influenced by all elements in a system (Rasmussen 1997). Rasmussen and Svedung (2000) outlined the AcciMap technique which draws upon the RMF and graphically represents the system-wide failures, decisions and actions involved in accidents (Waterson et al. 2017). The AcciMap output is a model of the contributory factors and their interrelationships across the system. AcciMap has been applied in a diverse range of safety-critical domains (Hulme et al. 2019), including healthcare, aerospcace, led outdoor recreation, transportation (i.e., road, rail and aviation), emergency response, maritime and civial engieering. Recent studies have synthesised multiple AcciMap outputs to identify common patterns of accident aetiology (Salmon et al. 2020a) and have used the method to highlight how various systems thinking theoretical principles contribute to accident causation (Hulme et al. 2020).

2.2.2 STAMP-CAST

The Systems-Theoretic Accident Model and Processes (STAMP) model (Leveson 2004) takes the view that accidents result from the inadequate control or enforcement of safety-related constraints, when disturbances, failures and/or dysfunctional interactions between components are not handled by existing control mechanisms. STAMP describes safety as a control and feedback issue that is managed through a hierarchical control structure, with the primary goal of enforcing constraints on the actors, organisations and technologies across the sociotechnical system. Various forms of control are considered, including legislative, organisational, managerial, operational, manufacturing and social controls (Leveson et al. 2009). The overall performance and safety of systems is therefore dictated not only by appropriately designed and engineered components, but also by policies, procedures, shared values and other aspects of the organisational culture. The STAMP model has associated risk and hazard assessment (i.e., Systems Theoretic Process Analysis; STPA) and accident analysis (i.e., Casual Analysis based on Systems Theory; CAST) methods. When used for accident analysis purposes, applying CAST involves developing a control structure model of the system under analysis, and then using the associated taxonomy to identify control and feedback failures that

contributed to the accident. Leveson's (2004) classification of control flaws includes failures related to: (i) the inadequate enforcement of safety constraints (control actions); (ii) the inadequate execution of control actions; and, (iii) inadequate or missing feedback.

2.2.3 AcciNet

AcciNet (Salmon et al. 2020b) is a new accident analysis method that is based on three fundamental tenets of accident causation. First, that accidents are created due to an interacting network of human and technological factors (e.g., Dekker 2011); second, that the interacting network of causal behaviours involved in accidents includes both normal performances and failure (e.g., Hollnagel 2012); and third, that emergent behaviours play a critical role in accident causation (e.g., Leveson et al. 2009). The capability of AcciNet to incorporate the interaction between work undertaken in line with procedures with abnormal or suboptimal tasks and behaviours is distinctive feature of the approach relative to other systems-based accident analysis methods. Methodologically, AcciNet uses a task network that conceptualises the interrelationships between various key tasks within a system (Stanton et al. 2018). Analysts identify the contributory factors and instances of normal performances and failure underpinning an adverse incident of interest (e.g., obtained from in-situ data, organisational data and information systems, official investigation reports), and then code each factor using the AcciNet classification scheme. The output visualises a network of contributory factors that includes the interaction between instances of normal performance and failure.

As a point of reference, AcciNet was developed as part of a new integrated safety management toolkit that contains both prospective risk assessment (i.e., the Networked Hazard Analysis and Risk Management System; Net-HARMS) (Dallat et al. 2018; Hulme et al. 2021a) and retrospective incident analysis (i.e., AcciNet) methods. In a similar manner to the RMF (Rasmussen 1997; Svedung and Rasmussen 2002) and control structure model in STAMP (Leveson 2004), the development and application of a task network supports the incident analysis activities contained within the new integrated toolkit. The task network is an essential component of the analysis and is used to identify the actors, organisations and contributory factors within the boundary of the work system describe (see Stanton et al. 2018 for further information on network analysis methods).

2.3 Study aims and hypotheses

From the above it is concluded that there is a clear a pressing need for more reliability and validity work as it is applied to systems-based accident analysis methods. The three accident analysis methods described, and that are subjected to the validity testing forthcoming, sit at the cutting-edge of the systems-based safety era. AcciNet, which is the most recent method out of the three, requires further formal independent examination as well as a comparison against established approaches to evaluate its capacity in supporting analysts with the accurate modelling of adverse incidents. Thus, the purpose of this study, which is to test the criterion-referenced concurrent validity of AcciMap, STAMP-CAST and AcciNet, will bring us closer in being able to address the gaps identified. The criterion-referenced concurrent validity of AcciMap, STAMP-CAST and AcciNet was not known *a priori*. The null hypotheses (H_0) was therefore stated such that there would be no difference in validity between the three methods or their individual phases – each would perform equally against the expert benchmark (see section 3.5). The alternate hypothesis (H_1) was that a difference would be found between the three methods or phases.

3.0 Methods

3.1 Study design

A test-retest workshop study design with HFE and safety experts was used to evaluate AcciMap, STAMP-CAST and AcciNet. Participants received training in all three methods and used them to perform an accident analysis on a fatal road collision (section 3.4.2 for further detail) on two separate occasions separated by four weeks (Time one (T1) and Time two (T2)). Two time points were used to establish more accurate validity measures through repeat exposure. The accuracy of the analyses, and thus the criterion-referenced concurrent validity of AcciMap, STAMP-CAST and AcciNet, was determined by comparing the participant analyses against an expert analysis, both at T1 and T2. Validity measures are highest when the participant and expert analyses were the same. Intra-rater and inter-rater reliability assessments were beyond the current scope and will be the focus of future work.

The methods and statistical techniques reported in this study follow the same approach to a recently published systems-based risk assessment reliability and validity study (Hulme et al. 2021c).

Data collection was undertaken in two separate locations, on the campus of the University of the Sunshine Coast (USC) and in Brisbane, Queensland, Australia. Two members of the research team (AH, PS) conducted a total of four, one-day workshops over the course of October and November 2020 at the above locations. There were two workshops at T1, one held at USC and one in Brisbane during October 2020. The same workshop format was repeated one month later at T2 during November 2020. Ethical approval was granted by the USC Human Ethics Research Committee (A201374).

3.2 Participant identification and recruitment

To be eligible to take part in the study, participants were required to work in HFE and/or in a safety-related occupation. Both researchers and industry professionals were invited to take part. There was no requirement to have used the methods prior to the workshops as training was provided as part of the study design. Participants were identified through existing research networks and advertisements placed on social media platforms (i.e., LinkedIn, Twitter) specifying the eligibility requirements. E-mails were sent to various Australian organisations who employ HFE professionals with a role specialising in organisational health and safety management. Follow-up e-mails were sent two weeks later in the case of no response.

During the months of September and October 2020, potential participants completed a pre-screening questionnaire to solicit further information about their eligibility and workshop availability. There was no online component to the study. Out of a total of 71 potential participants who were initially contacted, 48 (67.6%) returned the pre-screening questionnaire. Further information about the study, including dates, locations and the materials required on the day of the first workshop, were sent to each registered participant via e-mail. Of the 48 participants who registered for the study, 43 (89.5%) attended the T1 workshop. In the four-week period between T1 and T2, seven participants withdrew.

Missing and incomplete data were discovered in two cases, and so a total of 34 participants were included in the T1 and T2 analyses.

3.3 Participant demographics

The demographics of the 34 participants who completed both phases of the study are presented in Table 1.

[insert Table 1 about here]

The primary occupation/profession, qualification, work sector and context, and length of time employed in the current role of the 34 participants can be viewed in Table 2.

[insert Table 2 about here]

3.4 Materials

3.4.1 Method workshops

Electronic spreadsheets, analysis templates and guidance documents to support the application of all three accident analysis methods were made available to participants on the morning of the workshops (Electronic Supplementary Materials; ESM 1-3). Analyses were performed using a personal computer with spreadsheet processing software. The workshops at T1 and T2 were designed to be identical. At the start of the T1 workshop, two members of the research team (AH, PS) delivered a presentation covering an introduction to systems thinking applied to safety management and accident analysis. Participants then received training in the use of AcciMap, STAMP-CAST and AcciNet, including a background to the methods, an overview of existing applications and were allocated 30 minutes per method to undertake a practice analysis. The practice analyses involved 10 groups of

three to four participants applying each method to analyse the cause of a notable maritime disaster (the capsizing of MS Herald of Free Enterprise) (MAIB 2015). During each practice session, participants were free to ask questions and seek further clarification around each method if required, particularly as each method requires fewer or more steps to apply. The workshop could only progress to the validity exercise when all participants indicated that they understood and felt confident that they could independently use a given method. The three one-hour validity analyses were performed under exam-style conditions and immediately followed each practice session (i.e., AcciMap training, then AcciMap practice followed by the Uber-Volvo analysis). Participants were not permitted to work together, and questions were limited to aspects of the method. Upon completion of all data collection activities, a total of 204 hours' worth of analyses had been performed.

3.4.2 Case study example

The Uber-Volvo incident which occurred on 18th March 2018 in Arizona, USA, was selected as the case study example (NTSB 2018). The incident involved a collision between an Uber-Volvo automated vehicle and a pedestrian, resulting in a single fatality. This incident was chosen as, first, there is comprehensive information available online and in the official investigation report (NTSB 2018); second, the accident is compatible with all the systems methods under examination as it contained multiple systemic failures; and third, the models, networks and analyses of the incident were used to generate the expert analysis (Royal Automobile Club (RAC), 2019; Stanton et al. 2019; Salmon et al. 2020b). A detailed written description of the Uber-Volvo incident was provided to each participant, and a short presentation was delivered during the workshops covering the critical events leading up to the fatality. Participants were not expected to identify the contributory factors from the official Uber-Volvo collision incident report to support their analyses. This is a routine component of all (indeed any) method. Any influence of this generic step needed to be removed in order to isolate the effects of the methods themselves. The contributory factors were, therefore, provided to each participant and served as an experimental control.

3.4.3 AcciMap materials and procedure

Materials to support the AcciMap validity analysis included an ActorMap of the Uber-Volvo system, an eight level RMF handout, a document containing a list of 37 contributory factors, and electronic spreadsheet template. Eight system levels were represented as the expert analysis contained an additional two levels beyond the traditional six, including 'National committees' and 'International influences'. Participants used the electronic spreadsheet template to record each contributory factor (from the list provided) under what they considered to be the correct level of the Uber-Volvo system. Following this, participants placed all factors onto the RMF handout as a basis to physically draw relationships between them. The 72 AcciMap outputs obtained following T1 and T2 were individually coded and reproduced electronically to support the validity comparison. The AcciMap validity exercise therefore contained two phases: (i) the contributory factor assignment phase; and, (ii) the relationship description phase. ESM1 includes the AcciMap participant and expert analysis templates.

3.4.4 STAMP-CAST materials and procedure

Materials to support the STAMP-CAST validity analysis included a control structure model of the Uber-Volvo system, the CAST classification scheme of control flaws, a document containing a list of 17 contributory factors (from report: RAC, 2019), and an electronic spreadsheet template. The first three columns of the electronic spreadsheet template were prefilled with the agents and the control/feedback mechanisms from the control structure model, and participants were required to assign a contributory factor (from the list provided) and then classify it into a control or feedback flaw using the CAST classification scheme. The STAMP-CAST validity exercise therefore contained two phases: (i) the contributory factor assignment phase; and, (ii) the factor classification phase. ESM2 includes the STAMP-CAST participant and expert analysis templates.

3.4.5 AcciNet materials and procedure

Materials to support the AcciNet validity analysis included a task network of the Uber-Volvo system, the AcciNet classification scheme, a document containing a list of 35 contributory factors (refined from AcciMap), and an electronic spreadsheet template. The first two columns of the electronic

spreadsheet template were prefilled with the tasks and agents from the task network, and participants were required to assign each of the 35 contributory factors to the tasks indicated and then classify them using the AcciNet classification scheme. The AcciNet validity exercise therefore contained two phases: (i) the contributory factor assignment phase; and, (ii) the factor classification phase. ESM3 includes the AcciNet participant and expert analysis templates.

3.5 Expert analyses

The expert AcciMap, STAMP-CAST and AcciNet analyses were obtained from published work that contained the models, networks and analyses required to perform a criterion-referenced concurrent validity assessment (NTSB 2018; Stanton et al. 2019; Salmon et al. 2020b).

3.6 Signal Detection Theory

The Signal Detection Theory (SDT) paradigm (Green and Swets 1966) was used to assess the criterion-referenced concurrent validity of AcciMap, STAMP-CAST and AcciNet. SDT uses a confusion matrix to organise the collection of binary classification data (Figure 1). The two dimensions in Figure 1 contain the predicted and observed data for two identical sets of classes.

[insert Figure 1 about here]

To support the SDT paradigm, each participant's contributory factor assignment results and relationship descriptions (in the case of AcciMap); and, contributory factor assignment results and factor classifications (in the case of STAMP-CAST and AcciNet), were transposed adjacent to the corresponding expert analysis using electronic spreadsheet software. The SDT paradigm followed the approach of 'does the participant analysis predict the expert analysis'? Thus, if results were identical between analyses, a hit was recorded using the SDT paradigm (i.e., a true positive). If results were different, both a false alarm (predicted by the participant but not in the expert referent analysis) and miss (not predicted by the participant but in the expert referent analysis) were recorded, referred to as

a false positive and false negative, respectively. The exact number of correct rejections differed between each method. For example, in the case of AcciNet, there were 19 tasks. If one of the 19 tasks was correctly classified, then 18 tasks were correctly rejected. Likewise, there are 10 items in the AcciNet classification scheme, and so for each correct factor classification, there were nine correct rejections. Correct rejections are analogous to true negatives.

Based on the SDT confusion matrix, the hit rate (HR), or sensitivity of a given accident analysis method (Equation 1), and the false alarm rate (FAR) (Equation 2), were computed.

Eq. 1

$$HR = \frac{H}{H + M}$$

Eq. 2

$$FAR = \frac{FA}{FA + CR}$$

HR and FAR are expressed along a standardised scale ranging from 0.0 to 1.0. A higher and lower HR and FAR, respectively, indicates a greater level of agreement between a participant's analysis and the expert analysis. Reliability and validity studies and reviews have applied and identified thresholds in the region of 70-80% to indicate an acceptable level of agreement between predicted and observed data (Olsen 2013; Goode et al. 2017; Thoroman et al. 2020). Accordingly, the following thresholds were used, with >80% and <20% indicating the acceptable levels of agreement between participant analyses and the expert analysis for HR and FAR, respectively:

- <0.2 is a low HR/FAR;
- 0.2 – 0.4 is low to moderate HR/FAR;
- 0.4 – 0.6 is a moderate HR/FAR;
- 0.6 – 0.8 is moderate to high HR/FAR; and,
- >0.8 is a high HR/FAR.

Whilst HR and FAR are useful for evaluating the accuracy of a single prediction, such rates remain limited insofar as they only account for two of four categories in the confusion matrix (Chicco and Jurman 2020). As a means of considering all four values in the confusion matrix, Matthews Correlation Coefficient (MCC) (Matthews 1975), a special case of the ϕ (phi) coefficient, analysed the complete SDT data (Equation 3).

Eq. 3

$$MCC = \frac{H\ x\ CR - FA\ x\ M}{\sqrt{(H + FA)(H + M)(CR + FA)(CR + M)}}$$

MCC measures the strength of a correlation between the predicted positive and negative values, in this case the participant analyses at T1 and T2, and the observed positive and negative values, in this case the expert analysis. MCC is thus a reliable statistical approach that produces a relatively 'high' score, albeit only if the prediction obtains respectable scores in all four matrix categories (Chicco and Jurman 2020). Like other correlation coefficients, including Pearson product-moment correlation coefficient, intraclass or rank, MCC is normalised on a scale from -1.0 to +1.0 (Taylor 1990), where:

- +1.0/-1.0 is a perfect positive/ideal negative correlation;
- +0.8/-0.8 is a strong positive/negative correlation;
- +0.5/-0.5 is a moderate positive/negative correlation;
- +0.2/-0.2 is a weak positive/negative correlation; and,
- 0.0 means no relationship between a set of variables or responses.

The criterion-referenced concurrent validity of AcciMap, STAMP-CAST, and AcciNet is higher as the MCC score approaches positive 1.0.

3.7 Data presentation

The results of the comparative analyses, including HR, FAR and MCC, were graphed using box plots. Viewed in Figure 2, the three methods each contained two phases: (i) the contributory factor assignment phase; and, (ii) the relationship description/factor classification phase, of which a further four individual comparisons were made per phase between participant results and the expert analysis

using the SDT paradigm. This produced a total of eight individual HR and FAR analyses for each method. Next, the individual contributory factor assignment and relationship description/factor classification phases at T1 and T2 were combined. Combining method phases was used to evaluate the overall HR and FAR of each method at both time points. Finally, MCC was produced by combining HR with FAR at T1 and T2, thus indicating the criterion-referenced concurrent validity of the methods.

It is noted that the relationship description phase of AcciMap is not directly comparable with the second phase of STAMP-CAST and AcciNet, the latter two methods of which are similar as they both attempt to classify factors with a classification scheme. Nevertheless, including a comparison between and across individual method phases, as well as applying HR and FAR metrics to these phases, was necessary to understand the relative contribution of these component phases and metrics to the overall MCC validity scores.

[insert Figure 2 about here]

3.8 Data analysis

Paired sample t-tests were used to determine differences in HR and FAR between the contributory factor assignment and relationship description/factor classification phases of AcciMap, STAMP-CAST and AcciNet. One-way repeated measures analysis of variance (ANOVA) tests were used to determine differences in: (i) HR and FAR between the three methods when the contributory factor assignment and relationship description/factor classification phases were combined at T1 and T2; and, (ii) MCC between the three methods when HR and FAR were combined at T1 and T2. HR, FAR and MCC represented the continuous dependent variable and the accident analysis method as the categorical independent variable across three levels. The significance of Mauchly's test of sphericity was evaluated during each ANOVA run. Sphericity was assumed if $p = >.05$, otherwise the Greenhouse-Geisser correction test was selected to avoid distorting variance calculations. Skewness

and kurtosis z-values, the significance of the Shapiro-Wilk test and histograms were investigated prior to analyses to check that data were approximately normally distributed. In case of unequal data distribution following the above assumption checks, or should outliers be detected as a result of natural response variability, non-parametric tests included the Wilcoxon signed-rank test and Friedman's repeated measures comparison test. The Bonferroni correction was applied to all post-hoc tests. The mean (*M*) and standard deviation (*SD*) were presented for normally distributed data, whereas the median (Mdn) and interquartile range (IQR) were reported for data that were non-normally distributed. A value of $p = <.05$ was considered significant for all tests. Analyses were undertaken using the IBM SPSS statistics package for Windows (version 26, 2019).

The HR, FAR and correlation coefficient thresholds (section 2.6) are integral to the interpretation of any observed differences between methods and phases. Statistical significance testing only informs of whether the result is real or due to chance and is influenced by sample size. Thus, the thresholds and correlation limits point to how important the results are in a practical sense. HR, FAR and MCC are expressed along an intuitive and standardised scale and so interpretations of differences in effect size were based on the original units of measurement. The unstandardized effect size categories for HR and FAR were defined as negligible (<.05), small (.05 – .10), moderate (.10 – .15) and large (≥.15).

4.0 Results

4.1 Participant expertise ratings

Figure 3 shows the self-reported level of expertise within the participant group across nine dimensions prior to receiving methods training. A similar trend in the self-reported level of expertise across AcciMap, STAMP-CAST and AcciNet is observed, with a majority of participants having no experience to a low level of experience. This was pronounced for AcciNet. As the more established of the techniques, AcciMap and STAMP-CAST did not fare much better than AcciNet in terms of the self-reported levels of expertise prior to the provision of training. For example, only two (6%) participants self-reported a medium or high level of expertise in STAMP-CAST and 24 (71%) reported none or a low level of expertise in AcciMap. Over two-thirds (23; 68%) of participants in

the study sample reported a medium to high level of expertise in systems thinking/general systems theory and 20 (59%) reported a medium to high level of expertise in accident causation and analysis. This was to be expected as all participants worked in HFE and/or in a safety-related occupation.

[insert Figure 3 about here]

4.2 Individual phase within method comparison

4.2.1 AcciMap

Figure 4 shows HR and FAR at T1 and T2 for the individual phases of AcciMap. HR was moderate to high for the contributory factor assignment phase at T1 and T2, and low to moderate for the relationship description phase at T1 and T2. Differences between phases were large (Figure 4; panel A and B). A Wilcoxon signed rank test found a significant difference ($Z = 5.087$; $p = <.001$) in HR between the contributory factor assignment (Mdn [IQR] = .70 [.11]) and relationship description (Mdn [IQR] = .30 [.22]) phase of AcciMap at T1. There was no instance by which HR for the relationship description phase was higher relative to the paired HR for the contributory factor assignment phase (positive ranks). A paired sample *t*-test also found a significant difference ($t(33) = 9.933$; $p = <.001$) in HR between the contributory factor assignment (*M* [SD] = .64 [.09]) and relationship description (*M* [SD] = .40 [.14]) phase at T2.

FAR was low for both phases of AcciMap at T1 and T2. Differences between phases were negligible (Figure 4; panel C and D). No significant difference ($Z = .432$; $p = .665$) was found in FAR between the contributory factor assignment (Mdn [IQR] = .06 [.03]) and relationship description (Mdn [IQR] = .04 [.05]) phase at T1. Conversely, a significant difference ($Z = 4.788$; $p = <.001$) was found in FAR between the contributory factor assignment (Mdn [IQR] = .07 [.02]) and relationship description (Mdn [IQR] = .04 [.04]) phase at T2.

[insert Figure 4 about here]

4.2.2 STAMP-CAST

Figure 5 shows HR and FAR at T1 and T2 for the individual phases of STAMP-CAST. HR was moderate to high for the contributory factor assignment phase at T1 and T2, and moderate and low to moderate for the factor classification phase at T1 and T2, respectively. Differences between phases were large (Figure 5; panel A and B). A Wilcoxon signed-rank test found a significant difference (Z = 4.470; p = <.001) in HR between the contributory factor assignment (Mdn [IQR] = .71 [.12]) and factor classification (Mdn [IQR] = .45 [.18]) phase at T1. A significant difference (Z = 5.059; p = <.001) was also found between the contributory factor assignment (Mdn [IQR] = .74 [.15]) and factor classification (Mdn [IQR] = .35 [.18]) phase at T2. Further, there was only one (2.7%) instance by which HR for the factor classification phase was higher relative to the paired HR for the contributory factor assignment phase at T2 (positive ranks).

FAR was low for both phases of STAMP-CAST at T1 and T2. Differences between phases were moderate (Figure 5; panel C and D). A significant difference (Z = 4.507; p = <.001) was found in FAR between the contributory factor assignment (Mdn [IQR] = .02 [.02]) and factor classification (Mdn [IQR] = .14 [.12]) phase at T1. A significant difference (Z = 4.863; p = <.001) was also found in FAR between the contributory factor assignment (Mdn [IQR] = .02 [.02]) and factor classification (Mdn [IQR] = .16 [.11]) phase at T2.

[insert Figure 5 about here]

4.2.3 AcciNet

Figure 6 shows HR and FAR at T1 and T2 for the individual phases of AcciNet. HR was moderate for the contributory factor assignment phase at T1 and T2, and moderate and moderate to high for the

factor classification phase at T1 and T2, respectively. Differences between phases were small (Figure 6; panel A and B). A paired sample *t*-test found a significant difference ($t(33) = 2.377$; $p = .023$) in HR between the contributory factor assignment (*M* [SD] = .52 [.14]) and factor classification (*M* [SD] = .57 [.09]) phase at T1. A significant difference ($t(33) = 2.275$; $p = .030$) was also found between the contributory factor assignment (*M* [SD] = .56 [.15]) and factor classification (*M* [SD] = .61 [.08]) phase at T2.

FAR was low for both phases of AcciNet at T1 and T2. Differences between phases were negligible (Figure 6; panel C and D). A Wilcoxon signed-rank test found a significant difference ($Z = 3.300$; $p = .001$) in FAR between the contributory factor assignment (Mdn [IQR] = .04 [.03]) and factor classification (Mdn [IQR] = .07 [.04]) phase of AcciNet at T1. A significant difference ($Z = 3.438$; $p = .001$) was also found in FAR between the contributory factor assignment (Mdn [IQR] = .04 [.02]) and factor classification (Mdn [IQR] = .06 [.03]) phase at T2.

[insert Figure 6 about here]

4.3 Combined phase between method comparison

Figure 7 shows HR and FAR between the three methods when the contributory factor assignment and relationship description/factor classification phases were combined at T1 and T2. HR was moderate for AcciMap, STAMP-CAST and AcciNet at T1 and T2. The difference between AcciMap and STAMP-CAST was small at T1 and negligible at T2. The difference between AcciMap and AcciNet was small at T1 and T2. The difference between STAMP-CAST and AcciNet was negligible at T1 and T2 (Figure 7; panel A and B). A one-way repeated measures ANOVA found a significant difference ($F(2, 66) = 11.564$; $p = <.001$; $\eta_p^2 = .259$) in HR between AcciMap (*M* [*SD*] = .47 [.11]), STAMP-CAST (*M* [*SD*] = .55 [.10]) and AcciNet (*M* [*SD*] = .55 [.09]) at T1. Post-hoc testing with the Bonferroni correction revealed a significant difference between AcciMap and STAMP-CAST ($p = .002$), and AcciMap and AcciNet ($p = <.001$). No significant difference was found between STAMP-

CAST and AcciNet (p = 1.00). A one-way repeated measures ANOVA with a Greenhouse-Geisser correction also found a significant difference (F(1.557, 51.391) = 10.253; p = .001; η_p^2 = .237) in HR between AcciMap (M [SD] = .51 [.10]), STAMP-CAST (M [SD] = .54 [.09]) and AcciNet (M [SD] = .59 [.10]) at T2. Post-hoc testing revealed a significant difference between AcciMap and AcciNet (p = <.001). No significant difference was found between AcciMap and STAMP-CAST (p = .317), and STAMP-CAST and AcciNet (p = .051).

FAR was low for all methods at T1 and T2. The difference between methods was negligible (Figure 7; panel C and D). A comparison of the repeated measures using Friedman's test found no significant difference (χ^2(2) = .602; p = .740) in FAR between AcciMap (Mdn [IQR] = .05 [.04]), STAMP-CAST (Mdn [IQR] = .06 [.04]) and AcciNet (Mdn [IQR] = .05 [.02]) at T1. Conversely, a significant difference (χ^2(2) = 21.339; p = <.001) was found in FAR between AcciMap (Mdn [IQR] = .04 [.04]), STAMP-CAST (Mdn [IQR] = .06 [.03]) and AcciNet (Mdn [IQR] = .04 [.02]) at T2. Post-hoc testing with the Bonferroni correction revealed a significant difference between AcciMap and STAMP-CAST (p = <.001), and STAMP-CAST and AcciNet (p = <.001). No significant difference was found between AcciMap and AcciNet (p = 1.00).

[insert Figure 7 about here]

4.4 MCC between method comparison

Figure 8 shows MCC between the three methods when HR and FAR were combined at T1 and T2. A weak to moderate positive correlation coefficient was found for AcciMap T1, whereas a moderate positive correlation coefficient was found for STAMP-CAST and AcciNet. A weak to moderate positive correlation coefficient was found for AcciMap and STAMP-CAST and T2, whereas a moderate positive correlation coefficient was again found for AcciNet. A comparison of the repeated measures using Friedman's test found a significant difference (χ^2(2) = 8.226; p = .016) in MCC between AcciMap (Mdn [IQR] = .45 [.18]), STAMP-CAST (Mdn [IQR] = .53 [.18]) and AcciNet

(Mdn [IQR] = .50 [.12]) at T1. Post-hoc testing with the Bonferroni correction revealed that MCC was significantly different between AcciMap and AcciNet (p = .039), and AcciMap and STAMP-CAST (p = .046). No significant difference was found between STAMP-CAST and AcciNet (p = .100). A one-way repeated measures ANOVA with a Greenhouse-Geisser correction found a significant difference (F(1.447, 47.741) = 6.210; p = .008; η_p^2 = .158) in MCC between AcciMap (M [SD] = .47 [.12]), STAMP-CAST (M [SD] = .48 [.12]) and AcciNet (M [SD] = .54 [.11]) at T2. Post-hoc testing revealed that AcciNet MCC was significantly different from AcciMap (p = <.001) and STAMP-CAST (p = .024). No significant difference was found between AcciMap and STAMP-CAST (p = 1.00).

[insert Figure 8 about here]

5.0 Discussion

The purpose of this study was to test the criterion-referenced concurrent validity of AcciMap, STAMP-CAST and AcciNet. The rationale for the study was based on a recognised need for reliability and validity testing in HFE and safety science generally (Stanton and Young 1999; Kanis 2000; Annett 2002; Stanton and Young 2003; Kanis 2014; Stanton 2014; Waterson et al. 2015; Shorrock and Williams 2016; Stanton 2016; Salmon et al. 2020c; Hulme et al. 2021b; Hulme et al. 2021c) as well as accident investigation, reporting and analysis specifically (Olsen 2013; Goode et al. 2017; Salmon et al. 2017; Goncalves Filho et al. 2019; Thoroman et al. 2020; Goncalves Filho et al. 2021). The following discussion follows the same format as the results, covering the individual phase within method comparison, the combined phase between method comparison, and the overall criterion-referenced concurrent validity of AcciMap, STAMP-CAST and AcciNet. Implications and directions for future research are discussed.

5.1 Individual phase within method comparison

The results of the individual method phase comparison for AcciMap indicated a significant difference in HR between the contributory factor assignment phase and relationship description phase at T1 and T2. The difference in HR was large in effect, demonstrated by a median difference of .40 at T1 and a mean difference of .24 at T2 (Figure 4; panel A and B). The IQR for the relationship description phase was equally twice as large relative to the IQR for the contributory factor assignment phase at both timepoints. These findings reveal that not only was the description of relationships between factors a more inaccurate phase of analysis compared to the assignment of factors across RMF levels, but there was also a higher level of variability between analysts when identifying the correct relationships. Specifically, one third (32%) of the participants demonstrated a moderate level of agreement with the expert analysis, however a majority (68%) produced a low to moderate or low HR. This finding is concerning as it means that the modelling of relationships between factors with one of the most recognised systems-based accident analysis methods may produce results that do not reflect the objective reality of causation. Indeed, the accurate modelling of interactions between contributory factors is arguably the most important component of systems-based approaches that informs an understanding of accident aetiology (Salmon et al. 2020a). FAR was considerably low at T1 and T2 for both phases of AcciMap (Figure 4; panel C and D), primarily due to the high number of correct rejections for each successful factor assignment or relationship description. The significant difference in FAR at T2 may be a result of participants being more conservative when depicting relationships, thus recording fewer false alarms, albeit the median difference of .04 was negligible.

A similar pattern to AcciMap is observed for STAMP-CAST as the contributory factor assignment phase was associated with a significantly higher HR relative to the factor classification phase at T1 and T2. The difference was also large in magnitude; at T1 and T2 the median difference was .26 and .40, respectively (Figure 5; panel A and B). This finding shows that the participants achieved greater levels of accuracy in terms of assigning contributory factors to the appropriate control and feedback mechanisms described in the Uber-Volvo control structure model than they were at classifying failures using the STAMP-CAST classification scheme. In other words, participants were able to identify where controls and feedback mechanisms had failed across the system, but they were unable

to accurately classify these failures using the STAMP-CAST classification scheme. This finding adds further weight to other literature which has raised questions regarding the usability of the STAMP-CAST classification scheme (Salmon et al. 2012; Stanton et al. 2019). The original classification scheme may not be as applicable in this case compared to its use in highly engineered work systems that contain tightly coupled automated process control mechanisms and feedback loops. According to the HR and FAR thresholds of agreement between the participant and expert analyses, FAR for STAMP-CAST was found to be acceptable for both phases and timepoints (Figure 5; panel C and D). The reason for this is attributable to the high number of correct rejections that were identified for every true positive using the SDT paradigm. Nevertheless, the significant albeit moderate difference in the median FAR at T1 and T2 between phases suggests that the factor classification phase resulted in a higher number of false alarms relative to the contributory factor assignment phase.

The results of the individual method phase comparison for AcciNet show a trend in the opposite direction to AcciMap and STAMP-CAST. Specifically, a significantly lower HR at T1 and T2 was found for the contributory factor assignment phase relative to the factor classification phase (Figure 6; panel A and B). This means that participants found it difficult to accurately assign contributory factors to the appropriate tasks in the task network. The mean difference of .05 at T1 and T2 was, however, negligible with little practical significance. The IQR was considerably smaller for the factor classification phase when contrasted with the IQR for the contributory factor assignment phase, particularly at T2, pointing to more consistent results between analysts. These findings show that, compared to the expert analysis, participants were less accurate when assigning contributory factors to various tasks within the AcciNet task network than they were at classifying factors with the AcciNet classification scheme. Methodologically, the significantly higher HR for the second phase is likely due to the fact that AcciNet has the capacity to incorporate the interaction between normal performance and failure in the analysis. Thus, if a given task within the Uber-Volvo task network was found not to have any contributory factors associated with it, then participants classified that task as normal using the classification scheme. As a consequence, the sensitivity of the method was enhanced. A significant difference in FAR was found between the contributory factor assignment and

factor classification phase at T1 and T2 (Figure 6; panel C and D). The median difference was .03 and .02, respectively, which is not important in a practical sense and indicates the presence of a high number of correct rejections.

When considering the above results as a whole, the take-home message is that none of the methods achieved an acceptable level of agreement with the expert analysis in terms of HR for the contributory factor assignment or relationship description/factor classification phase at T1 or T2. For systems-based accident analysis methods that sit at the cutting edge of HFE and safety science, these results bring into the question the practical utility of such approaches when used to accurately describe and classify contributory factors and model relationships.

5.2 Combined phase between method comparison

This article has thus far provided a breakdown of how the individual phases of each method contribute to HR and FAR at T1 and T2. Figure 2 visualises the intricate process of analytical fragmentation. Accordingly, Figure 7 shows the HR and FAR results of the between method comparison analyses when the contributory factor assignment and relationship description/factor classification phases were combined at both timepoints. There was a moderate level agreement in HR between the participant and expert analyses for AcciMap, STAMP-CAST and AcciNet at T1 and T2; however, all methods fell considerably short of the 80% acceptable threshold (Figure 7; panel A and B). As FAR was acceptably low at T1 and T2 (Figure 7; panel C and D), it is evident that these moderate and substandard combined phase analysis results are primarily explained by the high volume of misses reported during the application of the individual method phases. Indeed, when the results in Figure 7 are placed in context with the individual method phase findings in Figures 4-6, it is possible to understand how the combined phase HR and FAR results were obtained. For example, whilst AcciNet produced the lowest HR for the contributory factor assignment phase relative to AcciMap and STAMP-CAST, it was also associated with the highest HR for the contributory factor classification phase at T1 and T2. The reverse is true for AcciMap and STAMP-CAST, thus restoring the balance in the combined phase analysis results. Overall, whilst the differences were statistically

significant between AcciMap and STAMP-CAST at T1, and AcciMap and AcciNet at T1 and T2, the size of the effect was negligible to small.

5.3 MCC between method comparison

The MCC results indicate the overall criterion-referenced concurrent validity of AcciMap, STAMP-CAST and AcciNet. AcciNet was the only method that achieved a moderate positive correlation coefficient between the participant and expert analyses on both occasions, a result that was significantly different to AcciMap at T1, and AcciMap and STAMP-CAST at T2. Despite the significant differences, the three methods achieved comparable, moderate and generally substandard results. The number of false alarms and misses identified prevented the methods from achieving higher and therefore stronger positive correlations with the expert analyses.

5.4 Summary of the main findings

Table 3 includes a summary and interpretation of the main study findings.

[insert Table 3 about here]

5.5 Recommendations

Aside from the overall validity scores obtained for each method, the findings of this study show that there was a large degree of variability in individual participant scores. This result was also found in recent study that examined the reliability and validity of systems-based risk assessment methods (Hulme et al. 2021c). For example, the range in the correlation coefficient scores for AcciMap, STAMP-CAST and AcciNet was .63, .48 and .42 at T1, respectively. Accordingly, some of the participants achieved relatively high levels of accuracy whereas others did not. Further, as seen in Figure 3, all 34 (100%) participants indicated that they lacked experience with AcciNet, yet after receiving training and applying the method in the workshop, 17 (50%) participants achieved a moderate positive correlation coefficient (MCC $\geq$.50, $<$.80), with the remaining half producing a

weak positive correlation coefficient (MCC ≥.20, <.50). The same trend is observed for AcciMap and STAMP-CAST.

What the above finding means by way of recommendation is that further training in the use of systems-based accident analysis methods is required prior to their use in practice. The emphasis of training should be placed on those phases of the analyses that performed relatively poorly, such as the description of relationships between factors in AcciMap, the classification of factors with the STAMP-CAST classification of control flaws, and the assignment of factors to tasks when using AcciNet, including training in the development and use of task networks (Stanton et al. 2018). Whilst training was provided as part of the study design, it is likely that the specific procedures associated with each method necessitate much more time to learn and master from what can be achieved in a one-day workshop. Shorrock and Williams (2016) also explain that methods requiring extensive time on activities that offer diminishing returns, especially 'invisible' desk-based non-user activities such as formatting, drawing diagrams and developing networks, are harder to sell and justify in practice. Given that AcciMap, STAMP-CAST and AcciNet almost perfectly fit this description, and in light of the results in Figure 3, it is clear that a majority of the participants were not familiar with the underpinning methods and tools used to support the application of each method. Indeed, systems-based accident analysis methods involve multiple parts and phases, including risk and safety management frameworks (e.g., Rasmussen 1997), hierarchical system models and control structures (e.g., Leveson 2004), task networks (e.g., Salmon et al. 2021b; Stanton et al. 2018), and classification schemes and taxonomies (e.g., Goode et al. 2017). In some cases, these elements are even refined further from existing HFE and safety techniques. From the above it can be concluded that there is a need for future research to first identify the underlying reasons for the observed differences in performance between analysts, which in turn, may offer new insights to inform appropriate method training deliverables. In doing so it will be possible to enhance the validity of systems-based HFE and safety science methods more widely.

Finally, it is worth noting that this study did not focus on method usability (i.e., how easy a method is to use) or utility (i.e., how useful are the results/outputs). Whilst there is an overlap between

reliability, validity, usability and utility, each of which are important concepts to consider and examine in and of themselves (Waterson et a. 2017), the present study was only concerned with determining the accuracy of the incident analysis results. Thus, the current study is a necessary piece of a much larger methods testing and comparison puzzle, and essentially addressed the question – do systems-based accident analysis methods produce outputs that reflect the objective reality of causation as deemed by an expert analysis. Given that usability and utility were not directly assessed, it is difficult to provide a hard recommendation as to which method to adopt in practice, especially given that they all achieved comparable validity results when applied to a specific case study example (i.e., a road collision incident in the 'road transport system'). However, when reflecting on the individual phase within method comparison analyses, the description of relationships between contributory factors in AcciMap at both T1 and T2 was associated with the lowest HR (sensitivity) relative to the referent expert analysis and when compared to the individual phases of STAMP-CAST and AcciNet. Specifically, a median HR of .30 at T1 and a mean HR of .40 at T2 for the relationship description phase significantly affected the HR of AcciMap in the combined phase between method comparison analyses. It is recommended that should AcciMap be used to analyse adverse incidents, a process of verification is applied to the relationships described across all levels of the RMF/system. This can be achieved in a number of ways, including the use of a subject matter expert workshops to establish agreement and consensus, or corroboration from independent analyses of the same incident to support a cross-validation. Analysts should also pay close attention to the details reported in the official investigation report(s) and seek out further evidence if required to inform their causal reasoning when linking contributory factors.

From this point forward, there is a need for further reliability and validity testing of systems-based accident analysis methods in various safety-critical domains to evaluate validity generalisation, as well as research applications that assess method usability and utility in context of the obtained stability and accuracy measures. Future work of this kind would move the HFE and safety science discipline closer in being able to support researchers and practitioners in deciding which systems-theoretic technique to adopt in practice.

5.6 Study limitations and research considerations

A number of study limitations and considerations to assist with the interpretation of the findings are disclosed. First, whilst participants were required to describe the relationships between factors in AcciMap, the control and feedback mechanisms in STAMP and the links between tasks in AcciNet were already specified in the expert analyses. The validity of STAMP-CAST and AcciNet would have been negatively affected if participants were asked to develop their own STAMP model and AcciNet task network of the Uber-Volvo system. Second, the self-reported level of expertise of the participant group across various method dimensions was low, with few exceptions. Although a total of six hours' worth of training was provided, the validity of the methods may have been influenced by the level of expertise. An interesting line of future inquiry would be to examine the validity of systems-based accident analysis methods amongst a group of highly trained experts in the methods. Third, the volume of correct rejections in this study was high compared to hits, misses and false alarms. Future studies could reduce the correct rejection frequency though design, specifically by creating a ruleset as to which factors could conceivably link to the remaining list of factors (in the case of AcciMap); or, which classification categories and modes in the classification scheme could reasonably apply to the particular contributory factor or failure identified (in the case of STAMP-CAST and AcciNet). In taking this approach, the absolute number of correct rejections would be more reflective of the true result; however, serious due consideration among the research team as to which relationships and categories to ignore is required. Fourth, the thresholds of agreement used in this study, and specifically the high threshold bracket (>.8), still admits 20% misclassifications, which could lead to a disaster. Thus, if one of the methods produces a moderate level of agreement with the referent expert analysis but succeeds in capturing the most vital factors, then it may be worth selecting that method over another which hits 'high' but overlooks the most critical factors. Future work should incorporate a qualitative dimension into the thresholds to evaluate method utility alongside measures of stability and accuracy. Fifth, the reliability and validity findings in this study may have been different had the methods been applied in another system and to a different incident. There is a

need for further work to examine the possibility of identifying and quantifying such differences when systems-based methods are applied elsewhere and to different domains.

6.0 Conclusion

This study has tested the criterion-referenced concurrent validity of AcciMap, STAMP-CAST and AcciNet. The findings of the overall between method analyses indicate that AcciMap, STAMP-CAST and AcciNet were comparable in terms of their criterion-referenced concurrent validity. However, the capacity of systems-based methods to accurately model accident aetiology was questionable based on the weak to moderate positive correlations with a referent expert analyses. There is a pressing need for further reliability and validity testing in HFE and safety science. Future studies should apply and compare safety methods in various domains to evaluate validity generalisation.

Funding

This work was supported by an Australian Research Council (ARC) Discovery Project grant (grant number: DP180100806).

References

Ahmadi, O., Mortazavi, S.B., Khavanin, A., Mokarami, H. (2019). Validity and consistency assessment of accident analysis methods in the petroleum industry. International Journal of Occupational Safety and Ergonomics 25(3): 355-361.

Annett, J. (2002). A note on the validity and reliability of ergonomics methods. Theoretical Issues in Ergonomics Science 3(2): 228-232.

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

Dallat, C., Salmon, P.M., Goode, N. (2018). Identifying risks and emergent risks across sociotechnical systems: The NETworked hazard analysis and risk management system (NET-HARMS). Theoretical Issues in Ergonomics Science 19(4): 456-482.

Dekker, S. (2011). Drift Into failure: From hunting broken parts to understanding complex systems. Surrey, UK, Ashgate.

Goncalves Filho, A.P., Jun, G.T., Waterson, P. (2019). Four studies, two methods, one accident – An examination of the reliability and validity of Accimap and STAMP for accident analysis. Safety Science 113: 310-317.

Goncalves Filho, A.P., Waterson, P., Jun, G.T. (2021). Improving accident analysis in construction – Development of a contributing factor classification framework and evaluation of its validity and reliability. Safety Science 140: 105303.

Goode, N., Salmon, P.M., Taylor, N.Z., Lenné, M.G., Finch, C.F. (2017). Developing a contributing factor classification scheme for Rasmussen's AcciMap: Reliability and validity evaluation. Applied Ergonomics 64: 14-26.

Green, D.M., Swets, J.A. (1966). Signal detection theory and psychophysics. Ney York, USA, Wiley New York.

Greenwood, M., Woods, H.M. (1919). A report on the incidence of industrial accidents upon individuals with special reference to multiple accidents, Reports of the Industrial Fatigue Research Board 4, Darling and son, Limited, printers.

Heinrich, H.W. (1931). Industrial Accident Prevention. A Scientific Approach. New York, USA, McGraw-Hill.

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

Hollnagel, E. (2012). FRAM, the Functional Resonance Analysis Method: Modelling complex socio-technical systems. Farnham, England, Ashgate Publishing, Ltd.

Hulme, A., McLean, S., Dallat, C., Walker, G.H., Waterson, P., Stanton, N.A., Salmon, P.M. (2021a). Systems thinking-based risk assessment methods applied to sports performance: A comparison of STPA, EAST-BL, and Net-HARMS in the context of elite women's road cycling. Applied Ergonomics 91: 103297.

Hulme, A., Stanton, N.A., Walker, G.H., Waterson, P., Salmon, P.M. (2020). Complexity theory in accident causation: Using AcciMap to identify the systems thinking tenets in 11 catastrophes. Ergonomics: 1-18.

Hulme, A., Stanton, N.A., Walker, G.H., Waterson, P., Salmon, P.M. (2021b). Testing the reliability and validity of Net-HARMS: A new systems-based risk assessment method in HFE. In: Black N.L., Neumann W.P., Noy I. (eds) Proceedings of the 21st Congress of the International Ergonomics Association (IEA 2021). IEA 2021. Lecture Notes in Networks and Systems, vol 219. Springer, Cham (doi: 10.1007/978-3-030-74602-5_51).

Hulme, A., Stanton, N.A., Walker, G.H., Waterson, P., Salmon, P.M. (2021c). Testing the reliability and validity of risk assessment methods in human factors and ergonomics. Ergonomics (in press, doi: 10.1080/00140139.2021.1962969).

Hulme, A., Stanton, N.A., Walker, G.H., Waterson, P., Salmon, P.M. (2019). What do applications of systems thinking accident analysis methods tell us about accident causation? A systematic review of applications between 1990 and 2018. Safety Science 117: 164-183.

Johnson, C.W., de Almeida, I.M. (2008). An investigation into the loss of the Brazilian space programme's launch vehicle VLS-1 V03. Safety Science 46(1): 38-53.

Kanis, H. (2000). Questioning validity in the area of ergonomics/human factors. Ergonomics 43(12): 1947-1965.

Kanis, H. (2014). Reliability and validity of findings in ergonomics research. Theoretical Issues in Ergonomics Science 15(1): 1-46.

Katsakiori, P., Sakellaropoulos, G., Manatakis, E. (2009). Towards an evaluation of accident investigation methods in terms of their alignment with accident causation models. Safety Science 47(7): 1007-1015.

Khanzode, V.V., Maiti, J., Ray, P.K. (2012). Occupational injury and accident research: A comprehensive review. Safety Science 50(5): 1355-1367.

Leveson, N. (2011). Applying systems thinking to analyze and learn from events. Safety Science 49(1): 55-64.

Leveson, N. (2004). A new accident model for engineering safer systems. Safety Science 42(4): 237-270.

Leveson, N., Dulac, N., Marais, K., Carroll, J. (2009). Moving beyond normal accidents and high reliability organizations: A systems approach to safety in complex systems. Organization studies 30(2-3): 227-249.

Leveson, N.G. (2016). Engineering a safer world: Systems thinking applied to safety. Cambridge, USA, The MIT Press.

Lindberg, A.K., Hansson, S.O., Rollenhagen, C. (2010). Learning from accidents – What more do we need to know? Safety Science 48(6): 714-721.

MAIB. (2015). Flooding and capsize of ro-ro passenger ferry Herald of Free Enterprise with loss of 193 lives. Available from https://www.gov.uk/maib-reports/flooding-and-subsequent-capsize-of-ro-ro-passenger-ferry-herald-of-free-enterprise-off-the-port-of-zeebrugge-belgium-with-loss-of-193-lives.

Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta -Protein Structure 405(2): 442-451.

NTSB (2018). National Transportation Safety Board Preliminary Report Highway: HWY18MH010. Available at: https://www.ntsb.gov/investigations/AccidentReports/Pages/HWY18MH010-prelim.aspx.

Olsen, N.S. (2013). Reliability studies of incident coding systems in high hazard industries: A narrative review of study methodology. Applied Ergonomics 44(2): 175-184.

Rasmussen, J. (1997). Risk management in a dynamic society: a modelling problem. Safety Science 27(2-3): 183-213.

Rasmussen, J., Suedung, I. (2000). Proactive risk management in a dynamic society, Swedish Rescue Services Agency.

Reason, J. (1990). Human error. New York, USA, Cambridge University Press.

Royal Automobile Club. (2019). Models and methods for collision analysis. Available at: https://www.racfoundation.org/research/safety/models-and-methods-for-collision-analysis

Salmon, P.M. (2016). Bridging the gap between ergonomics methods research and practice: Methodological Issues in Ergonomics Science Part II. Theoretical Issues in Ergonomics Science 17(5-6): 459-467.

Salmon, P.M., Cornelissen, M., Trotter, M.J. (2012). Systems-based accident analysis methods: A comparison of Accimap, HFACS, and STAMP. Safety Science 50(4): 1158-1170.

Salmon, P.M., Hulme, A., Walker, G.H., Waterson, P., Berber, E., Stanton, N.A. (2020a). The big picture on accident causation: A review, synthesis and meta-analysis of AcciMap studies. Safety Science 126: 104650.

Salmon, P.M., Hulme, A., Walker, G.H., Waterson, P., Stanton, N.A. (2020b). The Accident Network (AcciNet): A new accident analysis method for describing the interaction between normal performance and failure. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, SAGE Publications Sage CA: Los Angeles, CA.

Salmon, P.M., Read, G.J.M., Walker, G.H., Stevens, N.J., Hulme, A., McLean, S., Stanton, N.A. (2020c). Methodological issues in systems Human Factors and Ergonomics: Perspectives on the research–practice gap, reliability and validity, and prediction. Human Factors and Ergonomics in the Manufacturing and Service Industries: 1-14.

Salmon, P.M., Stanton, N.A., Lenne, M., Jenkins, D.P., Rafferty, L., Walker, G.H. (2011). Human factors methods and accident analysis: Practical guidance and case study applications. Surrey, England, Ashgate Publishing, Ltd.

Salmon, P.M., Walker, G.H., Read, G.J.M., Goode, N., Stanton, N.A. (2017). Fitting methods to paradigms: Are ergonomics methods fit for systems thinking? Ergonomics 60(2): 194-205.

Shappell, S.A., Wiegmann, D.A. (2001). Applying Reason: The Human Factors Analysis and Classification System (HFACS). Human Factors and Aerospace Safety.

Shorrock, S.T., Williams, C.A. (2016). Human factors and ergonomics methods in practice: Three fundamental constraints. Theoretical Issues in Ergonomics Science 17(5-6): 468-482.

Stanton, N.A. (2014). Commentary on the paper by Heimrich Kanis entitled 'Reliability and validity of findings in ergonomics research': where is the methodology in ergonomics methods? Theoretical Issues in Ergonomics Science 15(1): 55-61.

Stanton, N.A. (2016). On the reliability and validity of, and training in, ergonomics methods: A challenge revisited. Theoretical Issues in Ergonomics Science 17(4): 345-353.

Stanton, N.A., Salmon, P.M., Walker, G.H. (2018). Systems thinking in practice: Applications of the event analysis of systemic teamwork method. Boca Raton, Florida, USA, CRC Press.

Stanton, N.A., Salmon, P.M., Walker, G.H., Stanton, M. (2019). Models and methods for collision analysis: A comparison study based on the Uber collision with a pedestrian. Safety Science 120: 117-128.

Stanton, N.A., Young, M.S. (2003). Giving ergonomics away? The application of ergonomics methods by novices. Applied Ergonomics 34(5): 479-490.

Stanton, N.A., Young, M.S. (1999). What price ergonomics? Nature 399(6733): 197-198.
Svedung, I., Rasmussen, J. (2002). Graphic representation of accident scenarios: Mapping system structure and the causation of accidents. Safety Science 40(5): 397-417.

Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. Journal of Diagnostic Medical Sonography 6(1): 35-39.

Thoroman, B., Salmon, P.M., Goode, N. (2020). Evaluation of construct and criterion-referenced validity of a systems-thinking based near miss reporting form. Ergonomics 63(2): 210-224.

Underwood, P., Waterson, P. (2013). Systemic accident analysis: Examining the gap between research and practice. Accident Analysis & Prevention 55: 154-164.

Underwood, P., Waterson, P. (2014). Systems thinking, the Swiss Cheese Model and accident analysis: A comparative systemic analysis of the Grayrigg train derailment using the ATSB, AcciMap and STAMP models. Accident Analysis & Prevention 68: 75-94.

Walker, G.H., Stanton, N.A., Salmon, P.M., Jenkins, D.P., Rafferty, L. (2010). Translating concepts of complexity to the field of ergonomics. Ergonomics 53(10): 1175-1186.

Waterson, P., Jenkins, D.P., Salmon, P.M., Underwood, P. (2017). 'Remixing Rasmussen': The evolution of Accimaps within systemic accident analysis. Applied Ergonomics 59: 483-503.

Waterson, P., Robertson, M.M., Cooke, N.J., Militello, L., Roth, E., Stanton, N.A. (2015). Defining the methodological challenges and opportunities for an effective science of sociotechnical systems and safety. Ergonomics 58(4): 565-599.