Applied Ergonomics 44 (2013) 175–184

Contents lists available at SciVerse ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo

Reliability studies of incident coding systems in high hazard industries: A narrative review of study methodology

Nikki S. Olsen*

Department of Aviation, The University of New South Wales, Sydney, NSW 2052, Australia

ARTICLE INFO

Article history:
Received 16 January 2012
Accepted 28 June 2012

Keywords:
Reliability
Incident classification
Human error

ABSTRACT

This paper reviews the current literature on incident coding system reliability and discusses the methods applied in the conduct and measurement of reliability. The search strategy targeted three electronic databases using a list of search terms and the results were examined for relevance, including any additional relevant articles from the bibliographies. Twenty five papers met the relevance criteria and their methods are discussed. Disagreements in the selection of methods between reliability researchers are highlighted as are the effects of method selection on the outcome of the trials. The review provides evidence that the meaningfulness of and confidence in results is directly affected by the methodologies employed by the researcher during the preparation, conduct and analysis of the reliability study. Furthermore, the review highlights the heterogeneity of methodologies employed by researchers measuring reliability of incident coding techniques, reducing the ability to critically compare and appraise techniques being considered for the adoption of report coding and trend analysis by client organisations. It is recommended that future research focuses on the standardisation of reliability research and measurement within the incident coding domain.

1. Introduction

1.1. Incident reporting and analysis

Since the general acceptance of the prevalence of human error in high hazard industries such as aviation, rail and medicine, systems to analyse human error have existed with the aim of discovering how to eliminate, reduce or mitigate the errors. Accidents, incidents and events are reported retrospectively by personnel, supervisors and safety managers and stored in standard templates in electronic systems. Analysis is performed on the incident data to determine trends so that procedures, training and management can be altered in an attempt to eliminate the likelihood of or reduce the consequences of the error recurring. Alternatively, predictive assessments may be conducted by personnel during any part of the system lifecycle to highlight possible flaws that may result in future sources of human error.

The amount of data gathered by these systems, over time and over many departments of an organisation, requires that key pieces of information are stored in a simplified manner. By identifying the contributing factors in the incident reports and assigning standardised codes to these factors error trends can be analysed by automated means as well as clearly and quickly when viewed manually. Therefore at the heart of error reporting systems are predetermined codes in the form of taxonomies, lists, classification systems and models that provide a framework for investigators to assign codes to contributing factors of incident reports. These are compared with codes assigned to other reports for the analysis of human error trends.

Because the resultant trends are integral to the implementation of error elimination and mitigation measures in high hazard industries, researchers must ensure that the coding techniques are valid so that accurate and applicable trends are recorded. Validity refers to the testing of a tool, method or technique to ascertain whether it actually does what it says it does (RSSB, 2005). Several papers (Fleishman and Mumford, 1991; Kirwan, 1998; RSSB, 2005) identify a number of different types of validity that must be addressed in validating techniques, taxonomies, models and tools of the type reviewed in this paper. The specific type of validity that is focused on here is reliability.

Reliability refers to the extent to which a test, experiment or measuring procedure gives the same result(s) on repeated trials or applications (RSSB, 2005). In particular, intercoder reliability refers to the degree of agreement between a number of different analysts

* Tel.: +61 0404869261; fax: +61 351571593.
E-mail address: Nikki.Olsen@defence.gov.au.

0003-6870/$ – see front matter © 2012 Elsevier Ltd and The Ergonomics Society. All rights reserved.
http://dx.doi.org/10.1016/j.apergo.2012.06.009

classifying an error using a coding technique whereas intracoder reliability describes how one analyst will classify errors over time (RSSB, 2005). However, the terms reliability and agreement are still generally broad terms and require further definition to ensure their correct application to measurements within the incident coding domain.

Kozlowski and Hattrup (1992) distinguish between reliability and agreement in the following way:

> '*reliability* … references proportional consistency of variance among raters and is correlational in nature [whereas] in contrast, *agreement* references the interchangeability among raters [and] addresses the extent to which raters make essentially the same ratings'.

These definitions highlight that coders can be reliable in their coding assignments where the range of codes assigned by one coder is consistent with the range of codes assigned by another coder, even if the codes assigned to each individual event do not meet with consensus. On the other hand, where codes assigned to each individual event are the same between coders (that is, consensus is reached on the codes assigned to each individual event) then there is an agreement. Ross et al. (2004) and Tinsley and Weiss (1975) demonstrate that high intercoder reliability can be obtained even when there is little agreement between coders.

Given the nature of incident coding studies, reliability indexes are likely to be a biased estimation of the actual agreement between coders assigning codes to incident reports and this is potentially hazardous when such codes form the basis of safety mitigation measures. Therefore, as part of validating a technique, the researcher is ensuring that a high level of consensus can be reached between users of that technique, within certain contexts of use. It is for this reason why studies of this nature can be found within the literature.

1.2. Objectives of this review

While reliability studies are generally accepted as an appropriate tool for measuring consensus of coding and therefore an important part of establishing the validity of a technique, very few studies have actually been conducted on coding techniques, and even fewer have been published. Attempts to compare the reliability of numerous techniques have often been reported with very little detail of the trial methodology and results due to the emphasis being on the technique descriptions and other validation aspects (Beaubien and Baker, 2002; Kirwan, 1998; Loeb and Chang, 2003). No previous studies exist providing a detailed review of reliability studies or study methodology as applied to incident coding techniques.

Therefore this paper reviews the current literature on incident coding system reliability and discusses the methods applied in the conduct and measurement of agreement. Disagreements in the selection of methods between reliability researchers are highlighted as are the effects of method selection on the outcome of the trials. Finally the implications and limitations of the review are discussed and future research directions are suggested.

1.3. Terms used

The distinction between the terms reliability and agreement are extremely important to the future of studies in the incident coding domain, however it is clear in the reviewed literature to follow that in this domain the term reliability study refers to the consensus or agreement amongst coders and not the consistency of codes assigned. As a result, the term 'reliability study' will continue to be used for uniformity with the reviewed literature. Additionally the following concepts need clarification:

- technique levels: techniques might not simply be a list of codes available for assignment but may contain several levels within a taxonomy, model or other classification system. A typical technique may present a number of codes that represent each of the most often caused errors within a particular industry. Related error codes may then be contained within more general categories which in their own right may be assigned as a code, albeit less detailed than at the error level. Again, related categories may be categorised together in one or more taxonomies or models which can also be assigned as a code, but again codes assigned from this level are more general than codes assigned from the category or error level. Although different technique designs use different terms for each of these levels, this review uses the terms 'descriptor level', 'category level' and 'taxonomy level' for uniformity to describe the levels from which a particular code may be sourced. It is not unusual for a study to measure agreement on codes selected from the more specific 'descriptor level' and compare these results to agreement measures from codes taken from the more general 'category level' and 'taxonomy level'.
- report characteristics: a typical format for incident reports includes an analysis discussing the significance of the facts drawn from the investigation and a lists of findings (statements of all significant conditions, events or circumstances in the occurrence sequence that contributed to the occurrence) (Australian Government, 2009). Codes are then assigned to the findings from a coding technique.

2. Methods

Three electronic databases were searched: MEDLINE, Ergonomics Abstracts and Health and Safety Science Abstracts. Each database's controlled vocabulary search engines were used and then the first 200 results from each search were considered where the engine returned in excess of 200 entries. Table 1 lists all search terms used. The search strategy ensured that at least one term from each box was present in the title, abstract or body of each paper by the use of the Boolean operator "AND", however terms within each box were interchangeable using the Boolean operator "OR". For example, searches revealed papers that contained "taxonomy" AND "reliability" AND ("intercoder agreement" OR "interrater agreement") AND ("accident" OR "event"). Google Scholar was used in a similar way to find relevant papers, however with the absence of a controlled vocabulary search engine terms were manually typed using Google's guidelines for searching.

Considering the generalness of the search terms large numbers of hits were returned. However it was quickly clear that the majority of these hits were not relevant to the review. Therefore it was additionally necessary to read the titles and abstracts of returned entries to ensure that the terms that were found during the search were used in the context relevant to the review. In addition to

Table 1
Search terms.

Taxonomy Technique Model	Reliability Consensus
Intercoder agreement Interrater agreement Intercoder consistency Interrater consistency Intracoder agreement Intracoder consistency Intrarater agreement Intrarater consistency	Accident Incident Event Adverse event Hazard

electronic searches references of papers selected for review were manually searched by title and where further clarification of relevance was required, by searching abstracts. The search was limited to English language papers published prior to April 2012.

3. Findings

3.1. Inclusion and exclusion of papers

After completing the search and reading the titles and abstracts of the returned entries 35 papers were retrieved in full for further confirmation of relevance to the review. Of these 35 papers, 27 met the inclusion criteria of being published in peer reviewed journals or as technical reports for a professional safety department of a high hazard industry (for example, the Federal Aviation Administration, US and the Rail Safety Standards Board, UK). Three papers were excluded where only an abstract had been published and the full works could not be sourced from the authors (Hughes et al., 2007; West et al., 1991; Wiegmann and Shappell, 2001), one was excluded because it did not present any numerical results for analysis (van Vuuren et al., 1997), one was excluded because it reported on the same study as another already included in this review (Fagerlind et al., 2008) and one was excluded having been published in a book chapter describing unpublished papers with no method details (Wiegmann and Shappell, 2003).

3.2. Technique characteristics

Of the 27 papers, 13 (48%) of the papers outlined the use of the coding technique in the aviation industry, 5 (19%) in the medical field, 3 (11%) in the vehicle accident sector, 5 (19%) in the rail industry and 1 (4%) in the nuclear industry. Twenty different coding techniques were studied in 41 trials over the 27 papers. Additionally 13 (65%) appeared to be based on a theoretical model or principles of human error or behaviour (for example, Reason's (1990) Swiss Cheese Model). The remaining 7 (35%) had been constructed by subject matter experts sorting codes into categories and correlating the results to derive the final technique structure.

Only 1 (5%) of the techniques was solely designed for predictive use whereas 14 (70%) were solely designed for retrospective use. The remaining 5 (25%) techniques could be used both predictively and retrospectively.

Finally it is also noteworthy that while some techniques are trialed in their original forms, many have also been developed beyond those original forms for more effective use in other industries and are as such derivatives of the original technique. For example the Technique for the Retrospective and Predictive Analysis of Cognitive Errors (TRACEr), designed for air traffic control has at least two derivatives in the rail industry (TRACEr-Rail (RSSB, 2005) and TRACEr-RAV (Baysari et al., 2011)). Of the 20 techniques tested in the relevant papers, 13 (65%) were trialed in their original form and the remaining 7 (35%) were trialed as derivatives or later versions. Table 2 details the characteristics of the techniques trialed in the reviewed papers.

3.3. Reliability study characteristics

Twelve (44%) of the 27 papers incorporated a reliability study as part of the development of a new technique. Seven (58%) of these were only done after completion of the technique development whereas 5 (42%) of the 'technique development' papers used reliability studies during technique development to provide feedback on the development process. Ten (37%) papers used a reliability study in order to validate an existing technique whereas 8 (30%) papers used reliability studies to make comparisons (some papers had more than one aim). These comparisons included comparing the reliability of different professions of participants using the same technique (2 papers, 25%), comparing the reliability of trained users versus untrained users (1 paper, 13%), comparing the reliability of users from different nationalities (1 paper, 13%), comparing the reliability of an original technique versus its derivative (2 papers, 25%) and comparing the reliability of two different techniques (2 papers, 25%). One (4%) paper used a reliability study as part of a wider study to test hypotheses. Intercoder consensus was tested in 38 (93%) trials and intracoder consensus in 3 (7%) trials.

A total of 752 participants were reported in 39 trials (2 trials did not report the number of participants). Ten (26%) papers used students, 8 (21%) used subject matter experts (line workers, operators, organisation management), 6 (15%) used technique developers, their assistants or participants simply reported as 'coders', 5 (13%) used human factors and risk specialists and 7 (18%) used incident investigators and safety managers. Of the total 752 participants reported 584 (78%) of these were students however 345 of these students were military aviation students with aviation theoretical and practical knowledge. The modal number of participants for the reviewed studies is 7.

Of the 41 trials conducted, 25 (61%) used full accident, incident or event reports, 9 (22%) used abridged reports and 7 (17%) used scenarios or interview transcripts. Twenty three (59%) of the trials required participants to identify both the factors contributing to the incident as well as code the factors whereas 12 (29%) provided pre-identified factors and 6 (12%) did not report this information. Only one (4%) trial required participants to identify factors from an abridged report. The modal number of reports in the reviewed studies is 14.

Training was provided in 15 (56%) studies with 11 (73%) of these providing training of less than one day, 2 (13%) providing training of one day or greater and 2 (13%) providing self-paced training via distance learning format (workbook and email). Additionally, 11 (41%) studies provided materials other than the technique itself to participants. Four (36%) of these provided detailed definitions of the codes, 4 (36%) provided task descriptions or task analysis proformas, 2 (19%) provided flow charts, 1 (10%) provided access to the personnel involved in the incident, 2 (19%) provided photographs of the cockpit layout, 1 (10%) provided procedure manuals, engineering and medical reports and 1 (10%) provided a demonstration of the task via Microsoft Flight Simulator (some studies provided more than one type of extra material). Twelve (44%) studies provided only the technique and an answer form and 4 (15%) did not report the materials provided to the participants.

A variety of analysis tools and methods were employed in the studies. Seven (26%) studies used Kappa, 6 (22%) used the percentage of participants agreeing on the modal category, 5 (19%) used the Index of Concordance, 6 (22%) used a form of percentage agreement but did not specify the type and 7 (26%) used other methods (r_{wg}, Pearson's r, Yule's Q, Sensitivity matrices, Signal Detection Paradigm). Four (15%) studies used significance tests and 8 (30%) studies used comparison statistics to determine appropriate values of reliability. In most studies all data gathered was analysed to determine the technique reliability, however in 10 (37%) studies researchers employed methods to select specific data for analysis. Three (11%) studies employed a 'no choice' or 'missing' box in place of a non selection by a participant in order to count the non selection as an agreement where another participant had also made a non selection or as a disagreement where another participant had made a code selection. Table 3 tabulates for each paper the characteristics employed by that paper in the trial aim and preparation stages and Table 4 tabulates the paper conduct and analysis characteristics and the results of individual trials.

Table 2
Characteristics of techniques trialled in reliability studies reviewed.

Surname(s), year	Technique(s) reviewed and/or developed	Technique is the original or a derivative	Field of study	Technique is predictive or retrospective	Technique origin
af Wåhlberg, 2002	*No name*	Original	Automotive (bus)	Retrospective	This paper
Baker and Krokos, 2007	ACCERS	Original	Aviation (pilot)	Retrospective	Krokos and Baker (2005)
Baysari et al., 2009	HFACS	Original	Aviation (pilot)	Retrospective	Wiegmann and Shappell (2003)
Baysari et al., 2009	TRACEr	Original	Aviation (ATC)	Both	Shorrock and Kirwan (2002)
Baysari et al., 2011	TRACEr-Rail	Derivative	Rail	Retrospective	RSSB (2005)
Baysari et al., 2011	TRACEr- RAV	Derivative	Rail	Retrospective	This paper
Gibson, 2006	*Referred to as*: rail-specific HRA technique	Original	Rail	Retrospective	RSSB
Harris et al., 2005	SHERPA	Original	Aviation (pilot)	Predictive	Embrey (1986)
Isaac et al., 2003a	HERA-JANUS	Derivative	Aviation (ATC)	Both	Isaac et al. (2003b)
Jacobs et al., 2007	*No name*	Original	Medicine (Family)	Retrospective	This paper
Kaplan et al., 1998	MERS – TM	Derivative	Medicine (Transfusion)	Retrospective	This paper
Krokos and Baker, 2005	ACCERS	Original	Aviation (pilot)	Retrospective	This paper
Makeham et al., 2008	TAPS	Original	Medicine (GP)	Retrospective	This paper
O'Connor, 2008, O'Connor et al., 2010, O'Connor and Walker, 2011	DoD- HFACS	Derivative	Aviation (pilot)	Retrospective	Department of Defense (accessed 2011)
Olsen and Shorrock, 2010	HFACS – ADF	Derivative	Aviation –all jobs	Retrospective	Australian Government (2009)
Olsen, 2011	HFACS	Original	Aviation (pilot)	Retrospective	Wiegmann and Shappell (2003)
Pounds and Isaac, 2003	JANUS	Original	Aviation (ATC)	Retrospective	Isaac et al. (2003b)
Read et al., 2012	Contributing factors framework	Original	Rail	Retrospective	Safety Regulators' Panel (2009)
RSSB, 2005	TRACEr	Original	Rail	Retrospective	Shorrock and Kirwan (2002)
RSSB, 2005	TRACEr-lite	Derivative	Rail	Retrospective	Shorrock and Kirwan (2002)
Shorrock, 2002	TRACEr	Original	Aviation (ATC)	Both	This paper
Stanton et al., 2002	SHERPA	Original	Aviation (pilot)	Predictive	Embrey (1986)
Stanton et al., 2002	SHERPA	Original	Aviation (pilot)	Predictive	Embrey (1986)
Terhune, 1983	CALAX	Original	Automotive	Retrospective	This paper
Wallace et al., 2002	Observed cause and root cause analysis system	Original	Nuclear	Retrospective	UK Nuclear Industry
Wallace et al., 2002	SECAS	Derivative	Nuclear	Retrospective	This paper
Wallen Warner and Sandin, 2010	DREAM 3.0	Derivative	Automotive	Retrospective	Ljung (2002)
Woods, 2005	*No name*	Original	Medicine (paediatrics)	Retrospective	This paper
Zarbo et al., 2005	*Referred to as*: a taxonomy of anatomic pathology error	Original	Medicine (pathology)	Retrospective	This paper

4. Discussion

4.1. Trial aims

Essential to the methodology of reliability studies has been the clear outlining of the purpose of the study. The most commonly reported rationale for the publishing of a reliability study is as part of the development of a new incident coding technique. In at least two cases reliability studies have been used to improve a technique during development by completing a revision of the technique wording or structure after the trial (Baker and Krokos, 2007; Baysari et al., 2011). Surprisingly in the majority of published papers a reliability study has only been included post development. The papers highlight the importance of validating the technique, however suggest that no improvement is required based on the results of the validation. The evidence shows however that this is not accurate in many cases with most of the techniques only showing moderate to substantial (40–80%) reliability (see Tables 3 and 4).

While at least half of all studies appear to incorporate a reliability trial for the purpose of validating a technique almost as many studies have used a reliability trial to make comparisons in support of the flexibility of the technique. Most commonly tested comparisons are made between groups of participants divided based on their profession (for example as incident investigators, human factors specialists or line workers) (Isaac et al., 2003a; Olsen, 2011) or between participants who have received training in the technique and/or human factors and those that have not (Stanton and Stevenage, 1998). The aim of these comparisons has been to determine how effectively the technique can be employed by personnel with limited experience and training. This is common in small organisations which have limited resources. Other studies have compared the reliability of two different techniques to highlight the strengths and weaknesses of each so that organisations can select that which is more suited to their requirements. Others have compared the reliability of an original technique in a particular field to the reliability of a new or derivative technique in another field in support of the need for organisations to choose a field-specific technique for their organisation. Comparison studies such as these are regularly used by high hazard industries to select techniques for their safety departments thus highlighting their continuing importance.

4.2. Trial preparation

During the trial preparation stage decisions as to what type of information to code, how many information events to code, how many coders should there be and what characteristics those coders should have are paramount. At least one study reported that there was greater consensus in some areas among coders who were experienced in human factors (Isaac et al., 2003a). On the other hand two studies have been reported highlighting no significant impacts of experience or job performance between coders (Gibson, 2006; Olsen, 2011).

Table 3
Characteristics of the reliability studies: trial aims and preparation.

Surname(s), year	Aim of the study[a]	Trial # within study	Study tests the identification of factors also?	No. and type of reports	No. and profession of coders
af Wåhlberg, 2002	Development of a new technique	1	y	122 full reports	2 'coder and assistant'
Baker and Krokos, 2007	Development of a new technique	1	n	44 full reports	6 pilots
		2	n	28 full reports	5 pilots
Baysari et al., 2009	Comparison of two different techniques	1	n	1–2 full reports	4 students
		2	n	3 full reports	5 students
Baysari et al., 2011	Development of a new derivative and comparison with original	1	NR	6 abridged reports	25 rail industry workers
		2	NR	6 abridged reports	11 students
Gibson, 2006	Validate existing tool	1	NR	NR scenarios	8/18 rail HF and risk specialists
		2	NR	14 scenarios	
Harris et al., 2005	Validate existing tool in new industry	1	n	1 scenario	8 students
		2	n	57 reported errors	8 students
Isaac et al., 2003a	Validation of existing technique and comparison between professions23 æ	1	n	8 full reports	27/7 ATC SME, HF specialists, investigators etc
		2	n	7 full reports	
Jacobs et al., 2007	Development of a new technique	1	y	84 full reports	2 students
Kaplan et al., 1998	Development of a new technique	1	y	25 full reports	5 technique developers and hospital staff
Krokos and Baker, 2005	Development of a new technique	1	y	44 full reports	7/6 FAA and airline personnel
		2	y	28 full reports	
Makeham et al., 2008	Development of a new taxonomy and comparison with a pilot version	1	y	433 full reports	3 GP investigators
		2	y	132 full reports	3 GP investigators
O'Connor 2008	Validation of an existing derivative	1	y	2 full reports	123 military aviation students and 2 HF specialists
O'Connor et al., 2010	Validation of an existing derivative	1	y	1 interview transcript	18 military aviation students
O'Connor and Walker, 2011	Validation of an existing derivative	1	y	2 full reports	204 military aviation students
Olsen and Shorrock, 2010	Validation of an existing derivative	1	y	2 abridged reports	11/1/4 air traffic controllers
		2	y	63 full reports	
		3	y	5 full reports	
Olsen, 2011	Comparison between different participant professions	1	n	14 abridged reports	4 air traffic controllers 3 HF specialists
Pounds and Isaac, 2003	Comparison between different participant nationalities	1	y	7 full reports	7/NR incident investigators and safety managers
		2	n	32 errors extracted from transcript	
Read et al., 2012	Test hypotheses using technique	1	y	95 full reports	6 'coders' (only 2 coding each report from the pool of 6)
RSSB, 2005	Validate existing tool in new industry	1	n	2 abridged reports	4/6 HF specialists and accident investigators
		2	n	7 abridged reports	
Shorrock and Kirwan, 2002	Development of a new technique	1	y	4 full reports	9 HF specialists
Stanton and Stevenage, 1998	Comparison between trained users and untrained users	1	y	1 scenario	98/31/25 students
Stanton et al., 2002	Validation of existing technique	1	y	1 scenario	8 graduate engineering students
Terhune, 1983	Development of a new technique	1	y	100 full reports	2 'coders'
		2	y	28 full reports	4 accident investigators
Wallace et al., 2002	Development of a new taxonomy and comparison with a previous technique	1	y	28 full reports	3/9 'coders'
		2	y	12 full reports	
Wallen Warner and Sandin, 2010	Validation of an existing technique	1	y	4 scenarios	7 'coders'
Woods, 2005	Development of a new technique	1	NR	314 full reports	3 investigators
Zarbo et al., 2005	Development of a new technique	1	NR	430 abridged reports	NR

NR = not reported; HF = human factors; n = factors are preidentified for coding; y = factors must be identified by coder for coding.
[a] All studies test intercoder consensus except Stanton and Stevenage (1998) who test intracoder consensus and Olsen and Shorrock (2010) who test intracoder consensus in 1 of 3 trials in the paper.

The resources of the researcher has limited the decisions made during the trial preparation however it is reasonable to expect that many small organizations employing incident coding techniques in their safety department will also be of similarly limited resources. For example, it is observed that many researchers limit their participant numbers to a small convenience sample who individually code a small sample of incident reports, scenarios or transcripts. Where a small number of participants are concerned these tend to be subject matter experts, investigators and specialists. Where larger numbers of participants have been employed this is due to the use of university or military aviation students – a decision that has had to be weighed up against the use of participants more experienced in human factors and investigation theory.

Surprisingly there appears to be no correlation between the number of reports coded and whether those reports are full or abridged lengths (see Table 3). Full reports have been used in just

Table 4
Characteristics of the reliability studies: trial conduct and analysis.

Surname(s), year	Training provision and format	Materials provided in trial[a]	Type of analysis tool	Selected specific data for analysis[b]	Trial no/trial average result level[c]
af Wåhlberg, 2002	None or NR	Code definitions	% (modal category)	'no choice' box for non-selection	1/Almost perfect
Baker and Krokos, 2007	Less than 1 day group training	None	% (modal category)	–	1/Moderate 2/Substantial
Baysari et al., 2009	None or NR	Code definitions and flow charts	% (type not specified)	Reported when at least two coders identified a common error	1/Fair to Almost perfect 2/Moderate
Baysari et al., 2011	40 min group training	None	Index of Concordance	–	1/Substantial 2/Substantial
Gibson, 2006	None or NR	Extract of relevant task analysis	% (modal category)	Reported when the criteria for one order of magnitude either side of the median is met	1/NR 2/ Almost perfect
Harris et al., 2005	Less than 1 day group training	Task analysis, photographs	1/Signal detection paradigm, 2/Pearson's *r*	–	1/Almost perfect 2/Substantial
Isaac et al., 2003a	Trial 1: 1 day Trial 2: 5 days	Workbook, flowcharts and tasks	% (modal category)	–	1/Moderate 2/Substantial
Jacobs et al., 2007	Reported as none	None	Kappa	–	1/Substantial
Kaplan et al., 1998	Less than 1 day group training	NR	Yule's *Q*	–	1/Substantial
Krokos and Baker, 2005	Less than 1 day group training	None	% (modal category), Kappa	Reported where at least 60% of participants provided an assignment	1/Moderate 2/ Moderate
Makeham et al., 2008	None or NR	None	Kappa	Factors were coded only if 2 of 3 coders agreed on the selected factor	1/Moderate 2/Substantial
O'Connor, 2008	2 h group training	None	Within group rater agreement (r_{wg})	Reported when at least 50% of coders selected a code	1/Substantial
O'Connor et al., 2010	2 h group training	Personnel involved in incident	% (modal category)	Not included in analysis if only one participant selected the code	1/Substantial
O'Connor and Walker, 2011	5 h group training	Engineering, medical reports and procedures	Multirater Kappa free	Reported for codes the groups thought were casual and for codes the groups thought were not causal	1/Moderate
Olsen and Shorrock, 2010	Reported as none	None	Index of Concordance	'no choice' box for non selection	1/Fair 2/Fair 3/Fair
Olsen, 2011	Selfpaced workbook	Workbook with definitions and examples	Index of Concordance	–	1/Moderate
Pounds and Isaac, 2003	5 days group training	None	% (type not specified), Kappa	–	1/Substantial 2/Substantial
Read et al., 2012	Reported or none	None	Kappa	Calculated for only 4 of 5 pairs	1/Moderate to almost perfect
RSSB, 2005	None or NR	None	Index of Concordance	–	1/NR 2/Moderate
Shorrock and Kirwan, 2002	None or NR	None	# of analysts selecting most often chosen category	–	1/Moderate
Stanton and Stevenage, 1998	One group provided training	Hierarchical task analysis	Signal detection paradigm	–	1/Control: fair, SHERPA: almost perfect
Stanton et al., 2002	Less than 1 day group training	Hierarchical task analysis, demonstration using flight simulator, photos	Signal detection paradigm	Pooled error predictions compared to reference errors rather than individual predictions compared to reference errors	1/Substantial to almost perfect
Terhune, 1983	Less than 1 day group training	NR	% (type not specified)	100 reports coded with only the last 25 coded reports being analysed	1/Almost perfect 2/Substantial
Wallace et al., 2002	None or NR	None	Index of Concordance	–	1/Substantial 2/Almost perfect
Wallen Warner and Sandin, 2010	Self paced training via email	Technique manual	% (type not specified)	'no choice' box for non selection	1/Almost perfect
Woods, 2005	None or NR	NR	% (type not specified)	–	1/Substantial to almost perfect
Zarbo et al., 2005	None or NR	NR	% (type not specified), Kappa	–	1/Almost perfect

[a] Other than the technique taxonomy, model or list itself and answer forms; NR = not reported; % = percentage agreement; # = number.
[b] Data was reported or assumed to be analysed in its entirety unless specified.
[c] Refer to Table 5 for reliability result levels.

Table 5
Reliability result levels, adapted from Fleiss (1981).

Level	Kappa, Yule's Q and Pearson's r	% agreement and index of concordance	Signal detection paradigm	r_{wg}	Number of analysts selecting most often chosen category
			Percentage of responses that fall into 'hits' or 'correct rejections'		*Total number of analysts = 9*
Poor	≤0	0%	0%	0	0
Slight	0.1–0.2	>0–20%	>0–20%	0.1–0.2	1
Fair	0.21–0.4	21–40%	21–40%	0.21–0.4	2–3
Moderate	0.41–0.6	41–60%	41–60%	0.41–0.6	4–5
Substantial	0.61–0.8	61–80%	61–80%	0.61–0.8	6–7
Almost perfect	0.81–1	81–100%	81–100%	0.81–1	8–9

over half of the studies however the length and detail of full reports vary widely from industry to industry. While abridged reports have been used in one third of studies it has been argued that it is more realistic to provide coders with as much information as exists, including medical reports, technical reports, transcripts, event reports and procedures (O'Connor and Walker, 2011). In most studies where full reports have been provided participants have been required to identify relevant factors contributing to an incident as well as code the factor. The majority of researchers choosing not to supply full reports have instead provided pre-identified factors (often with abridged reports for contextual information) to ensure that only the coding of factors is tested and not the *selection* of factors to code being tested. There is evidence to suggest that this provides the most accurate assessment of coding reliability with the requirement to select the factor to code decreasing reliability by at least ten percent (Wallace et al., 2002).

4.3. Trial conduct

Whether to provide training to the participants, and how much training, appears to be one of the most contentious topics in reliability study methodology. Krippendorf (2004) argues that high agreement is only reached by giving extensive training, a principle which has been followed in a number of studies (O'Connor, 2008; O'Connor et al., 2010; Stanton and Stevenage, 1998; Wallen Warner and Sandin, 2010). In one study it was asserted that the optimal training period should be five days, however evidence for this assertion and an explanation were not provided (Isaac et al., 2003a). Furthermore another study showed evidence that training participants only improved reliability in some areas of the technique and not in others (Stanton and Stevenage, 1998).

Additionally no correlation could be found in this review between high reliability and the format of the training (by email, workbook or in group sessions) though conceivably this could also have an effect on the results where those trained in group sessions have the benefit of being party to discussion and ask questions of the trainer. The effect of such discussion during training may be similar in effect to the discussion between developer-coders where it has been argued that high reliability could be a result of participants learning each other's styles during the development meetings (Makeham et al., 2008). Furthermore where coders are permitted to discuss their code selections between each individual coding attempt evidence suggests that improvement again may be the result of the coders' ability to learn each other's style rather than agreement based solely on use of the technique (Terhune, 1983).

On the other hand it has been highlighted that in many organisations coding systems are more likely used by teams on investigation boards, over a number of weeks and with freedom to discuss aspects of the cases as well as coding selection (O'Connor and Walker, 2011). While this may be so in large organisations where safety boards investigate high-consequence accidents with many contributing factors, small organisations with a single investigator, coding minor event and incident reports do not necessarily have the resources or time for this manner of investigation. Balancing between providing an accurate context for the trial based on the context in which the technique is intended and testing reliability solely based on the technique is an art that many studies have had to negotiate.

Of final consideration are the materials provided to the participants in the trial. While the majority of studies provide only the technique and a coding form for participants, other studies provide manuals containing definitions (af Wåhlberg, 2002; Isaac et al., 2003a; Olsen, 2011; Wallen Warner and Sandin, 2010), task descriptions or task analysis proformas (Gibson, 2006; Stanton and Stevenage, 1998), flow charts (Isaac et al., 2003a) or interviews with personnel involved in the incident (O'Connor et al., 2010; Pounds and Isaac, 2003). These materials are designed to clarify the technique terminology, technique use and reported information and are an important tool in ensuring that participants are effectively able to use the technique as intended for accurate results. However there appears to be no correlation between the provision of extra materials and reliability result levels (see Table 4).

4.4. Analysis and reporting

By far the most debated decision in reliability study methodology is the selection of the analysis tool. In particular the debate over the use of Kappa (Cohen, 1960), a method which reduces the observed agreement by the portion of agreement that would be expected by chance alone, has been the subject of numerous papers (Byrt et al., 1993; Cicchetti and Feinstein, 1990; Feinstein and Cicchetti, 1990; Hubert, 1977; Ross et al., 2004; RSSB, 2005). Firstly, it has been argued that correcting for chance agreement is not appropriate in trials where participants do not start from a position of complete ignorance but make selections based on their professional knowledge, experience and on the technique's definitions and guidance (Olsen and Shorrock, 2010). Secondly, it has been noted that reported high Kappa results may be misleading with high levels of agreement on *unselected* codes (that is that participants could agree on which codes were *not* contributory) masking unacceptable levels of *selected* codes (codes that participants thought *were* contributory) (O'Connor et al., 2010). Finally, extremely low values of Kappa often result in reliability studies of incident coding despite high observed agreement. This is due to its susceptibility to prevalence, a condition in which the chance agreement value is high due to a skewed distribution of agreement in the marginals.

Similarly ambiguity and appropriateness of two measures of percentage agreement have been highlighted (Olsen and Shorrock, 2010; Ross et al., 2004). Half of reliability studies use percentage agreement as an indicator of reliability where the percentage of participants agreeing on the modal selected code is reported (Baker and Krokos, 2007; Isaac et al., 2003a). For example, if two of three coders agree on the modal selected code then the calculated

percentage agreement is 67%. This method shows clearly how many participants agree on that particular code, however does not provide any detail on how many other agreements there were- suppose if all other participants agreed on a second code? – or if all other participants selected all different codes? Four studies use the Index of Concordance (Martin and Bateson, 2001) to calculate agreement by dividing the total number of agreements between pairs with the total number of agreements possible between all pairs. Using the formula for the Index of Concordance, $IOC = A/(A + D)$ where A is the total number of agreements and D is the total number of disagreements, the agreement calculated where two of three coders agree is 33% (there is one agreement and two disagreements). This method is very harsh as it takes into account all disagreements between coder pairs however considers multiple agreements on different codes and can likewise be reported as a percentage. RSSB (2005) provides an example where the reported agreement differed by as much as 28% between agreement calculated using the Index of Concordance and percentage agreement on the modal category.

Similarly, it is important to distinguish between consensus and consistency as described in the introduction of this review when referring to agreement measures of reliability. One reliability study has provided evidence that a highly consistent pattern of codes can be derived from the use of a coding technique despite low reliability (Wallace et al., 2002) while one other study has inappropriately used frequencies to test reliability resulting in consistency, not consensus, being reported (Stanton and Stevenage, 1998).

Other methods of analysis that have been used include r_{wg} (James et al., 1993), Pearson's r, Yule's Q, sensitivity matrices and signal detection paradigm (Macmillan and Creelman, 1991). The latter two were used in studies of predictive techniques. Furthermore significance tests such as Chi^2 and McNemar Test have been in four of the twenty five studies and used to report significance of results however at least one study argues that statistical significance is not generally regarded as a useful method for interpreting Kappa as relatively low values of Kappa can still be significant (O'Connor and Walker, 2011). This argument can also hold true for percentage agreement measures in the incident coding domain.

Surprisingly only a small percentage of studies use comparison statistics to place their results in context. These values vary greatly with no two reports by different researchers using the same measure. On average the value used for satisfactory percentage agreement is 76% with accepted values as low as 70% (0.4 for Kappa) and as high as 88% (Wallace et al., 2002; Wallen Warner and Sandin, 2010). However no standard measure of acceptable agreement in coding consensus of incidents has been identified with measures taken instead from general benchmarks suggested by analysis tool developers or other social science fields.

No consistent method of selecting data for analysis has been identified either and the large variance in how data are selected has significant impacts on the results. Data selection methods have included the reporting of agreement only in instances where at least two raters identified a common error (Baysari et al., 2009), where at least a certain percentage of participants (typically 50–60%) provided an assignment (Krokos and Baker, 2005; O'Connor, 2008), if two of three coders agreed on a factor to be coded (Makeham et al., 2008) and by employing the criteria for one order of magnitude either side of the median (Gibson, 2006). These methods, along with disregarding codes where only one participant has selected a code (O'Connor et al., 2010) and codes assigned during the first 75% of the trial (Terhune, 1983), reduce the number of disagreements considered in analysis therefore artificially increasing the overall agreement reported for the technique. Possibly such selections are made to mitigate the reduction of agreement due to participants being required to select the factor as well as assign a code. However it is more appropriate to amend the conduct of the trial to include only preselected factors and would provide more accurate and genuine results.

Other studies have taken steps to ensure that the calculation of disagreements is as accurate as the calculation of agreements. To ensure this, where an uneven number of factors have been identified or codes have been assigned researchers have provided a 'no choice' or 'missing' box in place of the non selection (af Wåhlberg, 2002; Olsen and Shorrock, 2010; Wallen Warner and Sandin, 2010). Where other participants also have made a non selection, this is counted as an agreement; where they have made a selection this is counted as a disagreement. This method ensures that agreement is not reported as high based on the fact that data are not included because uneven selections have been made. Similarly, studies reporting Kappa results where Kappa is reported for selected codes and unselected codes separately also provide a clearer picture of whether the high agreement is due to the unselected codes or selected codes (O'Connor and Walker, 2011).

4.5. Inadequate reporting and unsupported conclusions

The importance of appropriate data selection and clear reporting of results cannot be overstated. However considering the lack of reliability studies in the field of incident coding it is unsurprising that the many of the few studies that exist are plagued with questionable data selection, poor reporting of results and underreporting of trial preparation and methodology. It is clear that the decisions made at all stages of the trial affects the outcome and places the results in context: without method details an accurate appreciation of technique reliability is impossible to achieve. At least two of the papers reviewed (Woods, 2005; Zarbo et al., 2005) presented very little information on the trial preparation, data selection and analysis of results and no information on the conduct of the trial. Two sources were also excluded from the review for reporting a single overall reliability result only with no method details at all (van Vuuren et al., 1997; Wiegmann and Shappell, 2003). Wallace et al. (2002) argues that there is no excuse for the exclusion of reliability studies in incident coding research however the argument could also be made for the underreporting of such studies.

Unsupported conclusions are also prevalent in reliability study publications. A number of studies report that the trialed technique was reliable or performed well however the published results suggest that the techniques produced on average moderate reliability (Krokos and Baker, 2005; Stanton and Stevenage, 1998). One study (Jacobs et al., 2007) also reported that their developed technique was more suitable for use in the Canadian context, however a comparison trial was not conducted using a non-Canadian technique, there were no apparent decisions in the technique development that were based on cultural variances and no explanations as to how the new technique was more appropriate to a Canadian user base were provided. A further study boasted good intra-coder reliability over time using a particular technique (Stanton and Stevenage, 1998) however the time period was less than two months, the same one report was coded each time and feedback provided in between each of the three coding tasks was given suggesting that the users' ability to learn answers was tested rather than their independent use of the technique. It is important that researchers conducting reliability studies do not go beyond the data they are presenting as it is the aim of the reliability study to provide as accurate assessment of the technique's ability to provide consensus as possible.

5. Implications

It is clear that there are very few reliability studies that have been published compared to the vast number of

techniques available for incident coding in high hazard industries. However it is also clear that those studies vary greatly in their methodological soundness, ability to be generalised and reporting adequacy. Limitations in the resources of researchers, participant availability and time constraints have reduced the number and experience of participants and resulted in the necessity for abbreviating reports. Although this may be a realistic representation of many small organisations it may reduce confidence in the reliability results. Sample size requirements for adequate confidence in reliability results has been explored in detail in Donner and Eliasziw (1987) and Shoukri et al. (2004).

Furthermore the meaningfulness of an increased pool of reliability data is likely to be reduced when results cannot be compared between study designs or when particular design decisions and methods are subject to such vigorous debate. Where an accurate measurement of reliability data cannot be obtained, poor selection of technique by organisations will dramatically affect trend analysis and error mitigation or elimination of errors. It is therefore doubtful whether the current pool of reliability data has been effective in assessing the consensus of incident coding techniques. The inability to compare studies, the use of controversial methods and inconsistencies in the selection of participants and reports not only make it impossible to determine an appropriate level of reliability in the incident coding industry, but also make it difficult to make judgements on the reliability of incident coding techniques individually.

6. Conclusion

This review is limited by the search parameters employed in the literature search. In particular the exclusion of conference proceedings, the limiting of the search to three electronic databases prior to April 2012 and non-English language publications reduce the volume of evidence which is already small and varied. Furthermore, the wide variances in methodologies employed in the reviewed studies, and the numerous techniques used and contexts they are used in has made comparisons difficult in such a small body of evidence. Additionally judgements made by the single reviewer may have affected the inclusion and exclusion of studies however the author has aimed to be very transparent on judgements made. Nevertheless time and resource constraints meant that limits had to be set as in all other reviews.

Despite these limitations, this review provides evidence that the meaningfulness of and confidence in results is directly affected by the methodologies employed by the researcher during the preparation, conduct and analysis of the reliability study. Furthermore, the great variance in methodologies reduces the ability of client organisations to effectively compare techniques being considered for the adoption of report coding and trend analysis, a highly critical step in the implementation of safety and risk measures. Future research should evaluate the effectiveness of the current pool of reliability data in the incident coding domain and if it is necessary to improve data effectiveness, focus on the standardisation of reliability research and measurement through the development of a methodological framework for reliability studies in the safety management field.

Acknowledgements

The author kindly thanks two anonymous reviewers and the journal editor for feedback and guidance on an earlier version of this paper.

References

af Wåhlberg, A.E., 2002. Characteristics of low speed accidents with buses in public transport. Accident Analysis & Prevention 34, 637–647.
Australian Government, 2009. Defence Aviation Safety Manual. Directorate of Defence Aviation and Air Force Safety.
Baker, D., Krokos, K., 2007. Development and validation of aviation causal contributors for error reporting systems (ACCERS). Human Factors: The Journal of the Human Factors and Ergonomics Society 49, 185–199.
Baysari, M.T., Caponecchia, C., McIntosh, A.S., Wilson, J.R., 2009. Classification of errors contributing to rail incidents and accidents: a comparison of two human error identification techniques. Safety Science 47, 948–957.
Baysari, M.T., Caponecchia, C., McIntosh, A.S., 2011. A Reliability and Usability Study of TRACEr-RAV: The Technique for the Retrospective Analysis of Cognitive Errors-for Rail. Australian version.
Beaubien, J., Baker, D., 2002. A Review of Selected Aviation Human Factors Taxonomies, Accident/Incident Reporting Systems, and Data Collection Tools. FAA report.
Byrt, T., Bishop, J., Carlin, J.B., 1993. Bias, prevalence and kappa. Journal of Clinical Epidemiology 46, 423–429.
Cicchetti, D.V., Feinstein, A.R., 1990. High agreement but low kappa: II. Resolving the paradoxes. Journal of Clinical Epidemiology 43, 551–558.
Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46.
Donner, A., Eliasziw, M., 1987. Sample size requirements for reliability studies. Statistics in Medicine 6, 441–448.
Embrey, D., 1986. SHERPA: a systematic human error reduction and prediction approach. Paper presented at the International Topical Meeting on Advances in Human Factors in Nuclear Power Systems, Knoxville, Tennesse.
Fagerlind, H., Bjorkman, K., Wallen Warner, H., Ljung Aust, M., Sandin, J., Morris, A., Talbot, R., Danton, R., Giustiniani, G., Shingo Usami, D., Parkkari, K., Jaensch, M., Verschragen, E., 2008. Development of an in-depth European accident causation database and the driving reliability and error analysis method, DREAM 3.0, Proceedings of 3rd International Conference on Expert Symposium on Accident Research (ESAR), Hanover, Germany.
Feinstein, A.R., Cicchetti, D.V., 1990. High agreement but low kappa: I. the problems of two paradoxes. Journal of Clinical Epidemiology 43, 543–549.
Fleishman, E., Mumford, D., 1991. Evaluating classifications of job behaviour: a construct validation of the ability requirement scales. Personnel Psychology 44, 523–575.
Fleiss, J., 1981. Statistical Methods for Rates and Proportions, 2nd ed. Academic Press Inc, NY.
Gibson, H., 2006. T270 Phase 3: User Trial of the Rail Specific HRA Technique. Gibson Ergonomics Ltd, UK.
Harris, D., Stanton, N.A., Marshall, A., Young, M., Demagalski, J., Salmon, P.M., 2005. Using sherpa to predict design-induced error on the flight deck. Aerospace Science and Technology 9, 525–532.
Hubert, L., 1977. Kappa revisited. Psychological Bulletin 84, 289–297.
Hughes, T., Heupel, K., Musselman, B., Hendrickson, E., 2007. Preliminary investigation of interrater reliability of the department of defense human factors analysis and classification system in USAF mishaps. [Abstract] Aviation, Space, and Environmental Medicine 78, 255.
Isaac, A., Lyons, M., Bove, T., Van Damme, D., 2003a. Validation of the Human Error in ATM (HERA-JANUS) Technique. EUROCONTROL.
Isaac, A., Shorrock, S., Kennedy, R., Kirwan, B., Andersen, H., Bove, T., 2003b. The Human Error in ATM Technique (HERA-JANUS). EUROCONTROL.
Jacobs, S., O'Beirne, M., Derflingher, L., Vlach, L., Rosser, W., Drummond, N., 2007. Errors and advserse events in family medicine. Canadian Family Physician 53, 270–276.
James, L., Demaree, R., Wolf, G., 1993. rwg: an assessment of within-group interrater agreement. Journal of Applied Psychology 78, 306–309.
Kaplan, H., Battles, J.B., Van Der Schaaf, T.W., Shea, C.E., Mercer, S.Q., 1998. Identification and classification of the causes of events in transfusion medicine. Transfusion 38, 1071–1081.
Kirwan, B., 1998. Human error identification techniques for risk assessment of high risk systems - part 1: review and evaluation of techniques. Applied Ergonomics 29, 157–177.
Kozlowski, S., Hattrup, K., 1992. A disagreement about within-group agreement: distenangling issues of consistency versus consensus. Journal of Applied Psychology 77, 161–167.
Krippendorf, K., 2004. Content Analysis: An Introduction to its Methodology, 2nd ed. Sage, Thousand Oaks, CA.
Krokos, K., Baker, D., 2005. Development of a Taxonomy of Causal Contributors for Use with ASAP Reporting Systems. Technical Report FAA Grant # 99-G-048.
Ljung, M., 2002. DREAM - Driving Reliability and Error Analysis Method (M.Sc. thesis), Linköping: Linköping University.
Loeb, J., Chang, A., 2003. Reduction of Adverse Events through Common Understanding and Common Reporting Tools: Towards an International Patient Safety Taxonomy. A Review of the Literature on Existing Classification Schemes for Adverse Events and Near Misses. World Health Organization report.
Macmillan, N., Creelman, C., 1991. Signal Detection Theory: a User's Guide. Cambridge University Press, Cambridge.
Makeham, M.A.B., Stromer, S., Bridges-Webb, C., Mira, M., Saltman, D.C., Cooper, C., Kidd, M.R., 2008. Patient safety events reported in general practice: a taxonomy. Quality and Safety in Health Care 17, 53–57.

Martin, P., Bateson, P., 2001. Measuring Behaviour: An Introductory Guide, second ed. Cambridge University Press, UK.
O'Connor, P., Walker, P., 2011. Evaluation of a human factors analysis and classification system as used by simulated mishap boards. Aviation, Space, and Environmental Medicine 82, 44–48.
O'Connor, P., Walliser, J., Philips, E., 2010. Evaluation of a human factors analysis and classification system used by trained raters. Aviation, Space, and Environmental Medicine 81, 957–960.
O'Connor, P., 2008. HFACS with an additional layer of granularity: validity and utility in accident analysis. Aviation, Space, and Environmental Medicine 79, 599–606.
Olsen, N., Shorrock, S., 2010. Evaluation of the HFACS-ADF safety classification system: inter-coder consensus and intra-coder consistency. Accident Analysis & Prevention 42, 437–444.
Olsen, N., 2011. Coding ATC incident data using HFACS: inter-coder consensus. Safety Science 49, 1365–1370.
Pounds, J., Isaac, A., 2003. Validation of the JANUS Technique: Causal Factors of Human Error in Operational Errors. FAA Report.
Rail Safety Regulators' Panel, 2009. Contributing Factors Framework Manual. Rail Safety Regulators' Panel, Fortitude Valley. Retrieved from. http://www.rsrp.asn.au/publications.cfm?pub_id=27 (accessed 1 April 2012).
Read, G., Lenné, M., Moss, S., 2012. Associations between task, training and social environmental factors and error types involved in rail incidents and accidents. Accident Analysis and Prevention 48, 416–422.
Reason, J., 1990. Human Error. Cambridge University Press, NY.
Ross, A., Wallace, B., Davies, J., 2004. Technical note: measurement issues in taxonomic reliability. Safety Science 42, 771–778.
RSSB, 2005. Rail-specific HRA Tool for Driving Tasks, United Kingdom.
Shorrock, S., Kirwan, B., 2002. Development and application of a human error identification tool for air traffic control. Applied Ergonomics 33, 319–336.
Shoukri, M., Asyali, M., Donner, A., 2004. Sample size requirements for the design of reliability study: review and new results. Statistics in Medicine 13, 251–271.
Stanton, N.A., Stevenage, V.S., 1998. Learning to predict human error: issues of acceptability, reliability and validity. Ergonomics 41, 1737–1756.
Stanton, N.A., Young, M., Salmon, P.M., Harris, D., Demagalski, J., Marshall, A., Waldman, T., Dekker, S., 2002. Predicting pilot error: assessing the performance of SHERPA. In: Johnson, C.W. (Ed.), 21st European Conference on Human Decision Making and Control. GIST Technical Report G2002-1, Glasgow, Scotland, p. 47.
Terhune, K.W., 1983. CALAX: a collision taxonomy for research and traffic records. Journal of Safety Research 14, 13–20.
Tinsley, H., Weiss, D., 1975. Interrater reliability and agreement of subjective judgments. Journal of Counseling Psychology 22, 358–376.
van Vuuren, W., Shea, C., van der Schaaf, T., 1997. The Development of an Incident Analysis Tool for the Medical Field. Eindhoven University of Technology, Eindhoven, The Netherlands.
Wallace, B., Ross, A., Davies, J., Wright, L., White, M., 2002. The creation of a new minor event coding system. Cognition, Technology and Work 4, 1–8.
Wallen Warner, H., Sandin, J., 2010. The intercoder agreement when using the driving reliability and error analysis method in road traffic accident investigations. Safety Science 48, 527–536.
West, R., Elander, J., French, D., 1991. Can road traffic accidents be reliably and usefully classified? In: Grayson, G., Lester, J. (Eds.), Behavioural Research in Road Safety. Transport and Road Research Laboratory, Crowthorne, pp. 59–67.
Wiegmann, D., Shappell, S., 2001. Assessing the reliability of the human factors analysis and classification system (HFACS) within the context of general aviation. Aviation, Space, and Environmental Medicine, 266.
Wiegmann, D., Shappell, S., 2003. A Human Error Approach to Aviation Accident Analysis: The Human Factors Analysis and Classification System. Ashgate, Aldershot.
Woods, D.M., 2005. Anatomy of a patient safety event: a pediatric patient safety taxonomy. Quality and Safety in Health Care 14, 422–427.
Zarbo, R., Meier, F., Raab, S., 2005. Error detection in anatomic pathology. Archives of Pathology and Laboratory Medicine 129, 1237–1245.