Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics AHFE 2014, Kraków, Poland 19-23 July 2014
Edited by T. Ahram, W. Karwowski and T. Marek

A test of a systems theory-based incident coding taxonomy for risk managers

Natassia Goode[1], Paul M. Salmon[1], Michael G. Lenné[2], Caroline F. Finch[3]

[1]University of the Sunshine Coast Accident Research
University of the Sunshine Coast
Maroochydore, Australia

[2]Monash University Accident Research Centre
Monash University
Melbourne, Australia

[3]Australian Centre for Research into Injury in Sport and its Prevention (ACRSIP)
Federation University Australia
Ballarat, Australia

ABSTRACT

Organizations need to be able to collect reliable and accurate data on the causal factors that lead to near misses and injury causing incidents in order to design appropriate, informed, safety interventions. The aim of this study was to test the inter-rater reliability of a prototype taxonomy for classifying the causal factors involved in incidents in the outdoor education and recreation sector. The taxonomy consists of three levels, where each category level breaks the previous one down into a finer level of detail. The study involved 14 respondents, who play a key role in risk management within their organization, using the taxonomy to code 10 detailed incident reports. The incident reports were composited from reports and enquiries into actual events, and ranged in injury severity from fractures to fatalities. Participants were asked to: 1) identify the causal factors involved in each incident; and 2) identify the code/s from the taxonomy which best described those causal factors. The study demonstrated that the taxonomy can be used by risk managers to identify and code causal factors across all levels of the led outdoor activity system. However, identifying appropriate codes at the second and third level of detail was problematic.

Keywords: Systems-thinking, accident analysis, incident reports

INTRODUCTION

There is an acknowledged risk of both severe and frequent injury associated with "led" (i.e. facilitated or instructed) outdoor education and recreation activities such as kayaking, rocking climbing and abseiling (Cessford, 2013; Dickson & Gray, 2012). In order to understand these risks, many organizations in the outdoor education and recreation sector collect incident reports. However, as in other domains (Gordon, Flin, & Mearns, 2005), most reporting systems do not capture a complete picture of the causal factors associated with incidents. Further, few providers have the capacity to analyze data at the aggregate level to identify trends (Goode, Finch, Cassell, Lenne, & Salmon, In Press). For learning from incidents to occur, risk managers must be able to collect data on the causal factors involved in near misses and injury causing incidents *and* the coding scheme must support the reliable and accurate analysis of the data collected.

The Understanding and Preventing Led Outdoor Accidents Data System (UPLOADS) project is an attempt to address this need via the following stages: 1) development of a theoretical framework for analyzing incidents during led outdoor activities (Salmon, Cornelissen, & Trotter, 2012; Salmon, Goode, Lenne, Cassell, & Finch, 2014; Salmon, Williamson, Lenné, Mitsopoulos-Rubens, & Rudin-Brown, 2010); 2) development of a prototype incident reporting, storage and analysis methods, including robust data coding systems and causal factor taxonomy; 3) implementation of the tool across participating organizations in Australia; and 4) development of a systems-based model of accident causation for the led outdoor activity domain to guide injury prevention efforts.

The current paper reports on the second phase, specifically examining the inter-rater reliability of a prototype version of the causal factor taxonomy. Inter-rater reliability refers to the level of agreement among two or more coders (Neuendorf, 2002); the higher the level of agreement among coders, the higher the inter-rater reliability of the coding taxonomy. Such a focus on demonstrating a high level of inter-rater reliability is important for a number of reasons. First, a high level of inter-rater reliability demonstrates that the classification system is logically organized and parsimonious (Fleishman & Quaintance, 1984). Second, as part of the UPLOADS project, organizations will contribute the data they collect to an industry dataset. This information will then be used by professional associations and government agencies to make evidence-based decisions about risk management issues that affect those involved in the provision of led outdoor activities. Finally, Human Factors methods, such as error prediction and coding schemes are often criticized due to the lack of reliability evidence associated with them (Stanton & Young, 1999, 2003). Testing coding schemes throughout their development life cycle is therefore a necessity to ensure that the end product is reliable.

The taxonomy is underpinned by a systems-theory model of accident causation, Rasmussen's (1997) Risk Management Framework (RRMF). Rasmussen's framework is underpinned by the idea that sociotechnical systems comprise various levels; actions and decisions across these levels interact with one another and contribute to the control of hazardous processes. In a series of previous studies (Salmon et al., 2012; Salmon et al., 2014), RRMF was adapted to describe the "led outdoor activity system" as a hierarchy across multiple levels including: government policy and budgeting; regulatory bodies and associations; activity centre planning, management and budgeting, local area government, parents and schools; technical and operational management; physical processes and instructor/participant activities; and equipment and surroundings. This framework has been validated through the analysis of case studies of fatal led outdoor incidents (Salmon et al., 2012; Salmon et al., 2010) and less severe injury causing incidents in the outdoors (Salmon et al., 2014).

Development of the prototype taxonomy was informed by three activities: (1) an analysis of 1014 led outdoor injury and near miss incidents (Salmon et al., 2014); (2) a review of the accident causation literature; and (3) a review of existing accident analysis methods. The intention was to develop a taxonomy that could be used by risk managers in the outdoor sector with minimal training to code their own incident data. In addition, the taxonomy needed to have enough detail that aggregate analyses of incident reports could immediately be used to generate meaningful injury prevention strategies, without further data coding.

Table 1 shows the taxonomy in the context of the adapted RRMF. The taxonomy consists of three levels of categories. The first level describes the outdoor activity 'system' in terms of the activity context; the key people involved in the activity; and the people and agencies that impact on how the activity is run. The second level breaks the first level categories down into more descriptive categories. The third level breaks the second level categories down into between 2 and 19 highly specific causal factors; giving a total of 325 highly specific causal factors. The three level structure was devised because incident reports contain varying amounts of detail and sometimes it is not possible to pin-point a highly specific causal factor. For example, "There was a problem with the equipment" can only be coded to "1: Equipment", while a more specific description of the problem, "Participant only brought thongs for the bushwalk" can be coded to: "1.2d: Lack of clothing/PPE." This 'decision gate' style system is a standard feature of coding taxonomies. However, in other domains, it has been found that increasing specificity in coding can result in a loss of inter-rater reliability (Finch et al., 2012; O'Conner, 2008).

In summary, the aim of this study was to test the inter-rater reliability of the prototype taxonomy for classifying the causal factors involved in incident reports for risk managers in the outdoor sector. Inter-rater reliability was examined for each level of the taxonomy (i.e. Level 1: The led outdoor activity system; Level 2: Descriptive categories; and Level 3: Specific causal factors).

METHOD

Sample

Managers and staff who play a key role in risk management within their organization were invited to participate through Victorian Outdoor Providers Network (VOPN) meetings and email lists. VOPN is collaboration between the six major Outdoor Education and Recreation employers in Victoria, Australia. VOPN collectively employ 600 staff and lead 32,000 students on Outdoor Education programs each year. Through these invitations, 17 people volunteered to participate.

Coding booklets were placed in blank envelopes along with a postage paid return envelope for the completed booklets. These were delivered by mail to the quarterly VOPN meeting and to people who volunteered via email. Completed booklets were returned directly to the University of the Sunshine Coast. Ethics approval for the study was granted by the University of the Sunshine Coast Human Ethics Committee.

Table 1: Overview of coding taxonomy in the context of Rasmussen's Risk Management Framework. Level 1 factors are shown in bold. Level 2 factors are presented below each corresponding Level 3 factor. Number of Level 3 codes indicated in brackets.

State and Federal Government policy and budgeting	10. Government 10.1 Budgetary constraints (2) 10.2 Infrastructure and land (2) 10.3 Policy and legislation (5)		
Regulatory bodies and associations	9. Regulatory bodies 9.1 Auditing (4) 9.2 Regulatory bodies (5)		
Activity Centre planning, management and budgeting; Local area govt; Parents and Schools	7. Activity Centre management 7.1 Activity training programs (11) 7.2 Organizational characteristics and constraints (11) 7.3 Practices (7) 7.4 Procedures (10) 7.5 Risk/hazard management systems (10)	8. Local area government, schools and parents 8.1 Local area government (3) 8.2 Schools (8) 8.3 Parents (7)	
Technical and Operational Supervision	6. Supervision/management 6.1 Planning and activity program (19) 6.2 Safety management (4) 6.3 Staff and staffing (7) 6.4 Supervision (10)		
Physical processes and instructor/ participant activities	3. Instructor 3.1 Communications (5) 3.2 Compliance (4) 3.3 Decision (4) 3.4 Demonstration (5) 3.5 Experience, qualifications and competence (5) 3.6 Leadership (3) 3.7 Mental condition (7) 3.8 Perception (3) 3.9 Physical condition (9) 3.10 Planning and preparation (7) 3.11 Safety (4) 3.12 Unsafe acts (6) 3.13 Violations (3)	4. Participant 4.1 Communications (4) 4.2 Compliance (2) 4.3 Decision (4) 4.4 Demonstration (3) 4.5 Experience and competence (4) 4.6 Mental condition (7) 4.7 Perception (3) 4.8 Physical condition (10) 4.9 Training and Practice (3) 4.10 Unsafe acts (6) 4.11 Violations (3)	5. Group (19)
Equipment and surroundings	1. Equipment 1.1 Activity equipment (7) 1.2 Clothing and PPE (7) 1.3 Documentation (5) 1.4 Food and drink (4) 1.5 Medication (3)	2. Environmental 2.1 Temperature (3) 2.2 Weather (7) 2.3 Miscellaneous (7) 2.4 Animals and insects (3) 2.5 Physical Environment (6) 2.6 Terrain (5) 2.7 Trees and Vegetation (3) 2.8 Water (7)	

Coding booklet content

Respondents were required to provide brief demographic information including their gender, age, current role and years of experience within the outdoor education and recreation sector, whether they currently lead activities as part of their role, and whether they have specific outdoor education and recreation qualifications.

In the next section of the booklet, the causal factor taxonomy was explained in detail. The codes were presented as a multi-level list, where Level 1 codes were numbered 1, 2, 3 etc.; Level 2 codes were numbered as 1.1, 1.2 1.3 etc.; and Level 3 codes were numbered as 1.1a, 1.1b, 1.1c etc. The three level structure of the causal factor taxonomy was explained, describing how each level breaks the previous level categories down into a finer level of detail.

Participants were instructed that the appropriate level of coding depended on the level of detail in the incident report, and that depending on the level of detail it was appropriate to choose a first level, second level or third level category to describe a causal factor. A number of examples were then provided, illustrating how the

different levels of the taxonomy apply to different levels of detail. The codes at each level of the framework were then described. In addition, Level 3 categories were described with reference to specific examples of causal factors.

The next section of the booklet presented 10 one page incident reports. Respondents were asked to: 1) identify the causal factors involved in each incident; and 2) identify the code/s from the taxonomy which best described those causal factors. In addition, respondents were instructed not to go beyond the details contained in the actual reports. Respondents were provided with a separate copy of the coding taxonomy to assist in the identification of appropriate codes.

The incident reports were reports of actual events. Eight reports were taken from the Australian Accident Register, which is an online publicly available depository of voluntary reports of accidents and serious near misses from the outdoors and adventure community (https://groups.google.com/forum/#!forum/australian-accident-register). Two reports were summaries of Coroner's inquests into outdoor education fatalities. The reports were selected for the study because they incorporated a range of human factors causes, had sufficient detail, and were easy to understand. In addition, they were selected to represent a range of injury severities (e.g. from fractures to fatal injuries), outdoor activities (e.g. kayaking, abseiling, rock climbing, swimming, canyoning, bushwalking) and participants (e.g. school and university students, young adults, older adults). In each report, any identifying details were changed to avoid referring to actual people and locations.

Data analysis

To assess the inter-rater reliability, the presence (coded 1) or absence (coded 0) of each code was recorded for each respondent for each incident. For each incident, the inter-rater reliability of each code was analysed using: 1) the within-group inter-rater reliability coefficient (r_{wg}); and 2) the percentage of respondents who selected the code. James, Demaree, and Wolf (1984, 1993) define r_{wg} as the proportional reduction in error variance of a distribution of obtained responses compared to a distribution representing a random response pattern in which the frequency of the responses is equal for each possible point on the scale. The equation for r_{wg} is:

$$r_{wg} = 1 - \left(\frac{S_x^2}{\sigma EU^2}\right)$$

where S_x^2 is the variance of the observed and σEU2 is the population variance of a discrete rectangular distribution of the responses. The equation for this is: σEU2 = (A2 − 1)/12, where A is the number of possible alternatives in the rating scale. In this case, there were 2 possible responses for each code: '1 and '0. Values of r_{wg} can vary from 0 to 1, where a score of 1 denotes perfect agreement between respondents. When the variance of the obtained ratings is random, then r_{wg} is equal to 0, reflecting no agreement between respondents.

There are no established criteria for interpreting r_{wg}. O'Connor (2008) suggests that $r_{wg} \geq .6$ indicates substantial agreement between the raters. However, in the current study, $r_{wg} \geq 0.6$ would only be attained if 13 out of 14 respondents selected or rejected a code (e.g. 92% agreement). Therefore, $r_{wg} \geq .45$ was used as an indicator of substantial agreement in the current study; this reflects an 85% level of agreement.

One criticism of r_{wg} is that it does not differentiate between agreement that a code should be rejected and agreement that a code is relevant. That is, if 100% of the sample selects or rejects the code, then r_{wg} will be 1. This is particularly problematic in large taxonomies, such as UPLOADS, where the majority of codes are unlikely to apply to any single incident. Consequently, r_{wg} will be artificially inflated by correct rejections. To overcome this problem in the current study for each incident r_{wg} was calculated only for items that were selected by at least one respondent. These values were then averaged across the 10 incidents. In addition, the percentage of respondents who selected each code was calculated for each incident (referred to as percent agreement). Again, these values were then averaged across the 10 incidents. This is also consistent with the advice of LeBreton and Senter (2008) which encourages researchers to use multiple indices to aid in interpreting their data and overcome the various limitations associated with reliance on a single index.

RESULTS

Sample

14 out of 17 complete booklets were returned, representing an 82% response rate. Respondents were predominantly male (9 male, 5 female). The mean age of the sample was 38.92 (SD = 7.05). The mean years' of experience working in the outdoor education and recreation sector was 15.69 (SD =6.12). All but one participant had outdoor recreation or education specific qualifications. Eight participants led activities as part of their current role. Five participants were employed in safety-specific management roles (e.g. Health and Safety Coordinator; Risk Manager); six were employed in management roles (e.g. Head of Department; Business Manager); and two were activity leader trainers. All respondents would be expected to report, investigate or analyze incidents as part of their role.

Number of factors identified at each level of the taxonomy per incident.

As Table 2 shows, there was a wide degree of variation in terms of the number of codes that participants identified for each incident. Two participants were particular outliers in relation to the rest of the group: one tended to consistently select only a few codes as relevant, while another tended to identify many codes across the taxonomy. The other participants tended to be quite similar to each other.

Table 2 also illustrates that compared to the total number of Level 3 codes (325), on average few Level 3 codes were selected as relevant to each incident. This provides a further justification for the data analysis strategy of only examining the reliability of codes that were selected by at least one respondent.

Table 2: Descriptive statistics of the number of factors identified at each level of the taxonomy per incident (M = Mean, SD = Standard deviation), n = 14

	Incident									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Level 1										
M	4.86	4.36	4.42	6.07	4.36	6.62	4.75	6.67	6.09	4
SD	1.16	1.78	2.34	1.21	1.82	1.39	1.60	1.87	1.81	1.84
Range	4-7	2-9	2-10	4-8	2-8	4-10	3-9	3-10	3-9	2-9
Level 2										
M	10.86	9.29	9.5	13.50	9.93	10.76	6.91	12.41	14.55	8.18
SD	5.12	6.14	6.82	5.27	5.94	4.59	3.90	4.46	6.83	5.64
Range	5-22	3-27	2-29	8-24	3-23	6-23	3-18	6-22	6-29	3-23
Level 3										
M	13.93	10.57	12.93	18.00	13.86	16.38	9.25	18.58	24.63	11.91
SD	6.84	9.19	10.38	11.05	9.76	12.33	8.30	6.69	16.80	10.80
Range	5-30	3-38	3-43	8-48	3-39	6-55	4-35	9-45	7-67	3-41

Level 1 inter-rater reliability

The mean r_{wg} for Level 1 codes was .44 with a standard deviation of .17; indicating an on average near substantial level of agreement for Level 1 codes. All codes at this Level were identified as playing a role in at least 7 incident reports. The inter-rater reliability summary data for Level 1 codes across all incidents is presented in Table 3. A number of Level 1 codes reached a substantial level of agreement, including "Environmental", "Instructor", "Regulatory bodies and associations" and "Government". The mean percentage agreement for these codes indicates that the codes relating to "Environmental" and "Instructor" were more likely to be selected as relevant by the majority of respondents, while the codes "Regulatory bodies and associations" and "Government" were more likely to be rejected by the majority. Regarding the use of the later codes, some respondents tended to extrapolate about the role of these entities in the incident, rather than relying directly on the details contained in the incident reports. For example, one respondent identified these codes across all incidents; stating that there no "enforceable standards" in the outdoor activity sector.

There was also a reasonable level of agreement regarding the use of the code "Supervision/management", "Activity centre management" and "Local area government, schools or parents." The first two codes were selected as relevant in all incidents, with an average percentage of agreement of over 80% of the sample. In comparison, "Local area government, schools or parents" was rejected in the majority of incidents; again some respondents tended to extrapolate about the role of these entities in the incidents beyond the information

contained in the incident reports.

There appears to be disagreement over the use of the codes "Equipment", "Participant" and "Group". The disagreement over the use of the codes "Equipment" and "Participant" appears to stem from confusion over how the incorrect use of equipment should be classified (i.e. is it a factor relating to the equipment or the person using it?). Similarly, respondents had trouble differentiating between factors that affected all members of the activity group and therefore should be classified as "Group" in comparison to factors that affected single members of the group and therefore should be classified as "Participant".

Table 3: Summary of results for Level 1 codes inter-rater reliability across all incidents (n = 14).

		r_{wg}		% agreement	
Level 1 code	**# incidents where this factor was selected**	**Mean**	**SD**	**Mean**	**SD**
1. Equipment	9	0.29	0.35	72.93	19.12
2. Environmental	10	0.45	0.34	59.21	35.35
3. Instructor	10	0.77	0.28	88.75	22.96
4. Participant	10	0.32	0.36	54.96	30.55
5. Group	7	0.36	0.28	30.53	26.82
6. Supervision/management	10	0.32	0.37	81.74	33.81
7. Activity centre management	10	0.30	0.32	80.39	71.62
8. Local area government, schools or parents	10	0.39	0.28	31.50	28.80
9. Regulatory bodies and associations	9	0.64	0.21	20.59	30.23
10. Government	9	0.49	0.31	21.64	23.81

Level 2 inter-rater reliability

The mean r_{wg} for Level 2 codes was .37 with a standard deviation of .15; indicating a reasonable level of agreement for Level 1 codes. Not all Level 2 codes were utilized; in particular factors relating "Animals and insects" were not identified as playing a role in any incident.

The inter-rater reliability summary data for Level 2 codes across all incidents is presented in Table 4. A number of codes at this level reached a substantial level of agreement; however, this is largely attributable to agreement that codes can be rejected, rather than selected. Of the codes that had a mean $r_{wg} \geq .45$, there were only two codes that were both selected by more than 60% of the sample and reached a substantial level of agreement ("9.1 Auditing" and "3.9 Physical Condition").

The codes that appear particularly problematic mainly fall under the higher level categories "Instructor" and "Supervision/management". In relation to the higher level category "Instructor", there appeared to be little agreement (selection or rejection) across incidents concerning codes relating to mental conditions and decisions (e.g. "3.3 Decision", "3.5 Experience, qualifications and competence", "3.6 Leadership", "3.8 Perception"). In relation to "Supervision/management" the codes "6.1 Planning & Activity Program", "6.2 Safety management", "6.3 Staff and Staffing", "6.4 Supervision" tended to be used interchangeably to describe similar factors identified from the incident reports, with little distinction between the codes.

Table 4: Summary of results for Level 2 codes inter-rater reliability across all incidents (n = 14)

		r_{wg}		% agreement	
Level 2 code	**# incidents where this factor was selected**	**Mean**	**SD**	**Mean**	**SD**
1.1 Activity equipment	8	0.22	0.26	41.99	25.85
1.2 Clothing and PPE	6	0.30	0.22	32.15	25.31
1.3 Documentation	3	0.51	0.24	13.50	8.30
1.4 Food and drink	3	0.36	0.20	57.14	37.80
1.5 Medication	1	0.18		75.00	
2.1 Temperature	5	0.35	0.20	60.65	33.45
2.2 Weather	6	0.31	0.23	38.74	30.30
2.3 Miscellaneous	1	0.71		7.14	
2.4 Animals and insects	0				
2.5 Physical environment	5	0.43	0.28	25.39	26.66
2.6 Terrain	6	0.44	0.33	22.37	23.24
2.7 Trees and vegetation	2	0.26	0.19	22.62	8.42
2.8 Water	3	0.05	0.07	60.75	15.79
3.10 Planning and preparation	9	0.20	0.32	52.09	26.35
3.11 Safety	9	0.12	0.22	36.97	15.53
3.12 Unsafe acts	7	0.37	0.26	19.20	10.31
3.13 Violations	5	0.55	0.11	11.93	3.27
3.2 Compliance	7	0.33	0.33	28.83	23.60
3.3 Decision	8	0.25	0.29	51.49	28.78
3.4 Demonstration	6	0.35	0.29	23.72	18.44
3.5 Experience, qualifications and competence	10	0.34	0.28	38.47	30.66
3.6 Leadership	10	0.36	0.30	32.05	27.81
3.7 Mental condition	5	0.51	0.21	25.06	30.45
3.8 Perception	10	0.16	0.21	34.78	17.76
3.9 Physical condition	2	0.56	0.62	85.71	20.20
4.10 Unsafe acts	4	0.42	0.35	19.78	16.74
4.1 Communications	7	0.36	0.35	33.97	28.70
4.11 Violations	2	0.59	0.17	10.71	5.05
4.2 Compliance	4	0.41	0.33	18.34	13.20
4.3 Decision	7	0.50	0.27	15.58	12.61
4.4 Demonstration	0				
4.5 Experience and competence	7	0.10	0.11	32.35	9.78
4.6 Mental condition	4	0.60	0.22	10.71	7.14
4.7 Perception	7	0.40	0.29	18.31	10.68
4.8 Physical condition	7	0.29	0.25	29.81	22.54
4.9 Training and Practice	8	0.39	0.28	18.99	11.51
6.1 Planning and activity program	10	0.26	0.31	66.15	23.38
6.2 Safety management	10	0.29	0.37	41.64	28.76
6.3 Staff and Staffing	8	0.15	0.18	47.10	23.06
6.4 Supervision	10	0.11	0.21	33.65	12.36
7.1 Activity and training programs	10	0.30	0.24	37.63	27.65
7.2 Organizational characteristics and constraints	9	0.45	0.26	16.08	10.09
7.3 Practices	9	0.29	0.31	26.82	19.24
7.4 Procedures	9	0.38	0.22	53.26	34.52
7.5 Risk/hazard management systems	10	0.37	0.28	39.12	31.70
8.2 Schools	8	0.50	0.21	29.94	32.95
8.3 Parents	4	0.49	0.26	14.40	9.08
9.1 Auditing	3	0.50	0.33	57.18	44.36
9.2 Regulatory bodies	10	0.50	0.24	14.35	9.34
10.1 Budgetary constraints	5	0.44	0.28	16.71	11.64
10.2 Infrastructure and land	3	0.44	0.37	17.71	15.59
10.3 Policy and legislation	8	0.61	0.22	15.44	21.75

Level 3 inter-rater reliability

Due to the large number of Level 3 codes, the analysis for Level 3 codes focuses on Incident 1 (the pattern of results described below was similar across all incidents).

For Incident 1, the mean r_{wg} for selected codes was .51 with a standard deviation of .24. However, this is largely attributable to agreement over the rejection of codes, rather than the selection. On average, 17.20% of the sample agreed on the selected codes. Only four codes were selected by more than 60% of respondents ("3.9a Fatigue", "3.7c Mental fatigue", "3.2b Failed to follow instructions" and "6.3d Poor rostering"). Only one of these codes reached a substantial level of agreement ("3.9a Fatigue", which had a r_{wg} of 1, reflecting unanimous agreement).

A high level of agreement on only a few codes would potentially be acceptable if overall participants selected few Level 3 codes. However, overall 81 Level 3 codes were selected. On average, respondents selected 14 Level 3 codes with a range of 5 to 28 codes. Given that a substantial level of agreement was reached on so few codes, this indicates poor inter-rater reliability at Level 3.

The large number of codes selected by respondents is primarily due to the selection of multiple codes to describe the same factor. For example, one participant coded "helmet not worn" as "1.1a Equipment not used properly", "1.1b Failure to use equipment", and 1.2a Activity clothing/Personal Protective Equipment not used" and "3.2b failed to follow procedures." Similarly, another participant selected three codes to describe the factor "should have cancelled activity due to being tired" – "3.3c Wrong/inappropriate decision", "3.10a Failure to cancel/abandon activity" and "3.11b Failure to abort activity." Across participants the factor "instructor was tired" tended to coded as both "3.7c Mental fatigue" and "3.9a Fatigue."

DISCUSSION

The aim of this study was to test the inter-rater reliability of a prototype taxonomy for classifying the causal factors involved in incident reports from risk managers from the outdoor education and recreation sector. Inter-rater reliability was examined for each level of the taxonomy (i.e. Level 1: The led outdoor activity system; Level 2: Descriptive categories; and Level 3: Specific causal factors). The findings show that reasonable levels of inter-rater reliability were achieved for Level 1 of the taxonomy relating to codes that were selected as causal to the incident. However, it is evident that there was some confusion concerning the use of "Equipment", "Participant" and "Group." For Level 2, there were reasonable levels of inter-rater reliability in the majority of the codes that were not considered to be causal to the incident. However, for the small subset of codes which the majority of respondents selected, acceptable levels of inter-rater reliability were not achieved. For Level 3 codes the same pattern of results were observed as for Level 2, with even less agreement over the codes that applied to the incident. Obviously, reliably classifying the causes of an incident is equally, if not more important, than reliably rejecting potential causal factors. This pattern of results replicates findings from other domains, where it has been shown that increasing specificity in coding can result in a loss of inter-rater reliability (Finch et al., 2012; O'Conner, 2008).

Neuendorf (2002) identified four threats to inter-rater reliability: 1) a poorly executed coding scheme; 2) inadequate coder training; 3) coder fatigue; and 4) the presence of a rogue coder. In addition, inadequately detailed reports and disparate backgrounds of coders (Finch et al., 2012) may also threaten inter-rater reliability. To various extents, all of the threats identified by Neuendorf (2002) can be said to have played a role in the results of this study. Firstly, it is clear from the results that the taxonomy contains too much detail, with codes that are indistinct to the novice coder. For example, many coders had difficulty distinguishing between "incorrect use of equipment" and "broken equipment". While the countermeasures that would be required to address these issues would differ considerably, it appears unlikely that coders with minimal training can reliably make these fine-grain distinctions when coding incident reports. A longer period of systematic training with feedback may help address this problem.

In addition, it is also evident that there are a number of codes that are not sufficiently distinct, as respondents often selected multiple codes to describe the same factor. In particular, the sub-categories within "Supervision/management" and "Activity Centre management" were often applied to describe the same factor. These codes need to be revised so they are mutually exclusive or combined.

Secondly, an examination of the factors classified in relation to each code suggests that there were some disagreements that could easily be resolved through further training. For example, the code "Equipment" (and sub-categories at level 2 and 3) was intended to be used to classify any causal factor relating to the equipment used during the activity. However, some participants consistently coded "the incorrect use of equipment" as a factor relating to "Participants". This confusion (i.e. is it a factor relating to the equipment or the person using it?) could easily be resolved through further training. Through the results of this study we have identified a number of codes that require disambiguation through further training.

Thirdly, many of the respondents complained about the time required to complete the task. The researchers anticipated that it would take approximately 20 minutes to code each incident; many participants reported that it required over an hour for each incident. This is probably attributable to respondents' lack of familiarity with the taxonomy combined with the number of codes specified at Level 3. Clearly, the number of codes in the taxonomy needs to be reduced to accommodate novice coders.

Finally, one of the main issues identified was the tendency of some respondents to go beyond the details contained in the reports, effectively coding factors based on what their personal experience told them would be involved in the incidents. Respondents were instructed only to code the causal factors that were explicitly stated within the reports. However, one respondent in particular, and two to three other respondents to a less degree, tended to code factors relating to "Government" and "Regulatory bodies and associations" in relation to all incidents, stating that there are no "enforceable standards" in the outdoor activity sector. Similarly, other respondents tended to identify issues with "Activity Centre management" that were not explicitly stated in the incident reports, reasoning that if the instructor had these problems then there must be a problem at higher levels within the organization. While this demonstrates that respondents have adopted a "systems-approach" to understanding incident causation, it was inappropriate in the context of the aims of the present study because the coded information was not present in the incident reports. Potentially, the coding framework needs criteria that can be used to determine whether there is evidence enough to support the attribution of a causal factor.

The study demonstrates the importance of testing coding schemes throughout their development, rather than post development. The next iteration of the coding scheme will be informed by this and other studies, ensuring that the final scheme achieves acceptable levels of reliability and validity. Further testing of the coding scheme will involve test-retest paradigms to examine the reliability of coders coding over time and also validity testing where coders' performance is assessed against an expert standard. This need to test methodological reliability and validity throughout the development process extends to all forms of human factors methodologies (Annett, 2002) and is recommended as a key line of inquiry for human factors and safety science research.

In conclusion, the study demonstrates that in developing a coding taxonomy there is a clear tension between categories that are broad and reliable yet provides insufficient information, and categories that are highly detailed and have low levels of reliability. The intention was to develop categories that could be used by risk managers in the outdoor sector with minimal training, and that also had enough detail that aggregate analyses of incident reports could immediately be used to generate meaningful injury prevention strategies, without further data coding. It appears that each goal cannot be achieved without some cost to the other. While Level 1 of the framework showed reasonable levels of reliability, it is too broad to be of use for developing injury prevention strategies (e.g. 'equipment' being identified as the key causal factor in 1000 incidents does not shed much light on prevention strategies). Similarly, Level 3 of the taxonomy appears to be too large and unwieldy to be of practical use. Potentially, Level 2 of the tested taxonomy may provide a middle ground between these two extremes; however, significant work is required to revise these codes so that they are parsimonious and discrete. As with any human factors method, the coding scheme has to achieve acceptable levels of reliability before it can be used in the real world.

REFERENCES

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

Cessford, G. (2013). National Incident Database 2012 Report. New Zealand: Mountain Safety Council.

Dickson, T. J., & Gray, T. 2012). Risk Management in the Outdoors: A Whole-of-Organisation Approach for Education, Sport and Recreation. Cambridge University Press: Cambridge, GB.

Finch, C.F., Orchard, J.W., Twomey, D.M., Saleem, M.S., Ekegren, C., Lloyd, D. G. & Elliott, B.C. (2012). Coding OSICS sports injury diagnoses in epidemiological studies: does the background of the coder matter? *British Journal of Sports Medicine*.

Fleishman, E. A., & Quaintance, M. K. (1984). Taxonomies of Human Performance: The Description of Human Tasks. Orlando, FL: Academic Press.

Goode, N., Finch, C., Cassell, E., Lenne, M., & Salmon, P. (In Press). What would you like? Identifying the required characteristics of an industry-wide incident reporting and learning system for the led outdoor activity sector. *Australian Journal of Outdoor Education*.

Gordon, R., Flin, R., & Mearns, K. (2005). Designing and evaluating a human factors investigation tool HFIT) for accident analysis. *Safety Science, 43(*3), 147-171.

Neuendorf, K. A. (2002). *The Content Analysis Guidebook.* United States of America: Sage Publications Inc.

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

Salmon, P., 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., Goode, N., Lenne, M., Cassell, E., & Finch, C. (2014). Injury causation in the great outdoors: a systems analysis of led outdoor activity injury incidents. *Safety Science, 63*, 111-120.

Salmon, P., Williamson, A., Lenné, M., Mitsopoulos-Rubens, E., & Rudin-Brown, C. M. (2010). Systems-based accident analysis in the led outdoor activity domain: application and evaluation of a risk management framework. *Ergonomics, 53(*8), 927-939.

Stanton, N. A., & Young, M. S. (1999). What price ergonomics? *Nature, 399*, 197-198.

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