Which code is it? Inter-rater reliability of systems theory-based causal factor taxonomy for the outdoor sector

Figure 1: UPLOADS revised taxonomy in the context of the adapted RRMF

2 Method

2.1 Sample

Human factor experts were invited to participate through the email lists maintained by the research team. Coding booklets were then emailed to these participants with a request to email the results back to the same University of the Sunshine Coast email address. Ethics approval for the study was granted by the University of the Sunshine Coast Human Ethics Committee (A/14/604).

2.2 Content of Coding booklet

The first section of the coding booklet asked participants for basic demographics information (e.g. age, gender, organisation, organisational role, type and length of human factor expertise).

The second section explained the structure of the contributory factor taxonomy. Participants were informed that multiple codes could apply to the same factor. For example, if both a person and equipment were described to be at fault, then both should be coded to their level of the taxonomy. Participants were also asked to only use the taxonomy to code contributory factors that were explicitly mentioned in the incident report, and avoid speculation

The third section of the booklet contained three incident reports. After each report, a list of the contributory factors involved was presented (these were identified through the study reported in Goode, Salmon, Lenne and Finch, 2014). For example, "Hot Day" and "Activity Leader incorrect use of quickdraw". Participants had to select the code from the taxonomy that best described each factor. The incident reports described real-life events without any identifying details that referred to people or locations. The reports were adapted from the Australian Accident Register, an online publicly accessible voluntary report database of accidents and serious near misses from the LED community (https://groups.google.com/forum/#!forum/australian-accident-register). Reports were selected because they described a range of contributory factors in enough detail to be coded to Level 2 of the taxonomy.

2.3 Data analysis

To assess the inter-rater reliability the occurrence or absence of the Level 1 and Level 2 codes was recorded for each participant for each contributory factor involved in each incident.

The inter-rater reliability of each code was calculated using: 1) the within-group inter-rater reliability coefficient (r_{wg}); and 2) percent agreement. James, Demaree, and Wolf (1993) define r_{wg} as the comparative reduction in error variance of a distribution of 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: ; 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: $\sigma EU2 = (A^2 - 1)/12$, where A is the number of possible alternatives in the rating scale. For this study, there was only 2 two responses for each code '1' and '0'. R_{wg} ranges from 0 to 1, with 0 representing complete variance of scoring with no agreement among participants. A score of 1 represents agreement and no variance.

There is no standard method for interpreting r_{wg}. O'Connor (2008) suggested that $r_{wg} \geq .6$ was sufficient agreement amongst participants. However, in this study $r_{wg} \geq .6$ would only be achieved if 12 out of 13 participants selected the same code. For this study an $r_{wg} \geq .44$ was chosen as indicating a substantial level of agreement, as it represents an 85% agreement. This is above the 60-70% agreement that is generally acknowledged to acceptable in the majority of other reliability studies.

The use of r_{wg} has been criticised because it does not differentiate between agreement that a code should be rejected, and agreement that it should be selected. That is, r_{wg} will be ≥.44 when 85% of participants select a code, or when 85% reject it. This creates a problem when there is a large number of codes, because r_{wg} will be artificially inflated by correct rejections . This is the case in the UPLOADS taxonomy, where the Level 2 has 107 codes. To avoid this, the current study used two methods. First, only the codes that were selected by at least one participant had an r_{wg} calculated. Individual values were then averaged across the three incidents. Secondly, percent agreement was calculated. Unlike r_{wg}, percent agreement has the ability to differentiate between rejection and acceptance, as a value of 85% reflects that 85% of participants agree upon the same code and 15% do not.

3. Results

Level 2 Code	# Incidents	r_{wg} Mean	r_{wg} SD	% Mean	% SD
1.0 Activity Equipment and Resources	1	0.23	0.00	23.08	0.00
1.1 Documentation	3	0.15	0.26	30.77	15.38

3.1 Sample

13 out of 16 complete booklets were returned, representing an 81% response rate. Eight participants were male and 5 were female, with a mean age of 30.61 (SD= 11.05). The mean level of experience with Human Factors methods was 10.66 years (SD= 15.06).

3.2 Level 1 inter-rater reliability

On average, Level 1 codes had a mean r_{wg} .43 (SD = .30) across the three incidents, indicating a close to substantial level of inter-rater reliability. All codes at this level were identified as contributory factors in at least one incident.

The degree of inter-rater reliability varied across Level 1 codes, as shown in Table 1. Five codes reached an $r_{wg} \geq .44$. Three codes reached an 85% level of agreement: "Environment", "Leader" and "Participant". One code reached an 85% level of agreement that it was not involved: "Parents/Carers". One code, "Group Factors", had one incident where it was substantially agreed to be involved, and two incidents where it was substantially agreed to be not involved.

The r_{wg} for the remaining ten codes ranged from .20 to .30.

Level 1 code	No. Incident Selected	r_{wg} Mean	r_{wg} SD	% Mean	% SD
1. Equipment	3	0.20	0.43	61.54	27.74
2. Environment	1	1.00	0.00	100.00	0.00
3. Leader	3	1.00	0.00	100.00	0.00
4. Participant	3	0.81	0.33	94.87	8.88
5. Other People in Activity Group	3	0.30	0.12	20.51	4.44
6. Group Factors	3	0.52	0.15	35.90	42.37
7.Other People in Activity Environment	2	0.33	0.15	19.23	5.44
8. Supervisor/Field Manager	3	0.20	0.26	69.23	20.35
9. Higher Level Management	3	0.09	0.13	69.23	7.69
10. Local Area Government	1	0.06	0.00	38.46	0.00
11. Schools	2	0.38	0.44	38.46	43.51
12. Parents/Carers	1	0.44	0.00	15.38	0.00
13. Regulatory Bodies/Professional Association	2	0.46	0.33	15.38	10.88
14. State and Federal Government	1	0.23	0.00	23.08	27.74

Table 1: Summary of inter-rater reliability for Level 1 codes across all incidents (n = 13)

3.3 Level 2 inter-rater reliability

The mean r_{wg} for Level 2 codes was .47, with a standard deviation of .25, indicating substantial agreement.

The inter-rater reliability summary data for Level 2 codes across all incidents is presented in Table 2. Eight codes had an $r_{wg} \geq .44$. Two codes reached an 85% level of agreement: '2.5 Weather Conditions' and '3.4 Judgement and Decision Making'. Three codes reached an 85% level of agreement that they were not involved: '9.11 Other' '11.1 Communication', and '11.4 Legal Responsibility for safety of staff and students'. Three codes, '3.2 Compliance With Procedures, Violations & Unsafe Acts', '3.5 Mental And Physical Condition' and '9.7 Staffing and recruitment,' had mixed agreement for different incidents.

The r_{wg} for the remaining sixty codes ranged from zero to .38

Table 2: Summary of inter-rater reliability for Level 2 codes across all incidents (n = 13)

1.2 Equipment, Clothing & PPE	3	0.30	0.36	41.03	34.69

1.4 Medication (For Those Involved In The Activity)	1	0.38	0.44	19.23	16.32
1.5 Activity Equipment and Resources: Other	1	0.69	0.00	7.69	0.00
2.5 Weather Conditions	1	1.00	0.00	100	0.00
3.0 Activity Leader	1	0.69	0.00	7.69	0.00
3.1 Communication, Instruction & Demonstration	2	0.26	0.25	42.31	38.07
3.2 Compliance Procedures/Violations & Unsafe Acts	2	0.62	0.54	61.54	54.39
3.3 Experience, Qualifications, Competence	2	0.03	0.07	53.85	21.76
3.4 Judgement And Decision-Making	3	0.44	0.00	84.62	0.00
3.5 Mental And Physical Condition	2	0.69	0.00	50.00	59.83
3.6 Planning & Preparation	2	0.10	0.18	57.69	27.20
3.7 Situation Awareness	3	0.28	0.27	23.08	13.32
3.8 Supervision/Leadership Of Activity	2	0.69	0.00	7.69	0.00
4.0 Activity Participant	2	0.69	0.00	7.69	0.00
4.1 Communication & Following Instructions	3	0.20	0.43	66.67	23.50
4.2 Compliance Procedures/Violations/ Unsafe Acts	3	0.61	0.15	10.26	4.44
4.3 Experience, Competence	2	0.21	0.33	73.08	16.32
4.4 Judgment And Decision-Making	2	0.33	0.15	19.23	5.44
4.5 Mental And Physical Condition	2	0.21	0.33	73.08	16.32
4.6 Planning & Preparation For Activity/Trip	2	0.03	0.07	34.62	5.44
4.7 Situation Awareness	2	0.69	0.00	7.69	0.00
4.8 Activity Participant: Other	1	0.69	0.00	7.69	0.00
5.1 Communication & Following Instructions	2	0.69	0.00	7.69	0.00
5.2 Compliance Procedures/Violations & Unsafe Acts	1	0.69	0.00	7.69	0.00
5.3 Experience, Qualifications, Competence	1	0.69	0.00	7.69	0.00
5.4 Judgement And Decision-Making	2	0.69	0.00	7.69	0.00
5.6 Planning & Preparation For Activity/Trip	1	0.69	0.00	7.69	0.00
5.9 Situation Awareness	2	0.69	0.00	7.69	0.00
5.9 Other People in Activity Group: Other	1	0.69	0.00	7.69	0.00
6.1 Communication Within Group	2	0.69	0.00	7.69	0.00
6.4 Group Size	1	0.23	0.00	76.92	0.00
6.6 Teamwork	1	0.67	0.00	7.69	0.00
6.7 Time Pressure	2	0.68	0.02	7.69	0.00
6.8 Activity Group Factors: Other	1	0.69	0.00	7.69	0.00
7.1 Communication	2	0.69	0.00	7.69	0.00
7.4 Judgement And Decision-Making	2	0.41	0.03	15.38	0.00
7.6 Planning & Preparation	2	0.68	0.02	7.69	0.00
8.1 Activity Or Program Design	2	0.18	0.36	34.62	27.20
8.2 Communication	2	0.21	0.33	38.46	32.64
8.3 Compliance Procedures/Violations/ Unsafe Acts	1	0.69	0.00	7.69	0.00
8.4 Experience, Qualifications, Competence	1	0.69	0.00	7.69	0.00
8.5 Judgement And Decision-Making	1	0.00	0.00	53.85	0.00
8.7 Planning & Preparation For Activity	1	0.69	0.00	7.69	0.00
8.8 Supervision Of Activity Leaders And Other Staff	2	0.44	0.00	15.38	0.00
8.9 Supervision/Oversight Of Programs/Activities	1	0.00	0.00	38.46	0.00
9.1 Communication	2	0.33	0.15	19.23	5.44
9.3 Judgement And Decision-Making	1	0.69	0.00	7.69	0.00
9.5 Policies Procedures Activities /Emergencies	2	0.00	0.00	53.85	0.00
9.6 Risk Assessment And Management	2	0.69	0.00	7.69	0.00
9.7 Staffing And Recruitment	2	0.46	0.33	42.31	48.95
9.9 Supervision/Oversight Of Activities And Programs	1	0.44	0.00	15.38	0.00
9.10 Training And Evaluation Of Staff	1	0.69	0.00	7.69	0.00
9.11 Higher- Level Management: Other	1	0.44	0.00	15.38	0.00
10.2 Communication	1	0.08	0.00	30.77	0.00
10.5 Policies And Procedures	1	0.69	0.00	7.69	0.00
11.1 Communication	2	0.38	0.44	38.46	43.51
11.3 Judgement And Decision-Making	1	0.00	0.00	46.15	0.00
11.4 Legal Responsibility Safety Of Staff/ Students	1	0.44	0.00	15.38	0.00
11.5 Planning And Preparation For Activity/Trip	1	0.69	0.00	7.69	0.00
11.8 Schools: Other	1	0.69	0.00	7.69	0.00
12.3 Judgment And Decision-Making	1	0.69	0.00	7.69	0.00
12.5 Planning And Preparation For Activity/Trip	1	0.69	0.00	7.69	0.00
13.3 Communication	1	0.23	0.00	23.08	0.00
13.7 Standards and code of practice	1	0.69	0.00	7.69	0.00
14.1 Communication	1	0.44	0.00	15.38	0.00
14.4 Policies and legislation	1	0.69	0.00	7.69	0.00

4. Discussion

The aim of this study was to test the inter-rater reliability of the revised taxonomy of led outdoor activity incident contributory factors.

Overall, the inter-rater reliability of the revised taxonomy is greatly improved compared to the original (Goode et al., 2014). The original taxonomy did not reach a substantial level of inter-rater reliability for any of its three levels. However, despite these encouraging findings, only three codes out of fourteen in Level 1 and two codes out of sixty-eight in Level 2 reached 85% agreement. These results were due to confusion in Level 1 and 2 concerning the use of 'Equipment' 'Other People in Activity Group', 'Other People in Activity Environment', 'Supervisor/Field Management', 'Higher Level Management; and 'Local Area Government'.

The secondary aim of this study was to identify any issues prior to implementation in the LOA sector. 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, Finch et al. (2012) identified inadequately detailed reports and coders with diverse backgrounds as threats. The issues in the present study seem to be due to inadequate coder training. Specifically, participants had trouble categorising how specific entities fit into the taxonomy, and only coding the contributory factors that were listed.

In the current study, two different threats to inter-rater reliability were identified that signalled that participants had trouble categorising factors. First, three sets of Level 1 codes were commonly coded to the same contributory factor. The first set was "Higher level management" and "Supervisor/field manager". For example, the contributory factor "Field Manager failed to act upon Activity Leader's concerns about physical exhaustion" was coded to the Level 2 code "Supervision/Oversight of activities and programs" under both Level 1 codes by different participants. The second set was "Local Area Government", "State and Federal Government" and "Regulatory Bodies and Associations". For example, "National Parks and Wildlife Service did not explicitly communicate limits to activity providers" was coded to Level 2 codes within all three Level 1 codes by different participants. The third set was "Other People in Activity Group", "Schools", "Higher Level Management" and "Other People in Activity Environment". For example 'the school did not inform the activity provider of the activity participants' asthma condition" was coded to Level 2 codes within "Other People in Activity Group", 'Schools" and "Higher Level Management."

The second threat to inter-rater reliability identified was that multiple or different Level 2 codes within "Leader", "Participant", and "Activity Equipment and Resources" were used interchangeably for the same factors. For example, "The participant did not disclose his pre-existing injury to the activity providers" was coded to two codes within 'Activity Participant: '4.1 Communication, instruction & demonstration and 'Activity Participant: 4.6 Planning & Preparation' by different participants. The examples given for each code indicate that the latter would be most appropriate. However, this indicates that more information is required to train coders to distinguish between these two codes.

Previous research has found that 'expert' samples have trouble making fine grain distinctions in taxonomies. A study by O'Connor (2008), found that an 'expert' sample got confused between the levels of the Human Factors Analysis and Classification System (HFACS) taxonomy. The HFACS taxonomy was designed to be used by both experts and personnel in the aviation field to identify underlying human contributory factors in aviation accident investigations (O'Connor, 2008). Both the participants in this study and in the HFACS reliability study had trouble distinguishing between codes that represented a step higher or lower in the system that the taxonomy was testing, especially if they had similar names (O'Connor, 2008). Similar to this study, mistakes were also made when participants were required to place specific actors into the largely similar categories. This problem was also found in the initial UPLOADS reliability study of the taxonomy (Goode et al. 2014).

In the initial reliability study, Goode et al (2014) found that LOA risk managers went beyond the information provided in the incident report, and referred to their personal experience when coding the reports. The results from this study show that participants are still making assumptions. For example, "Parents/Carers" was selected to describe the contributory factor "Participant had run out of Ventolin" even though 'Parents/Carers' was not mentioned. This speculation is highly problematic, because risk managers are likely to have distinct experiences that may influence their selection of codes.

To increase the inter-rater reliability of the revised taxonomy for implementation into the LOA sector, two different sets of information should be provided to participants. First, further explanation is required

regarding the roles of the actors and artefacts within the LOA system. Secondly, further explanation of the second level is needed, to assist coders in distinguishing between concepts such as "judgement and decision making" and "situation awareness". Third, coders should be provided with examples of common assumptions (e.g., that participant had run out of Ventolin because the parents forgot) alongside an alternative explanation of that situation (e.g. participant had run out of Ventolin because it was stolen). This is an extension of Goode et al. (2014)'s suggestion of a criterion that determines whether there is enough evidence in the descriptions to code at that factor.

In conclusion, to improve inter-rater reliability participants need further training in identifying and classifying contributory factors, and how these factors are described by the taxonomy. In addition, they also need further instruction to discourage speculation. The next step before integrating the coding scheme into the UPLOADS incident reporting system is to test the revised taxonomy in a sample of LOA risk managers, to check for reliability issues in the second demographic of end-users. In addition, it is important to test whether reliability remains stable over time, i.e test-retest reliability. This ensures results are not a product of chance for one particular time point. Lastly, the validity of the taxonomy should be tested, i.e. whether participants are choosing the correct codes. This also ensures that the right risk interventions are formulated from the analysis. The results from this study show the importance of testing the reliability of a method before releasing the product to the end user. It also shows that a reliable method is an essential criteria to save money and reduce causalities.

Acknowledgements

This project was funded through the Australian Research Council Linkage scheme (Project LP110100037) in partnership with Australian Camp Association, Outdoor Educators' Association of South Australia, United Church Camping, Outdoors Victoria, Outdoor Council of Australia, Recreation South Australia, Outdoor Recreation Industry Council, Outdoors WA, YMCA Victoria, The Outdoor Education Group, Girl Guides Australia, Queensland Outdoor Recreation Federation, Christian Venues Association, Parks Victoria, Victoria Department of Planning and Community Development, Outdoor Education Australia and the Department of National Parks, Recreation, Sport and Racing Australia. Caroline Finch was funded by a National Health and Medical Research Council (NHMRC) Principal Research Fellowship (ID: 1058737). ACRISP is one of the International Research Centres for the Prevention of Injury and Protection of Athlete Health supported by the International Olympic Committee (IOC). Paul Salmon's contribution was funded through his Australian Research Council Future Fellowship (FT140100681).

References

Dekker, S. (2011). Drift into failure: From hunting broken components to understanding complex systems. U.K., Ashgate.

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*, 48(7).

James, L., R., Demaree, R, G., & Wolf, G. (1993). Rwg: An assessment of Within-Group Interrater Agreement. *Journal of Applied Psychology, 78*(2), 306-309.

Goode, N., Salmon, P, M., Lenne, M, G., & Finch, C, F. (2014). A test of a systems theory-based incident coding taxonomy for risk managers. *Proceedings of the 5th International Confrence on Applied Human Factors and Ergonomics*.

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

O'Connor, P. (2008). HFACS with an additional layer of granularity: validity and utility in accident analysis. *Aviation Space and Environmental Medicine, 76*(6), 599-606.

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

Reason, J. (1997). Managing the risks of organizational accidents. Aldershot, UK, Ashgate.

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.

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