Cite As: 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.

EVALUATION OF A HUMAN FACTORS ANALYSIS AND CLASSIFICATION SYSTEM AS USED BY SIMULATED MISHAP BOARDS

Paul O'Connor, PhD, MSc and Peter Walker, PhD, BSc

Running title: Evaluation of a human factors classification system

Manuscript metrics:

Word count for Abstract: 245

Word count for narrative text: 2,420

Number of references: 15

Number of Tables: 3

Number of Figures: 0

ABSTRACT

Background: The reliability of the Department of Defense Human Factors Analysis and Classification System (DOD-HFACS) has been examined when used by individuals working alone to classify the causes of summary, or partial, information about a mishap. However, following an actual mishap a team of investigators would work together to gather, and analyze, a large amount of information before identifying the causal factors and coding them with DOD-HFACS. ***Method***: 204 military Aviation Safety Officer students were divided into 30 groups. Each group was provided with evidence collected from one of two military aviation mishaps. DOD-HFACS was used to classify the mishap causal factors. ***Results:*** Averaged across the two mishaps, acceptable levels of reliability were only achieved for 56.9% of nanocodes. There were high levels of agreement regarding the factors that did not contribute to the incident (a mean agreement of 50% or greater between groups for 91.0% of unselected nanocodes), the level of agreement on the factors that did cause the incident as classified using DOD-HFACS were low (a mean agreement of 50% or greater between the groups for 14.6% of selected nanocodes). ***Discussion:*** Despite using teams to carry out the classification, the findings from this study are consistent with other studies of DOD-HFACS reliability with individuals. It is suggested that in addition to simplifying DOD-HFACS itself, consideration should be given to involving a human factors/organizational psychologist in mishap investigations to ensure the human factors issues are identified, and classified, in a consistent and reliable manner.

Key words: DOD-HFACS, reliability, human factors, mishap classification

INTRODUCTION

The collection of reliable accident data is essential for improving workplace safety. However, this is not an easy goal to achieve. Many mishap reporting systems are not built upon a firm theoretical framework (4), and do not provide a complete picture of the conditions under which the mishap occurred (13). Moreover, given that the majority of mishaps in high reliability industries can be attributed to human error (12), there is a need for organizations to be able to accurately capture the human causes of mishaps.

Recognizing the difficulties in accurately, and reliably classifying the human factors causes of aviation mishaps, Wiegmann and Shappell (14) developed the Human Factors Analysis and Classification System (HFACS). HFACS is based upon Reason's (13) organizational model of human error. The HFACS framework as a whole has been shown to have substantial levels of reliability between pairs of raters (15; as indicated by a Cohen's kappa of 0.71 between pairs of raters using the system; see the analysis section for a discussion of interpreting kappa coefficients).

The U.S Department of Defense (DOD) added an additional level of classification to HFACS that allows for the specific identification, and classification, of each mishap causal factor. For each HFACS category between 1 and 16 associated nanocodes were developed (there are a total of 147). This adaption to HFACS was called DOD-HFACS (for more details see reference 2).

Studies of the reliability of mishap classification systems, DOD-HFACS included, have examined the reliability of these systems by looking at the reliability between individuals who use the system alone to classify the causal factors from summary, or partial, information about mishaps (5, 8, 9). These studies all concluded that there were problems with the reliability of DOD-HFACS, and that that more

parsimony, increased mutual exclusivity, and training were required to utilize DOD-HFACS effectively. The Australian Defence Force (ADF) also developed a version of HFACS with an additional level of nanocodes called HFACS-ADF. The authors of a paper evaluating the use of HFACS-ADF as applied by typical end-users determined that the system was unreliable. It was concluded that "extensions of HFACS which include an additional layer of 'descriptors' have met with little success" (10, p.444). Although the findings from these studies provide important insights into the unreliability of HFACS with an additional layer of nanocodes, they are an inaccurate simulation of how a mishap classification system would be used by an organization following an actual mishap.

When organizations have a major accident, it is investigated by a team (rather than an individual working alone) of people who work together to sift through a large amount of information (rather than summary, or partial information) over a number of weeks (rather than hours). The investigation team is free to take time to discuss the possible causes of the mishap, and work together to identify, and classify, the causal factors using a mishap classification system. Given that the only studies in the literature examine reliability of mishap classification systems as used by individuals, the reliability of mishap classification systems when used by teams is unknown.

The purpose of this paper is to extend the earlier studies by O'Connor (8) and O'Connor et al (10) by examining the reliability of DOD-HFACS as applied in a manner that simulates how it would be used by a U.S Navy Aviation Mishap Board (AMB) following an actual U.S Naval aviation mishap. Further, comparisons will be made with the individual level of DOD-HFACS reliability reported in references 5, 8, and 9.

METHOD

Subjects

All 204 of the subjects were Aviation Safety Officer (ASO) students at the Navy/Marine Corps School of Aviation Safety, Pensacola, Florida. A total of 197 were naval aviators, and seven were aeromedical specialists. The ASO course is 23 days of instruction in safety programs, human factors, aerospace medicine, mishap investigation, mishap reporting, aerodynamics, and structures. As part of the 25 hours of human factors and aerospace medicine training received by the students, two hours were specifically devoted to hands-on training in the use of DOD-HFACS to investigate a mishap. The participants were separated into 30 groups of between six and eight students (mean= 6.8, st dev=0.61). The study was judged to be exempt from review by the Institutional Review Board Chairperson of the Naval Postgraduate School.

Procedure

Within the first two days of ASO school the students were separated into groups. This was not done randomly, but rather the aim was to ensure that each group consisted of individuals with a range of aviation experience. Each group was provided with all of the evidence from either an investigation of a helicopter mishap, or a tactical aviation (TACAIR) mishap. The details of the mishap cannot be presented here as they are based upon actual U.S Navy mishaps. However, the TACAIR mishap had fairly clear organizational failures in formal communication, whereas the helicopter mishap was largely centered on failures in crew resource management. The information that was provided to the groups included all of the information that would be collected as part of an actual mishap (e.g. engineering reports, medical reports on

those involved in the mishap, interview transcripts, copies of pertinent procedures, copies of formal communications).

The 'investigation' continued throughout ASO school and culminated with each group writing a Safety Investigation Report (SIR) that was then graded on the basis of completeness and accuracy. The SIR is the standard report that is written by a U.S Navy AMB following a class A aviation mishap (one in which the total cost of damage to property, aircraft, or exceeded $2,000,000; 1). "The purpose of an SIR is to report hazards that were causes of the mishap or were causes of damage or injury occurring in the course of the given mishap and to provide a means for submitting recommendations to eliminate those hazards" (7, p.102). A key part of the SIR is the classification of the causes of the mishap using DOD-HFACS.

Data analysis

The causes of the mishap, as identified by the DOD-HFACS coding of the mishap causal factors included in each group's SIR, were analyzed. The same method used by O'Connor et al (9) was used to calculate the reliability and agreement between the nanocodes selected, and rejected, by each group. Reliability between the groups was calculated using the multi-rater kappa free (κ_{free}). The multirater κ_{free} uses the same observed probability as Fleiss' (3) kappa, but the expected probability is 1/k (where k is equal to the number of categories; see 11 for more details). Multirater κ_{free} is appropriate for situations in which the rater does not know a priori the quantity of cases that should be distributed into each category. Multirater κ_{free} can take values of -1 to 1. A value of zero is indicative of agreement at chance, greater than zero better than chance, and less than zero worse than chance. The inter-group reliability of the nanocodes associated with each category was calculated.

Statistical significance is not generally regarded as a useful method for interpreting kappa as relatively low values of kappa can still be significant. Although not without criticism, a widely used interpretation was provided by Landis and Koch (6). They proposed that a kappa of less than 0 was indicative of poor agreement, between 0.0 and 0.20 indicates a slight agreement; between 0.21 and 0.40 a fair agreement; between 0.41 and 0.60 a moderate agreement; between 0.61 and 0.80 a substantial agreement; and between 0.81 and 1.00 almost perfect agreement.

The percentage agreement between the groups was independently examined for those nanocodes that the groups believed to be causal to the incident, and among the nanocodes that the groups did not think contributed to the incident. The reliability and agreement at the nanocodes level was evaluated by examining the selected and unselected nanocodes within each of the 20 DOD-HFACS categories. The reliability and agreement at the category level of DOD-HFACS was assessed by examining the selected and unselected categories within each of the four levels of DOD-HFACS. A category was deemed to be causal if at least one of the nanocodes in that category was selected as causal to the incident.

RESULTS

A total of 13 groups analyzed the helicopter mishap, and 17 groups analyzed the TACAIR mishap. Tables I and II summarize the reliability and agreement amongst the groups for each mishap.

INSERT TABLES I AND II

Table III provides a comparison of the percentage of categories for which the reliability at the nanocode was substantial or higher (a kappa of greater than 0.6) compared to the inter-rater reliability between individual raters using DOD-HFACS reported by Hughes et al (5), O'Connor (8), and O'Connor et al (9).

INSERT TABLE III

DISCUSSION

It can be seen that the levels of inter-group reliability of the nanocodes associated with each category are comparable to the reliability of individuals using DOD-HFACS to identify mishap causal factors (5, 8, 9). The overall reliability of DOD-HFACS found in this study showed that there was substantial agreement between the groups (as indicated by a mean overall kappa of 0.65). However, examining the reliability of the nanocodes at the category level reveals that the acceptable overall reliability can be attributed to the high reliability in rejecting nanocodes that clearly did not apply to the mishaps.

As found by O'Connor (8), and O'Connor et al (9), although there were high levels of agreement among the nanocodes that were judged not to be causal to the mishap, there were much lower levels of agreement surrounding nanocodes that were selected as being causal to the mishap. It is possible that the advantage of having a group of people to consider the causes of the mishap was attenuated by the greater variability introduced by the necessity to consider a large amount of evidence rather than just reviewing the summary of a mishap.

As described by O'Connor et al (9), rejecting potential causes is an important early step in mishap investigation. However, the AMB must then go on to reliably use

DOD-HFACS to identify the nanocodes that contributed to the mishap. Therefore, this paper adds further support to the conclusions of O'Connor (8), and O'Connor et al (9) that a more parsimonious version of DOD-HFACS needs to be developed with greater mutual exclusivity between nanocodes.

There are numerous examples of overlapping nanocodes within a particular category (e.g. 'overconfidence' and 'complacency', 'checklist error' and 'procedural error'). There are also overlapping nanocodes from different categories. For example, the 'perceptual factors' nanocodes of expectancy (defined as when the individual's expects to perceive a certain reality and those expectations are strong enough to create a false perception of the expectation; 2) and the 'cognitive factor' nanocode of confusion (a state characterized by bewilderment, lack of clear thinking, or perceptual disorientation; 2). The effect of overlapping nanocodes (particularly across categories) is that the summary data presented to senior leadership is unreliable, and may result in limited resources being applied to address the wrong issues.

There are three weaknesses to the study reported in this paper. Firstly, we did not compare individual and groups using DOD-HFACS to identify the causal factors of the same mishaps. The reason for this is that given the amount of information that would have to be considered, the task would be overwhelming for a person working on their own, and a suitably qualified person could not be released from their job for the time necessary to carry out the task. Although the comparison between individuals and groups may be interesting in terms of comparing group and individual decision making, this evaluation has little practical value given that major mishap investigation are carried out by a group of people, rather than an individual operating alone.

Secondly, the members of the simulated AMBs were not from the same backgrounds as would make up an actual U. S Navy AMB. An actual AMB would

have, at a minimum: an aviation safety officer, a flight surgeon, an officer knowledgeable about aircraft maintenance, an officer knowledgeable about aircraft operations, and a senior member who is in-charge of the board (1). Given that all of the subjects in the simulated AMB had received training in human factors, and in the use DOD-HFACS, it could be argued that the reliability of the simulated AMBs in using DOD-HFACS was an overestimation as to what would be expected in a real AMB in which only two members of the (the safety officer and flight surgeon) have had any training in human factors or DOD-HFACS training.

The largest weakness of this study is that the reliability of DOD-HFACS was only examined with respect to two specific mishaps. Complex mishaps without clear causes will have a detrimental effect on the measured reliability of a mishap classification system. To illustrate, for the helicopter mishap, a key decision that had to be made by the groups investigating this mishap was whether failures at the supervisory level could also be attributed to failures on the part organization as a whole. As reflected by the low reliability at the organizational level, about half of the groups believed these failures could be attributed to the organization, while the other groups did not. There is no clear right or wrong answer. As possible causes of a mishap further back in the organization are examined it can become harder, and arguably more subjective, to link them directly to the mishap. However, perhaps by including an experienced human factors/organizational psychologist in a mishap investigation team more consistent may be achieved.

It is suggested that where human causes are suspected, a human factors/organizational psychologist should be utilized by AMBs. We believe that the background knowledge in human factors that is likely to be required in the investigation of a major mishap is beyond the small amount of human factors and

DOD-HFACS training received by the Aviation Safety Officer and flight surgeon. By involving an experienced human factors/organizational psychologist, with a background in aviation safety, we believe this will help ensure the pertinent human factors issues are identified, and classified, in a consistent and reliable manner.

CONCLUSION

The accurate classification of mishaps is necessary for tracking safety performance, and the effectiveness of safety measures. Although perhaps an improvement on the reliability of how the U.S. Navy (and other high-risk organizations) classifies the human factors causes of mishaps, it can be concluded that measures must to be taken to improve the reliability of DOD-HFACS. Further, in agreement with Olsen and Shorrock (10), we are doubtful as to whether typical end-users can use HFACS with an additional layer of nanocodes to reliably classify the causes of mishaps.

ACKNOWLEDGEMENTS

All opinions stated in this paper are those of the authors and do not necessarily represent the opinion or position of the U.S. Navy, the Navy/Marine Corps School of Aviation Safety, or the National University of Ireland, Galway.

REFERENCES

1. Chief of Naval Operations. Naval aviation safety program. OPNAVINST 3750.6R change transmittal 4. 2009; Retrieved 7 May 2010: from http://doni.daps.dla.mil/Directives/03000%20Naval%20Operations%20and%20Readiness/03700%20Flight%20and%20Air%20Space%20Support%20Services/3750.6R.pdf
2. DOD HFACS: A mishap investigation and data analysis tool. 2005; Retrieved on 9 September 2010 from www.uscg.mil/safety/docs/ergo_hfacs/hfacs.pdf
3. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971; 76: 378-382.
4. Gordon R, Flin R, Mearns K. Designing and evaluating a human factors investigation tool (HFIT) for accident analysis. Safety Science 2005; 43: 147-171.
5. Hughes TG, Heupel KA, Musselman BT, Hendrickson E. Preliminary investigation of the interrater reliability of the Department of Defense Human Factors Accident and Classification System in USAF mishaps [Abstract]. Aviat Space Environ Med 2007; 78: 255.
6. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33(1): 159-174.
7. Naval Safety Center. The Naval flight surgeon's pocket reference to aircraft mishap investigation. Norfolk, VA: Naval Safety Center, 2001: 102-116.
8. O'Connor P. HFACS with an additional level of granularity: validity and utility in accident analysis. Aviat Space Environ Med 2008; 79: 599-606.
9. O'Connor P, Walliser J , Philips E. Evaluation of a human factors analysis and classification system as used by trained raters HFACS with an additional level of

granularity: validity and utility in accident analysis. Aviat Space Environ Med 2010; 81:956-960.

10. Olsen NS, Shorrock ST. Evaluation of the HFACS-ADF safety classification system: Inter-coder consensus and intra-coder consistency. Acc Anal Prevent 2010; 42, 437-444.
11. Randolph JJ. Free-marginal multirater kappa: An alternative to Fleiss' fixed-marginal multirater kappa.2005. Retrieved on 25 February 2010 from http://www.eric.ed.gov/PDFS/ED490661.pdf
12. Reason J. Human error: models and management. Br Med J 2000; 320: 768-770.
13. Stoop J. Accident scenarios as a tool for safety enhancement strategies in transportation systems. In: Hale A, B. Wilpert B, Freitag M (Eds.). After the event: from accident to organisational learning. Oxford: Elsevier Science Ltd, 1997: 77–93
14. Wiegmann DA, Shappell SA. A human error approach to aviation accident analysis: the human factors analysis and classification system. Aldershot, UK: Ashgate; 2003: 122-147.
15. Wiegmann DA, Shappell SA. Human error analysis of commercial aviation accidents: application of the human factors analysis and classification system. Aviat Space Environ Med 2000; 72(11): 1006-1016.

Table I. Inter-group reliability and mean agreement for the helicopter mishap.

	κ_{free}	Mean % unselected	Mean % selected	Mean unselected		Mean selected	
				100% agreement (%)	≥ 50% agreement (%)	100% agreement (%)	≥ 50% agreement (%)
Level 1: Acts (5.2)*	0.53	23.1	76.9	0	25.0	0	75
Skill-based errors (1.2)	0.60	80.8	19.2	50.0	83.3	0	33.3
Judgment & decision errors (2.2)	0.21	64.1	35.9	0	66.7	0	33.3
Misperception errors (0.3)	0.08	69.2	30.8	0	100	0	0
Violations (1.6)	-0.08	46.2	53.8	0	0	0	100
Level 2: Preconditions (9.8)	0.61	59.0	41.0	28.6	85.7	28.6	42.9
Physical environment (0.2)	0.93	97.9	2.1	90.9	100	0	0
Technological environment (0)	1.00	100	0	100	100	0	0
Cognitive factors (1.4)	0.53	82.7	17.3	25.0	87.5	0	16.7
Psycho-behavioral factors (2.7)	0.69	82.1	17.9	53.3	86.7	0	28.6
Adverse physiological state (0)	1.00	100	0.0	100	100	0	0
Physical/mental limitations (0.2)	0.89	96.9	3.1	80.0	100	0	0
Perceptual factors (0.7)	0.79	93.7	6.3	63.6	100	0	0
Coordination/communication/ planning factor (4.6)	0.26	62.2	37.8	8.3	58.3	0	36.4
Self imposed stress (0.1)	0.95	98.7	1.3	83.3	100	0	0
Level 3: Supervision (3.6)	0.24	28.8	71.2	0	0.0	25.0	100
Inadequate supervision (1.6)	0.32	73.1	26.9	16.7	83.3	0	20.0
Planned inappropriate actions(0.6)	0.85	91.2	8.8	85.7	85.7	0	100
Failed to correct a known problem 0.7)	0.03	65.4	34.6	0	100	0	0
Supervisory violations (0.7)	0.51	82.7	17.3	50.0	100	0	0
Level 4: Organizational influence (1.7)	0.08	53.8	46.2	0	66.7	0	33.3
Resources/acquisition management 0.2)	0.91	97.4	2.6	88.9	100	0	0
Organizational climate (0.5)	0.68	89.2	10.8	60.0	100	0	0
Organizational processes (1.0)	0.55	83.3	16.7	33.3	83.3	0	25.0

* Numbers in parentheses represent the mean number of nanocodes selected by each group in this category or at this level.

Table II. Inter-group reliability and mean agreement for the TACAIR mishap.

	κ_{free}	Mean % unselected	Mean % selected	Mean unselected		Mean selected	
				100% agreement (%)	≥ 50% agreement (%)	100% agreement (%)	≥ 50% agreement (%)
Level 1: Acts (2.4)*	0.67	60.3	39.7	50.0	50.0	0	100
Skill-based errors (0.7)	0.85	88.2	11.8	83.3	83.3	0	100
Judgment & decision errors (1.7)	0.35	71.6	28.4	16.7	83.3	0	20.0
Misperception errors (0)	1.00	100.0	0	100	100	0	0.0
Violations (0)	1.00	100.0	0	100	100	0	0.0
Level 2: Preconditions (6.1)	0.48	70.6	29.4	12.5	87.5	12.5	25
Physical environment (0.1)	0.96	98.9	1.1	90.9	100	0	0
Technological environment (0.1)	0.94	98.5	1.5	87.5	100	0	0
Cognitive factors (0.6)	0.75	92.6	7.4	50.0	100	0	0
Psycho-behavioral factors (1.0)	0.81	93.3	6.7	66.7	100	0	0
Adverse physiological state (0.4)	0.90	97.4	2.6	62.5	100	33.3	0
Physical/mental limitations (0.2)	0.88	96.5	3.5	80.0	100	0	0
Perceptual factors (0)	1.00	100.0	0	100	100	0	0
Coordination/communication/ planning factor (3.6)	0.39	70.1	29.9	0	83.3	0	16.7
Self imposed stress (0.1)	0.36	63.7	36.3	66.7	100.0	0	0
Level 3: Supervision (4.0)	0.25	39.7	60.3	0	25.0	0	75.0
Inadequate supervision (2.2)	0.63	86.6	13.4	16.7	83.3	0	20.0
Planned inappropriate actions (0.9)	0.25	70.6	29.4	57.1	100	0	0
Failed to correct a known problem (0.6)	0.74	92.6	7.4	0.0	100	0	0
Supervisory violations (0.3)	0.92	98.0	2.0	50.0	100	0	0
Level 4: Organizational influence (3.9)	0.55	27.5	72.5	0.0	33.3	0	66.7
Resources/acquisition management (1.7)	0.52	81.0	19.0	33.3	88.9	0	16.7
Organizational climate (0.4)	0.69	91.8	8.2	20.0	100	0	0
Organizational processes (1.8)	0.27	70.6	29.4	0	83.3	0	16.7

* Numbers in parentheses represent the mean number of nanocodes selected by each group in this category or at this level.

Table III. Comparison of the mean percentage of inter-group nanocode reliability of a substantial agreement or higher (a kappa greater than 0.6) and the mean kappas in the current study and those of Hughes et al (6), O'Connor (9), and O'Connor et al (10).

	Hughes et al (6)	O'Connor (9)	O'Connor et al (10)	Current study
# mishap scenarios	48	2	1	2
Acts	Unreported	25.0	N/A*	37.5
Preconditions	Unreported	83.3	100	77.8
Supervision	Unreported	25.0	75.0	50.0
Organizational influence	Unreported	66.7	66.7	50.0
Total %	55.6	57.5	87.5	53.8
Mean Kappa	Unreported	0.67	0.75	0.65

*Data was analyzed by the authors on the basis of complete agreement at the Act level.