Evaluation of a human factors analysis and classification system as used by trained raters.

Title	Evaluation of a human factors analysis and classification system as used by trained raters.
Author(s)	O'Connor, Paul
Publication Date	2010-10

Cite as: O'Connor, P., Walliser, J., & Philips, E. (2010). Evaluation of a human factors analysis and classification system as used by trained raters. *Aviation, Space and Environmental Medicine, 81*:956-960.

Evaluation of a Human Factors Analysis and Classification System Used by Trained Raters

Paul O'Connor MSc, PhD, James Walliser, BSc, & Eric Philips BSc, MSc

Running title: Evaluation of a Human Factors Classification system

Manuscript metrics:

Word count for Abstract: 242

Word count for narrative text: 2,397

Number of references: 16

Number of Tables: 2

Number of Figures: 0

Abstract

Background: The U.S. Department of Defense (DoD) has utilized DoD Human Factors Analysis and Classification System (DoD-HFACS) to help identify and classify human factors that may have caused or contributed to aircraft mishaps since 2005. ***Method***: In this study 22 military officers used DOD-HFACS to classify information obtained from an interview with an individual who had been involved in an aviation incident in which the potential for serious injury had been high. ***Results:*** It was found that although the overall inter-rater reliability was generally acceptable (as reflected by a mean Fleiss' kappa of 0.76), and there were high levels of agreement regarding the factors that did not contribute to the incident (there was agreement of 50% or greater between raters for 84.4% of unselected nanocodes), the level of agreement on the factors that did cause the incident as classified using DOD-HFACS were lower than desirable (agreement of 50% or greater between raters that a particular nanocode was causal was found only for a mean of 22.5% of selected nanocodes). ***Discussion:*** The findings from this study are consistent with the small number of other studies reporting an evaluation of the reliability of DOD-HFACS. It is recommended that organizations must evaluate the reliability and validity of mishap coding systems, as applied by the proposed end-users, prior to the widespread adoption of a system. It is only through the accurate identification of mishap causal factors that informed decisions can be made to prevent future mishaps.

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

INTRODUCTION

The need for high-risk organizations to collect reliable and valid mishap data is crucial to improving workplace safety. Further, given that human error is a factor in 80 to 90 percent of all work-related of mishaps (12.13), mishap classification systems must be able to accurately and reliably capture the human factors causes of mishaps. The mishap analysis should allow an organization to draw the right conclusions and prevent similar mishaps from occurring in the future (part of what Reason (13) describes as a learning culture).

In an effort to reliably classify the human factors causes of mishaps, U.S. Naval aerospace medicine has utilized the Human Factors Analysis and Classification System (HFACS; 3). HFACS has a hierarchical structure based upon Reason's (12) organizational model of human error. A full discussion regarding the theory and structure of the HFACS taxonomy is beyond the scope of this paper. The reader is therefore referred to (3) for more detail. Pairs of well-trained experts have demonstrated acceptable levels of reliability using HFACS to classify the causes of aviation mishaps (6, 15). However, the lack of in-depth detail (granularity) of HFACS led to criticism of the system's ability to detect specific operational problems and suggest interventions (1).

The lack of granularity of HFACS was addressed by the U.S. Department of Defense (DoD; Navy, Marine Corps, Army, Air Force, Cost Guard, and Department of Homeland Security) Aviation Safety Improvement Task Force. The Task Force added an additional level of classification to HFACS that allowed for more detailed analysis. The purpose of the nanocodes is to allow the specific identification, and classification, of each specific mishap causal factor. For each HFACS category between 1 and 16 associated

nanocodes were developed (there are a total of 147). To illustrate, the six nanocodes associated with the category of 'skill based errors' are: 'inadvertent operation', 'checklist error', 'procedural error', 'overcontrol/undercontrol', 'breakdown in visual scan', and 'inadequate anti-g straining maneuver' (see 3 for more details and definitions of the nanocodes). This adaption to HFACS was called DOD-HFACS, and it was agreed in 2005 by the U.S. DoD to use DOD-HFACS to classify all mishaps (8).

Two studies of DOD-HFACS reliability have been reported in the literature. Hughes et al (5) examined the inter-rater reliability between four professional safety investigators and safety policy consultants. DOD-HFACS was used to classify 54 U.S. Air Force mishaps. It was found that only 52% of the nanocodes had a reliability greater than the 0.60 kappa coefficient recommended by Wiegmann and Shappell (16).

O'Connor (8) examined the inter-rater reliability and accuracy of 123 naval aviators who used DOD-HFACS to identify the human factors causes of two aviation mishap scenarios. It was found that the inter-rater reliability was acceptable for the majority of the nanocodes that were not considered to be causal to the mishap. However, for the small number of nanocodes of which at least half of the subjects thought applied to the mishap, acceptable levels of inter-rater reliability were not achieved. O'Connor (8) recommended that there was a need for more parsimony of the DOD-HFACS, increased mutual exclusivity of nanocodes, and training was required to use the system effectively.

The purpose of the current study was to utilize DOD-HFACS in a manner that is closer to how it would be applied by an investigator of a real mishap, than by the two studies described above, to assess the inter-rater reliability and accuracy of the system as used by trained users with a background in human factors. Some efforts were made to

increase the parsimony of the version of DOD-HFACS used in the current study. Although the version of DOD-HFACS used had the same nanocodes as the original DOD-HFACS, the definitions had been simplified by mishap analysts from the U.S. Naval Safety Center with the goal of making them clear to someone who was not a human factors expert (see 7 for the nanocode definitions).

METHOD

Subjects

A total of 22 military officers participated in the study. All of the subjects were junior military officers enrolled in the Human Systems Integration Master's program at the Naval Postgraduate School, Monterey, California. Prior relevant education completed as part of the Master's was: 12 week courses in 'human factors in design', and 'individual performance', and four weeks of a course in 'team performance.' The study was judged to be exempt from review by the Institutional Review Board (IRB) by the Vice-Chairman of the Naval Postgraduate School IRB.

Procedure

As a graded assignment within the course entitled 'survivability, habitability, environmental safety, and occupational health', the students were required to use DOD-HFACS to identify the human factors causes of an aviation incident. The assignment was given during the fourth week of the course. Four hours of classroom time was specifically devoted to hands-on training in the use of DOD-HFACS to investigate a mishap, and one

hour of instruction in the use of the Critical Incident Technique (CIT) interview (the training was carried out in a tutorial style).

The subjects carried out a CIT interview with a U.S. Navy officer who had been involved in a rotary wing flying incident where the potential for death or serious injury had been high. The CIT interview is a task analysis method used for evaluating systems and behavior in work environments. CIT interviews have been shown to be an effective method for eliciting information about mishaps, or near-misses, in operational environments (9, 10). The reader is referred to references 9, and 10 for detailed accounts of the CIT interview technique. The CIT interviews were carried out by groups of three or four students. Each group created a single transcript of the interview.

The mishap causal factors identified in the transcript were classified individually using the nanocodes from the revised DOD-HFACS. The nanocodes were presented in the form of a flip book (see 7) that was given to each subject. The subjects were provided with the following instructions for using DOD-HFACS, "use your group's interview transcript to carry out an analysis using DOD-HFACS. Work on your own. Identify a minimum of four [unsafe] ACTS. Then for each [unsafe] ACT identify the PRECONDITIONS, SUPERVISORY, and ORGANIZATIONAL failures for each individual [unsafe] ACT. Write down the nanocode chosen, and provide a three or four sentence justification for your choice of each nanocode."

Data analysis

The reliability between the raters was calculated using the multi-rater kappa free (κ_{free}). The multirater κ_{free} uses the same observed probability as Fleiss' (4) 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 percentage agreement between the raters was independently examined for those nanocodes that the respondents believed to be causal to the incident, and among the nanocodes that the respondents did not think contributed to the incident.

RESULTS

A total of 18 of the 22 responses were analyzed. To be included in the analysis the student must have followed the instructions, and received a grade of A minus or above. Of the 72 usable sets of causal factors (a set being defined as the precondition, supervisory, and organizational influence nanocodes that were selected in association with a particular act level nanocode), there were 12 act level nanocodes for which at least two raters agreed were causal to the mishap (see Table I). The mean number of nanocodes selected for each set of causal factors was 1.46 at the precondition level, 1.09 at the supervisory level, and 0.63 at the organizational influence level.

Table I. Frequency with which respondents agreed that the act level nanocodes were causal to the incident.

Category	Nanocode	Freq. identified
Skill based errors	Unintended operation of equipment	2
	Checklist not followed correctly	1
	Procedure not followed correctly	5
	Over control/under control	2
	Breakdown in visual scan	1
	Inadequate anti-g straining maneuver	0
Judgment & decision making errors	Inadequate real-time risk assessment	10
	Failure to prioritize tasks	2
	Rushed a necessary action	4
	Delayed a necessary action	2
	Ignored a caution/warning	3
	Wrong choice of action during an operations	3
Perception errors	Incorrect response to a misperception	0
Violations	Work-around violation	3
	Widespread/routine violation	18
	Extreme violation	16

Table II summarizes the reliability and agreement among the raters for the nanocodes and categories at each DOD-HFACS level (a category was deemed to be causal if at least one of the nanocodes in that category was selected as causal to the incident).

Table II. Mean rater reliability and agreement amongst the ratings for DOD-HFACS levels 2, 3, and 4.

	Mean κ_{free}	Mean % unselected	Mean % selected	Mean unselected		Mean selected	
				≥ 50% agreement (%)	100% agreement (%)	≥ 50% agreement (%)	100% agreement (%)
Level 2: Preconditions	**0.78**	**81.0**	**19.0**	**78.7**	**68.0**	**63.4**	**22.0**
Physical environment (11)*	1.00	99.5	0.5	100	98.5	7.7	0
Technological environment (8)	0.98	99.5	0.5	100	89.6	0	0
Cognitive factors (6)	1.00	100	0	100	98.6	0	0
Psycho-behavioral factors (12)	0.63	86.4	13.6	84.6	55.1	30.9	1.5
Adverse physiological state (8)	0.94	98.8	1.2	100	94.8	15.4	0.0
Physical/mental limitations (5)	1.00	100	0.0	100	100	0	0.0
Perceptual factors (11)	0.98	99.6	0.4	100	99.2	8.3	0.0
Coordination/communication/ planning factor (15)	0.72	81.3	18.7	90.5	68.0	12.9	3.2
Self imposed stress (16)	1.00	100	0	100	100	0	0
Level 3: Supervision	**-0.01**	**51.3**	**48.7**	**39.8**	**8.0**	**63.0**	**10.9**
Inadequate supervision (6)	0.61	85.5	14.5	77.0	50.2	34.3	5.7
Planned inappropriate actions (2)	-0.71	59.8	40.2	44.0	12.9	52.4	9.5
Failed to correct a known problem (7)	0.81	94.6	5.4	100	73.8	18.2	0
Supervisory violations (4)	0.72	90.9	9.1	71.1	54.1	23.5	0
Level 4: Organizational influence	**0.19**	**70.0**	**30.0**	**43.1**	**19.3**	**50.0**	**3.8**
Resources/acquisition management (9)	0.98	99.6	0.4	96.7	94.0	0	0
Organizational climate (5)	0.58	83.0	17.0	78.5	40.7	28.1	6.3
Organizational processes (6)	0.72	93.4	6.6	100	72.2	19.2	0

*values in parentheses represents the number of nanocodes in each category.

The mean κ_{free} column in Table II was calculated by taking the mean κ_{free} between the raters for each category for the set of selected, and unselected, nanocodes associated with each of the 12 act level nanocodes that at least two raters agreed were causal to the mishap. This averaging method was also used to calculate the mean values for all of the other columns in Table II. The mean percentage of unselected, and selected, nanocodes are in columns three and four of Table II respectively. Columns five and six of Table II show the mean number of nanocodes for which there was 50% or greater agreement and 100% agreement amongst raters for the unselected nanocodes respectively. The final two columns in Table II shows the mean number of selected nanocodes for which there was greater than 50% or greater agreement and 100% agreement respectively.

DISCUSSION

From the summary data presented in Table II it can be seen that acceptable levels of reliability were found for the majority of the categories. The authors believe that this finding is largely due to the fact that the majority of nanocodes are not selected. The high levels of agreement among the unselected nanocodes serves to mask the much lower levels of agreement surrounding nanocodes that were selected as being causal to the mishap (see Table II for the agreement levels for selected nanocodes). Similar conclusions were drawn by O'Connor (8). However, it should be noted that the subjects in the current study had received more training in DOD-HFACS and had received more

education in human factors and human performance limitations than the subjects in the O'Connor (8) study.

It could be argued that the fact that the subjects were basing their analysis on different CIT interviews introduces variability into the DOD-HFACS coding of the mishap. This is certainly possible. However, the interviewee carefully read the transcripts of each of the interviews, and found them to be very similar, with no major discrepancies.

The high levels of agreement between the majority of unselected nanocodes is unsurprising. To illustrate, if the physical environment was clearly not causal to a mishap, there is high levels of reliability and agreement between raters. Rejecting potential causes is an important early step in mishap investigating. "The first thing the Aviation Mishap Board must do is discuss everything that could possibly have led to the mishap, then reject those things too remote to consider" (p.6-20; 2). However, it is crucial for mishap investigators to then go on to reliably identify the causes of the mishap.

From Table II it can be seen that, similar to the findings of O'Connor (8), the levels of agreement between raters for selected nanocodes is much lower than would be desirable. This was the case despite the fact the raters were in complete agreement at the ACT level. The ability to reliably classify the cause of a mishap is fully as important as reliably rejecting potential mishap causal factors (8). Further, examining the reliability at the category level also suggests that coding the nanocodes seemed to have a detrimental effect on the reliability at the category level for the supervisory and organizational influence levels in this study. Li and Harris (6) make the point that mishap investigators may have difficulty identifying abstract concepts such as 'operational tempo/workload' and linking this back to the cause of the mishap. However, in studies utilizing HFACS, in

which there are no nanocodes, pairs of well trained raters were able to reach acceptable levels of agreement (6, 15).

To address the unreliability of selected nanocodes, we propose that each of the 147 nanocodes that are in DOD-HFACS should be scrutinized with respect to two questions. Firstly, is it possible for a user with moderate amounts of training to reliably make the distinction between a particular nanocode and other similar nanocodes? Secondly, does the distinction between similar nanocodes matter from a safety improvement perspective? By scrutinizing each nanocode in this manner we believe that the number of nanocodes can be greatly reduced through discarding or combining similar nanocodes. We believe that simplifying DOD-HFACS though the reduction of the number of nanocodes and increasing the mutual exclusivity of the remaining nanocodes will have a beneficial effect on the reliability of chosen nanocodes.

We believe that the reliability of DOD-HFACS at the category level could be further improved through changing the construction of the supporting documentation and associated training. The DoD HFACS flip book (7) should be separated into two distinct sections. The first section should only consist of the 20 DOD-HFACS categories and definitions. The nanocodes should be listed in the second part of the flip book. The training must emphasize that the appropriate category should be selected first, prior to identifying the particular causal factors at the nanocode level. These changes will help create a distinction between the two levels, and hopefully prevent users from focusing on the nanocode level.

The findings from this study are limited in that only a single incident was coded utilizing DoD-HFACS. Also, the respondents identified the causal factors on their own.

This does not reflect how a mishap is investigated in the 'real world' in which a mishap board would decide as a group on the causes of a mishap. Given that there is a large literature on the effects of teams on decision making, the reliability and validity of mishap classification systems as used by groups versus individuals is something that should be examined. Nevertheless, despite these limitations, the finding that trained raters were unable to use DOD-HFACS to reliably identify the causes of mishaps is consistent with the lack of reliability between users reported by Hughes et al (5) and O'Connor (8).

CONCLUSION

Problems with the reliability of mishap coding systems are not confined to DOD-HFACS, but have been recognized as an issue common to many mishap classification systems (14). It is recommended that organizations should evaluate the reliability and validity of mishap coding systems, as applied by the proposed end-users, prior to the widespread adoption of a system. Moreover, if changes are made to the coding system, the reliability and validity must be re-examined. It is only through the accurate identification of mishap causal factors that informed decisions can be made to prevent future 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, or the Naval Postgraduate School.

REFERENCES

1. Beaubien JM, Baker DP. A review of selected aviation human factors taxonomies, accident/incident reporting systems and data collection tools. Int J of Appl Aviat Stud 2002; 2: 11-36.
2. Chief of Naval Operations. Naval aviation safety program. OPNAVINST 3750.6R. 2001; Retrieved 7 May 2010: from http://doni.daps.dla.mil/Directives/ 03000 %20 Naval%20Operations%20and%20Readiness/03-700%20 Flight %20and%20Air %20Space%20Support%20Services/3750.6R.pdf
3. DoD HFACS: A mishap investigation and data analysis tool. 2005; Retrieved on 16 November 2009 from http://www.safetycenter.navy.mil/hfacs/downloadhfacs.pdf.
4. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971; 76: 378-382
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. Li W-C, Harris D. Pilot error and its relationship with higher organizational levels: HFACS analysis of 523 accidents. Aviat Space Environ Med 2006; 77: 1056-1061.
7. Naval Safety Center. Human factors analysis flip book. Retrieved on 4 May 2010 from http://www.safetycenter.navy.mil/aviation/aeromedical/downloads/ human_factor_analysis_flip-book.pdf
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, O'Dea A, Flin R. Identifying the team skills required by nuclear operations personnel. Int J Ind Ergon 2008; 38: 1028-1037.
10. O'Connor P, O'Dea A, Melton J. A methodology for identifying human error in U.S. Navy diving accidents. Hum Fac 2007; 49: 214-226.
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. Reason J. Achieving a safe culture: theory and practice. Work Stress 1998; 12: 293-306.
14. 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.
15. Wiegmann DA, Shappell SA. A human error approach to accident analysis. Aldershot, UK: Ashgate; 2003.
16. Wiegmann DA, Shappell SA. Human error analysis of commercial aviation accidents: application of the Human Factors Analysis and Classification System (HFACS). Aviat Space Environ Med 2001; 72: 1006 -17.