

**Cite as: O'Connor, P. (2008). HFACS with an additional level of granularity: validity and utility in accident analysis. *Aviation, Space and Environmental Medicine*, 79, 599-606.**

**A Preliminary Evaluation of the Effectiveness of the U.S. Department of Defense Human Factors Analysis and Classification System.**

**Paul O'Connor**

**Short title:** Evaluating DOD-HFACS

**Manuscript metrics:**

Word count for Abstract: 212

Word count for narrative text: 3,953

Number of references: 34

Number of Tables: 4

Number of Figures: 1

## **ABSTRACT**

This paper represents the first attempt to evaluate the effectiveness of the Department of Defense Human Factors Analysis and Classification System (DOD-HFACS). DOD-HFACS delineates the human factors causes of aviation mishaps, and is based upon Wiegmann and Shappell's (32) HFACS coding system. However, the main difference from HFACS is that DOD-HFACS includes an additional level of fine-grain classification. This layer consists of 147 detailed nanocodes that are to be used to identify the mishap causes.

The internal validity, external validity, and utilitarian criteria of DOD-HFACS were evaluated. A total of 123 naval aviators used DOD-HFACS to identify the human factors causes of two aviation mishap scenarios. There were reasonable levels of inter-rater reliability in the majority of the nanocodes that were not considered to be causal to the mishaps. However, for those nanocodes to which at least half of the raters thought applied to the mishaps, acceptable levels of inter-rater reliability were not achieved. Similarly, when compared to 'expert' reference ratings, there were acceptable levels of agreement for those nanocodes that were discarded, but not for the majority of the nanocodes that were identified by the experts as being causal to the mishap. Therefore, this study has identified that more parsimony, increased mutual exclusivity, and training are required to utilize DOD-HFACS effectively.

## **KEYWORDS**

Accident investigation, human error, Department of Defense Human Factors Analysis and Classification System.

## INTRODUCTION

Safety research has shown that human error, as opposed to mechanical failure, is a major cause of industrial and transportation accidents (24,33). Between 80% and 90% of all work related accidents and incidents can be attributed to human error (12,14,24). Similarly, human error accounts for more than 80% of U.S. Naval aviation mishaps (32). Therefore, the collection and accurate analysis of human factors accident data is essential for improving workplace safety (4,18,34).

Reason (25) identifies four critical elements of an effective safety culture — a reporting, just, flexible, and learning culture. However, underreporting, incomplete recordings, and incomplete information about conditions and contexts are common to many accident reporting systems and do not provide a complete picture of the conditions under which accidents result (9, 28). For learning to occur, organizations must collect reliable and accurate human factors data so that measures can be instituted to prevent similar mishaps from occurring in the future.

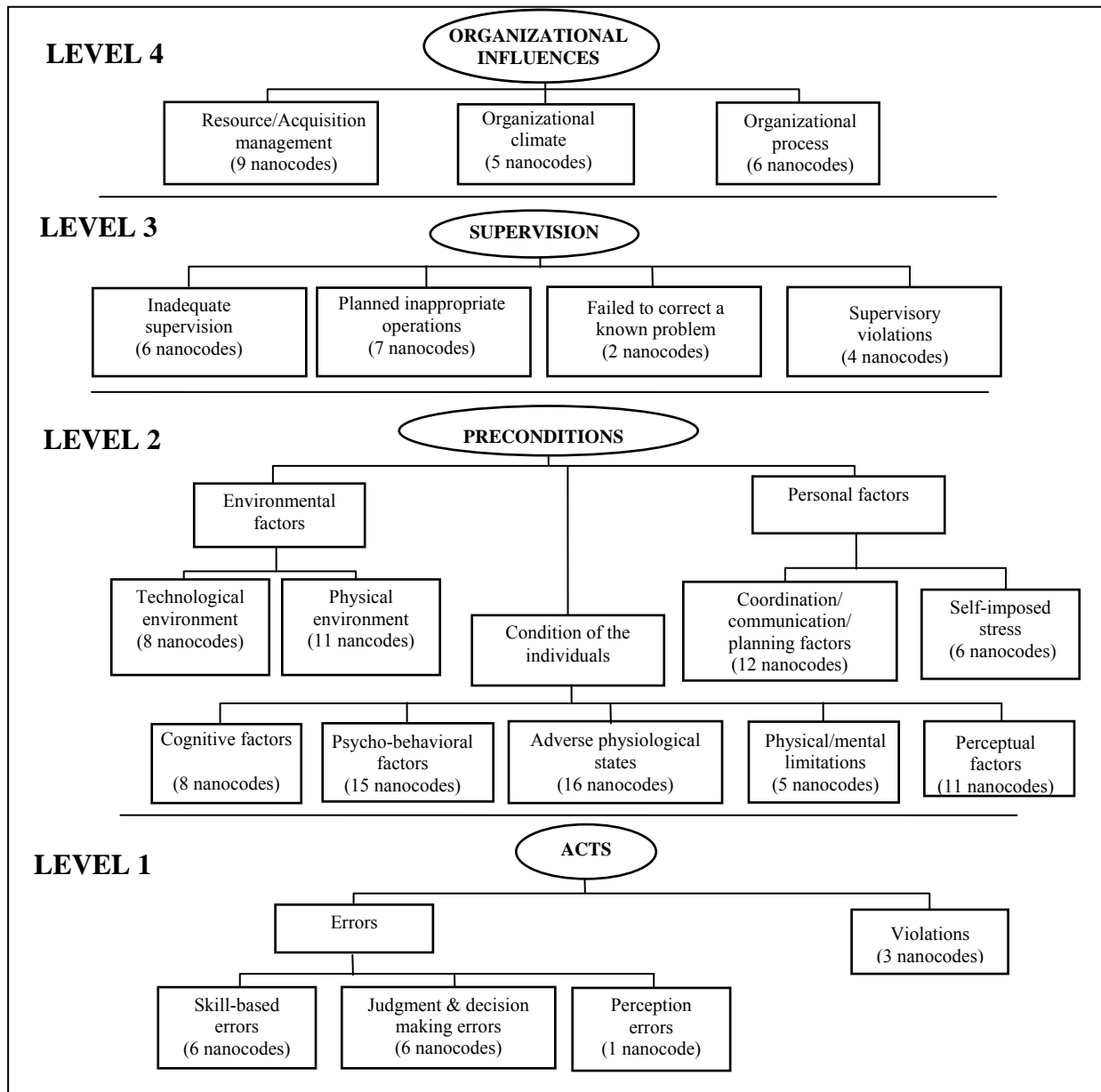
Navy aerospace medicine has been using the Human Factors Analysis and Classification System (HFACS) to analyze the human factors causes of aviation mishaps for a number of years (32). HFACS is derived from Reason's (24) organizational model of human error. HFACS has a clear hierarchical structure, and has been shown to have reasonable levels of reliability for aviation mishap classification when the responses of pairs of well-trained expert have been compared (10,21,26,27,29,31,32,33). The HFACS framework has been applied to the analysis of military (21,29,32) commercial (26,33), and general aviation (10,27,31) mishaps. However, HFACS has received some criticism in terms of the coding system being too coarse, or under-identified, in terms of detecting specific operational problems or to suggest interventions for those problems (2).

The issue of the lack of granularity of HFACS was addressed by the U.S. Department of Defense (DOD; Navy, Marine Corps, Army, Air Force, Coastguard, and Department of Homeland Security) Aviation Safety Improvement Task Force. The Task Force developed DOD-HFACS to meet the goal of creating a common DOD system for investigating the human factors causes of mishaps (5). In 2005 all members of the U.S. DOD signed a memorandum of agreement to use DOD-HFACS to investigate the human factors causes of aviation, ground, weapons, afloat, space, and off-duty mishaps (17).

The structure of DOD-HFACS is founded upon Wiegmann and Shappell's (32) HFACS coding system. However, as can be seen from Figure 1, at the category level there are a number of differences between HFACS and DOD-HFACS. These differences are:

- 'routine violations' and 'exceptional violations' have been dropped as a categories of 'violations' in DOD-HFACS;
- 'adverse mental state' has been dropped as a category of 'conditions of the individual' in DOD-HFACS;
- 'cognitive factors', 'psycho-behavioral factors', and 'perceptual factors' have been added as 'conditions of the individual' in DOD-HFACS;
- 'crew resource management' and 'personal readiness' have been dropped as categories in 'personnel factors' in DOD-HFACS; and
- 'coordination/communication/ planning factors' and 'self-imposed stress' have been added as categories in 'personnel factors' in DOD-HFACS.

Figure 1. Department of Defense Human Factors Accident Classification System (DOD-HFCAS).



In addition to the changes to the categories, the main difference between HFACS and DOD-HFACS is the inclusion of an additional level of fine-grain classification. Each DOD-HFACS category 4 has between one and 16 associated nanocodes (see Figure 1; there are a total of 147 nanocodes in DOD-HFACS). 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 5 for more details and definitions of the nanocodes).

The purpose of the study described in this paper was to carry out a preliminary assessment of the effectiveness of DOD-HFACS to classify the human factors causes of aviation mishaps. There is no standard method for evaluating the effectiveness of a classification system. To evaluate DOD-HFACS, the three major criteria identified by Fleishman and colleagues (6,7) were used: internal validity, external validity, and utilitarian criteria. These criteria were employed by Beaubien and Baker (2) to evaluate the utility of eight different aviation human factors taxonomies.

Internal validity refers to the extent to which a classification system is logically organized and parsimonious (7). Taxonomies that have high internal validity can be reliably used by novices with minimal training. To assess internal validity, the inter-rater reliability of DOD-HFACS, as used by student Aviation Safety Officers (ASO), to classify the human factors causes of a mishap was evaluated. The ASO is a naval aviator whose primary roles are to advise the Commanding Officer on all aviation safety matters and assist in coordinating the investigation of an aviation mishap. The ASO is trained for this role at the Navy/Marine Corps School of Aviation Safety. The ASO course is 23 days of instruction in safety programs, human factors, aerospace medicine, mishap investigation, mishap reporting, aerodynamics, and structures (3).

External validity is the extent to which a taxonomy achieves the objectives for which it was designed (7). The purpose of DOD-HFACS is to provide a common human error categorization system, and is intended for use by all persons who investigate, report and analyze DOD mishaps to “accurately capture and recreate the complex layers of human error in context with the individual, environment, team, and mishap or event” (p.3; 5). An ASO must be able to use DOD-HFACS with a high degree of accuracy to identify the human factors causes of aviation

mishaps. The accuracy was assessed by measuring the extent to which the participants' ratings matched those of an 'expert' reference rating of the human factors causes of the mishap.

Utilitarian criteria is the extent to which a taxonomy is useful and efficient. For DOD-HFACS to be useful and efficient in allowing mishap trends to be tracked over time, it is crucial that the system can be used with acceptable levels of reliability and accuracy.

## **METHODS**

### *Participants*

The participants were 123 U.S. Navy and Marine Corps aviators who were ASO students at the Navy/Marine Corps School of Aviation Safety. A total of 58 were fixed wing aviators, and 65 rotary wing aviators. Of the participants, 97.5% were male, and 2.5% were female with a mean of 7.4 years (st dev= 4.0 years) as a naval aviator and 1,586 flight hours (st dev= 913 hours).

### *Procedure*

As a graded assignment as part of the ASO school curriculum, the students were required to use DOD-HFACS to identify the human factors causes of either a fixed or rotary wing mishap scenario. The assignment was given during the third week of the four week ASO school. 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. When the assignment were returned to the students, they signed a written consent if they agreed to their assignment being used for research.

### *Scenarios*

Two mishaps (one fixed and one rotary wing) were selected from summaries obtained from the U.S. Naval Safety Center of all the U.S. Navy mishaps that had occurred between 1990 and 2000. These particular mishaps were selected for the study because they had a range of human factors causes, sufficient detail, and were easy to understand. Real mishaps were used to ensure that the scenario was realistic. Unfortunately, as these are accounts of actual U.S. Navy mishaps, it is not possible to describe the scenarios in this paper.

### *'Expert' reference ratings*

Reference ratings using DOD-HFACS to delineate the human factors causes of the mishaps were established by two human factors 'experts'. Both of the experts had Ph.D.'s in psychology, were knowledgeable about DOD-HFACS, and experienced in investigating the human factors causes of accidents in a range of high-risk industries. The final conclusion as to the most appropriate nanocodes used to classify the mishaps was reached by consensus between the experts.

### *Analysis*

To assess the inter-rater reliability, the presence (coded 1) or absence (coded 0) of each nanocode was recorded for each rater. To avoid over-representation, the use of each nanocode was counted a maximum of once per rater. The within-group inter-rater reliability coefficient ( $r_{wg}$ ) was used to analyze the inter-rater agreement at the nanocode level. James, Demaree and Wolf (15,16) define  $r_{wg}$  as the proportional reduction in error variance of a distribution of obtained

ratings compared to a distribution representing a random response pattern. Thus, the equation for  $r_{wg}$  is:

$$r_{wg} = 1 - (S_x^2 / \sigma_{EU}^2)$$

where  $S_x^2$  equals the variance of the observed ratings, and  $\sigma_{EU}^2$  equals the population variance of a discrete rectangular distribution of ratings. The equation for this is:

$$\sigma_{EU}^2 = (A^2 - 1) / 12$$

where A is number of possible alternatives in the rating scale. Values of  $r_{wg}$  can vary from 0 to 1. Similar to Cohen's Kappa, when the variance of the obtained ratings is random, then  $r_{wg} = 0$ , reflecting no agreement among raters. However when there is total agreement between the raters, then  $r_{wg} = 1$ . An example of the use of this measure was to assess inter-rater reliability of multiple raters using behavioral marker systems to evaluate the crew resource management skills of aviators (11,20,23).

There are no established criteria for interpreting the value of  $r_{wg}$ . Therefore, for interpretive purposes,  $r_{wg} \geq 0.6$  was taken to be an indication of substantial agreement between the raters. This means that the variance of the rating distributions is at least 60% smaller than the variance associated with a random response pattern, and is consistent to the Landis and Koch (19) widely used guidance for interpreting Kappa.

To assess agreement between the raters and the expert ratings, agreement (coded 1) or disagreement (coded 0) with the expert ratings for each nanocode was recorded. As with the inter-rater reliability, to avoid over-representation, the use of each nanocode was counted a maximum of once per rater. To evaluate the level of agreement between the experts and the raters, the chi-square test was used to assess whether the raters were agreeing with the experts raters

significantly more than would be expected by chance. Given that a chi-square test would be used to evaluate each individual nanocode, to reduce the likelihood of making a Type I error, the significance level was set at the 1% level.

## **RESULTS**

A total of 58 responses were obtained for the fixed-wing mishap, and 65 responses were obtained for the rotary wing mishap. The inter-rater reliability, and the agreement with the reference ratings are discussed below.

### *Inter-rater reliability*

The inter-rater reliability summary data for the fixed wing mishap are shown in Table I, and for the rotary wing mishap are shown in Table II.

Table I. Inter-rater agreement results for the fixed wing scenario (n= 58 raters).

Level		Mean % agreement	St dev % agreement	Mean reliability	St dev reliability	% $r_{wg} \geq .6$	# majority selected	# majority rejected
1	Skill-based errors	84.5	15.2	0.54	0.42	50.0	2 (0)*	4 (3)
	Judgment & decision errors	73.6	16.6	0.31	0.30	16.7	2 (0)	4 (1)
	Misperception errors	74.1	-	0.22	-	0.0	0 (0)	1 (0)
	Violations	81.0	5.2	0.38	0.13	0.0	0 (0)	3 (0)
2	Physical environment	99.7	1.0	0.99	0.04	100.0	0 (0)	11 (11)
	Technological environment	96.8	7.8	0.89	0.25	87.5	0 (0)	8 (7)
	Cognitive factors	76.7	16.2	0.37	0.35	25.0	2 (0)	6 (2)
	Psycho-behavioral factors	91.0	10.5	0.71	0.29	73.3	0 (0)	15 (11)
	Adverse physiological state	97.5	6.4	0.92	0.21	87.5	0 (0)	16 (14)
	Physical/mental limitations	96.6	6.8	0.88	0.23	80.0	0 (0)	5 (4)
	Perceptual factors	98.3	4.6	0.94	0.16	90.9	0 (0)	11 (10)
	Coordination/communication /planning factor	87.1	14.6	0.62	0.35	58.3	0 (0)	12 (7)
Self imposed stress	99.1	2.1	0.97	0.08	100.0	0 (0)	6 (6)	
3	Inadequate supervision	84.5	12.4	0.52	0.32	50.0	1 (0)	5 (3)
	Planned inappropriate actions	85.0	15.8	0.57	0.33	71.4	0 (0)	7 (5)
	Failed to correct a known problem	85.3	3.7	0.49	0.11	0.0	0 (0)	2 (0)
	Supervisory violations	82.8	21.3	0.56	0.41	75.0	0 (0)	4 (3)
4	Resources/acquisition management	98.1	3.3	0.93	0.12	100.0	0 (0)	9 (9)
	Organizational climate	92.1	8.6	0.73	0.27	80.0	0 (0)	5 (3)
	Organizational processes	79.6	12.0	0.39	0.24	16.7	0 (0)	6 (1)

\* brackets indicate the number of nanocodes in the category in which  $r_{wg} \geq .6$ .

Table II. Inter-rater agreement results for the rotary wing scenario (n= 65 raters).

Level		Mean % agreement	St dev % agreement	Mean reliability	St dev reliability	% $r_{wg} \geq .6$	# majority selected	# majority rejected
1	Skill-based errors	91.8	10.00	0.73	0.32	66.7	0 (0)*	6 (4)
	Judgment & decision errors	84.6	12.98	0.53	0.37	33.3	1 (0)	5 (2)
	Misperception errors	93.8	-	0.77	-	100.0	0 (0)	1 (1)
	Violations	73.8	20.35	0.33	0.31	33.3	1 (0)	2 (1)
2	Physical environment	99.9	0.46	0.99	0.02	100.0	0 (0)	11 (11)
	Technological environment	99.2	1.42	0.97	0.06	100.0	0 (0)	8 (8)
	Cognitive factors	95.0	9.55	0.84	0.28	87.5	0 (0)	8 (7)
	Psycho-behavioral factors	84.1	16.73	0.56	0.38	53.3	3 (0)	12 (8)
	Adverse physiological state	97.5	8.00	0.93	0.22	93.8	0 (0)	16 (15)
	Physical/mental limitations	97.8	2.57	0.92	0.10	100.0	0 (0)	5 (5)
	Perceptual factors	97.6	5.48	0.92	0.18	90.9	0 (0)	11 (10)
	Coordination/communication /planning factor	84.2	14.13	0.53	0.36	50.0	2 (0)	10 (6)
3	Self imposed stress	92.1	14.48	0.76	0.39	83.3	0 (0)	6 (5)
	Inadequate supervision	84.4	8.28	0.49	0.22	33.3	1 (0)	5 (2)
	Planned inappropriate actions	91.0	13.80	0.73	0.33	85.7	0 (0)	7(6)
	Failed to correct a known problem	72.3	4.35	0.19	0.44	0.0	1 (0)	1 (0)
4	Supervisory violations	87.3	20.42	0.68	0.46	75.0	0 (0)	4 (3)
	Resources/acquisition management	99.1	1.74	0.97	0.07	100.0	0 (0)	9 (9)
	Organizational climate	85.5	12.00	0.54	0.33	40.0	1 (0)	4 (2)
	Organizational processes	89.2	4.96	0.62	0.16	50.0	0 (0)	6 (3)

\* brackets indicate the number of nanocodes in the category in which  $r_{wg} \geq .6$ .

There were only seven nanocodes in which 50% or greater of the participants agreed to select the nanocode in the fixed wing mishap (see Table I). In level one the only nanocodes for which greater than half of the participants agreed to select were: ‘checklist error’ (70.7%,  $r_{wg}=0.16$ ), ‘breakdown in visual scan’ (77.6%,  $r_{wg}=0.29$ ), ‘caution/warning ignored’ (50%,  $r_{wg}=0$ ), and ‘decision making during operations’ (58.6%,  $r_{wg}=0.01$ ); in level two ‘channelized attention’ (70.7%,  $r_{wg}=0.16$ ), and ‘cognitive task oversaturation’ (55.2%,  $r_{wg}=0$ ); and in level four, ‘leadership oversight inadequate’ (63.8%,  $r_{wg}=0.06$ ). The levels of inter-rater reliability were higher in the 139 nanocodes that the majority of participants agreed did not apply to the mishap.

The inter-rater reliability exceeded  $r_{wg}=0.6$  in 72% of these rejected nanocodes (see Table I). However, as evidenced from the low inter-rater reliabilities in certain categories, deciding whether or not to reject was more difficult for raters for those nanocodes from the ‘judgment and decision errors’, ‘cognitive factors’, ‘coordination/communication/planning factor’, and ‘organizational process’ categories (see Table I).

In the rotary wing mishap, there were ten nanocodes in which 50% or greater of the participants agreed to select the nanocode (see Table II). In level one the nanocodes in which at least half of the participants agreed to select were: ‘inadequate real-time risk assessment’ (67.7%,  $r_{wg}=0.11$ ), and ‘extreme violation’ (81.5%,  $r_{wg}=0.39$ ); in level two ‘overconfidence’ (86.2%,  $r_{wg}=0.52$ ), ‘complacency’ (80.0%,  $r_{wg}=0.35$ ), ‘more aggressive than necessary’ (69.2%,  $r_{wg}=0.13$ ), ‘lack of assertiveness’ (64.67%,  $r_{wg}=0.07$ ), and ‘command oversight inadequate’ (87.7%,  $r_{wg}=0.56$ ); in level three ‘failure to identify/correct risky behavior’ (69.2%,  $r_{wg}=0.13$ ); and in level four ‘organizational culture allows for unsafe mission demands/pressure’ (67.7%,  $r_{wg}=0.11$ ).

Similar to the fixed wing mishap, of the 137 nanocodes of which the participants thought did not apply, the inter-rater reliability exceeded  $r_{wg}=0.6$  in 88% of the nanocodes. Those

categories in which there were lower levels of agreement as to whether the nanocodes did not apply included: 'judgment and decision errors', 'psycho-behavioral factors', 'coordination/communication/planning factor', 'inadequate supervision', 'organizational climate', and 'organizational processes' (see Table II).

*Agreement with reference ratings.*

The 'experts' identified 21 nanocodes that applied to the fixed wing mishap, and 26 that applied to the rotary wing mishap. Table III shows those nanocodes for which a chi-square test indicated that, when compared to chance, there was significant agreement or disagreement ( $\chi^2 > 6.63$ ,  $df=1$ ,  $p < .01$ ). From Table III it can be seen that for the fixed wing mishap, there were significantly greater than chance agreement with the reference ratings in six (29%) of the nanocodes selected by the experts, significant disagreement with the reference ratings in eight (38%) of the selected nanocodes, and nonsignificant differences from chance in seven (33%) of the selected nanocodes. For the rotary wing mishap there were significantly greater than chance agreement for nine (35%) of the nanocodes selected by the experts, significant disagreement with the experts in nine (35%) of the nanocodes, and a nonsignificant difference for eight (38%) of the nanocodes.

Table III. Agreement with reference rating for which the nanocodes were selected as causal.

Category	Nanocode	% agree	Sig. agree	n.s.	Sig. disagree
<b>Fixed wing</b>					
Skill-based errors	Checklist error	70.7	X		
	Breakdown in visual scan	77.6	X		
Judgment & decision errors	Task misprioritization	74.1	X		
	Caution/warning- ignored	50.0		X	
	Decision making during operation	58.6		X	
Violations	Violation routine/widespread	13.8			X
Cognitive factors	Inattention	44.8		X	
	Channelized attention	70.7	X		
	Cognitive task oversaturation	55.2		X	
Psycho-behavioral factors	Emotional state	34.5		X	
Physical/mental limitations	Technical/procedural knowledge	15.5			X
Perceptual factors	Misinterpreted/misread instrument	15.5			X
Coordination/communications/ planning factor	Cross-monitoring performance	32.8			X
	Communicating critical information	46.6		X	
Inadequate supervision	Leadership oversight inadequate	63.8	X		
	Supervision- lack of feedback	17.2			X
Failed to correct a known problem	Personnel management	17.2			X
Supervisory violations	Directed violation	10.3			X
	Currency	48.3		X	
Organizational climate	Organizational values/culture	22.4			X
Organizational processes	Program & policy risk assessment	89.6	X		
<b>Rotary wing</b>					
Skill-based errors	Procedural error	18.8			X
	Overcontrol/ undercontrol	23.4			X
Judgment & decision errors	Risk assessment during operation	68.8	X		
	Caution/warning ignored	23.4			X
Violations	Violation routine/widespread	48.4		X	
	Violation- lack of discipline	81.3	X		
Cognitive factors	Inattention	28.1			X
Psycho-behavioral factors	Emotional state	46.9		X	
	Personality style	50		X	
	Overconfidence	85.9	X		
	Pressing	31.3			X
	Complacency	79.6	X		
	Overaggressive	68.8	X		
Adverse physiology	Fatigue- physiological/mental	32.8			X
Coordination/communications/ planning factor	Crew/ team leadership	71.9	X		
	Cross-monitoring performance	34.4		X	
	Assertiveness	65.6		X	
	Communicating critical information	17.2			X
	Mission briefing	37.5		X	
Inadequate supervision	Leadership oversight inadequate	89.1	X		
	Supervision- lack of feedback	29.7			X
Planned inappropriate actions	Crew composition	40.6		X	
Failed to correct a known problem	Personnel management	68.8	X		
	Operations management	25			X
Supervisory violations	Supervision- discipline enforcement	42.2		X	
Organizational climate	Organizational values/culture	67.2	X		

For those nanocodes that were rejected by the experts, there were much higher levels of agreement between the reference ratings and the participants' responses, than for those nanocodes that were selected by the experts. From Table IV it can be seen that there was a significantly greater than chance agreement between the reference ratings and the participants for 98% of the rejected nanocodes in the fixed wing mishap, and 99% in the rotary wing mishap.

Table IV. Agreement with reference rating when the nanocodes were rejected as being causal to the mishap.

Level	Category	Fixed wing				Rotary wing			
		mean % agree	st dev	sig	ns.	mean % agree	st dev	sig	n.s.
1	Skill-based errors	89.7	16.4	3	1	98.0	2.3	4	0
	Judgment & decision errors	84.5	10.3	3	0	91.0	10.8	4	0
	Misperception errors	74.1	-	1	0	93.8	-	1	0
	Violations	78.4	3.7	2	0	89.1	-	1	0
2	Physical environment	99.7	1.0	11	0	99.9	0.5	11	0
	Technological environment	96.8	7.8	8	0	99.2	1.4	8	0
	Cognitive factors	86.6	9.8	5	0	98.2	2.9	7	0
	Psycho-behavioral factors	89.9	18.1	14	0	94.6	7.3	9	0
	Adverse physiological state	97.5	6.4	16	0	99.6	0.9	15	0
	Physical/mental limitations	99.6	0.9	4	0	98.1	2.0	5	0
	Perceptual factors	99.7	0.7	10	0	97.6	5.6	11	0
	Coordination/communication/planning factor	87.0	19.4	10	0	94.4	5.2	7	0
Self imposed stress	99.1	2.1	6	0	91.9	14.7	5	1	
3	Inadequate supervision	86.8	6.5	4	0	86.7	6.4	4	0
	Planned inappropriate actions	85.0	15.8	7	1	96.4	2.4	6	0
	Failed to correct a known problem	87.9	-	1	0	-	-	0	0
	Supervisory violations	94.8	7.3	2	0	97.4	3.3	3	0
4	Resources/acquisition management	98.1	3.3	9	0	99.1	1.8	9	0
	Organizational climate	95.7	3.3	4	0	89.8	7.8	4	0
	Organizational processes	83.8	6.9	5	0	89.1	5.0	6	0

## DISCUSSION

### *Internal validity*

As described in the results section, there were reasonable levels of inter-rater reliability in the majority of the nanocodes that were not considered to be causal to the mishaps. However, for the small subset of nanocodes to which at least half of the raters thought applied to the mishaps, acceptable levels of inter-rater reliability were not achieved. Obviously, reliably classifying the cause of a mishap is equally, if not more important, than reliably rejecting potential mishap causal factors.

An explanation for the high reliability in rejected nanocodes is that for many it is clear that they do not apply to the mishap. To illustrate, if weather, a failure of technology, and medical or physiological conditions were clearly not a factor, then 35 nanocodes can quickly be disregarded by the raters. However, the lack of a consensus between raters for the selected nanocode is concerning. It is suggested that the reason for this finding is that raters may pick from a range of similar nanocodes that have overlapping concepts. To illustrate, the definition of the nanocode overconfidence is: “overconfidence is a factor when the individual overvalues or overestimates personal capability, the capability of others or the capability of aircraft/vehicle or equipment and this creates an unsafe situation” (p. 9; 5). The definition of the nanocode complacency is: “complacency is a factor when the individual’s state of reduced conscious attention due to an attitude of overconfidence, under-motivation or sense that others have the situation under control leads to an unsafe situation” (p. 9; 5). Some evidence for the effect of conceptually overlapping nanocodes is that for those categories in which more than half of the raters selected a particular nanocode, there tended to be lower levels of reliability for the rejected nanocodes from the same category than for those categories in which no nanocodes were selected.

Although there have been arguments that HFACS utilizes under-specified labels (2), it would appear that DOD-HFACS includes labels that are over-specified. A balance must be struck between an error taxonomy that is so small that it provides insufficient information, and so big that it is unwieldy and has low levels of reliability. The result of the over-specification is that it may not be possible to achieve acceptable levels of inter-rater reliability, even with extensive rater training. An error taxonomy should be parsimonious, and consists of discrete subcategories that are sufficiently distinct to lead to high levels of inter-rater agreement (23). Taxonomies that use mutually exclusive and exhaustive descriptors have higher internal validity than those that do not (2).

To improve the parsimony of DOD-HFACS it is recommended that the individual nanocodes be subjected to detailed expert scrutiny to establish whether they could be removed, or combined with other nanocodes. The author believes that it would be possible to improve the parsimony of DOD-HFACS by reducing the number of nanocodes by at least a third. The refinement of error classification system is a necessary process for increasing reliability. To illustrate, it took four iterative processes, and a number of reliability studies, to develop the current version of HFACS (32).

The level of inter-rater reliability could also be improved through more user training and exposure to DOD-HFACS. A large body of research exists on training observers to reliably use behavioral markers (1). It may be possible to adapt some of the principles of such training to instruct mishap investigators in using DOD-HFACS.

To summarize, it is suggested that reducing the number of nanocodes, increasing the mutual exclusivity of the remaining nanocodes, and increasing the exposure and training of DOD-HFACS users would have beneficial effects on the internal validity of the system.

### *External validity*

As stated in the introduction, the purpose of DOD-HFACS is to provide a common human error categorization system to accurately capture the human factors causes of all DOD mishaps. Similar to the inter-rater reliability, when compared to the reference ratings, there were high levels of agreement for the nanocodes rejected by the ASO students, but low level of agreement with the selected reference ratings. Also, although the reference rating may not have been selected, nanocodes that were conceptually similar were chosen. Therefore, improvements in the parsimony and mutual exclusivity of the nanocodes would also have benefits for external validity of DOD-HFACS.

### *Utilitarian criteria*

Utilitarian criteria is concerned with the extent to which a taxonomy is useful and efficient. A reliable U.S. DOD system to classify the human factors causes of all mishap classification system would be beneficial to all departments of the U.S. DOD. It would allow comparisons of mishap causal factors across services and domains. However, if DOD-HFACS is to be used to collect human factors causes of mishaps in domains other than military aviation, care should be taken to ensure that the system is suitable for this purpose.

Just as there have been problems with taking crew resource management training developed for one domain and applying it to another (8,13), the same may be true of applying a human error taxonomy to a domain for which it was not developed. There are a number of considerations that should be addressed if DOD-HFACS is to be used to classify mishaps in non-military aviation domains.

It is crucial that subject matter experts from the new domain to which the error taxonomy is to be used are consulted. This is necessary for a number of reasons. Firstly, to avoid, or

translate, domain specific language that is not understood by operators in the new domain. To illustrate, although crew resource management is a widely used and understood concept in aviation, the converse is true in U.S. Navy diving. Therefore, it was unsurprising that in an examination of the analysis of 263 navy diving mishaps, O'Connor et al (23) did not find a single example of the use of the crew resource management subcategory of the error classification system used by the U.S. Navy. Secondly, some of the nanocodes may be redundant in the new domain. For example, it is possible to think of a number of domains in which the effects of g-forces, hypoxia, or evolved gas disorders may never be applicable. Other domains may require specific nanocodes that are relevant to the particular types of operations being carried out. For example, preconditions such as contaminated breathing gas, dangerous marine life, or contaminated water may be required to classify navy diving accidents. Finally, care should also be taken in utilizing an error taxonomy developed for use in a domain with a mature safety culture, and applying it to other domains with a less developed safety culture in which human error is less widely understood and reported by operators. To illustrate, an error taxonomy developed for use in aviation, may not be useful to classify mishaps in industries such as fishing or construction.

Another factor that has a detrimental effect on the usefulness of an error classification system is the need for extensive training. For example, to reach acceptable levels of inter-rater agreement using HFACS to classify the causes of aviation accident, Li and Harris (21) provided 10 hours of training, and Shappell et al (26) and Wiegmann et al (31) provided 16 hours of instruction including lectures and practice using the framework. Although the amount of training time is not extreme, it is far in excess of the two hours of classroom instruction in DOD-HFACS that were provided to raters in the current study. Further, given the added layer of complexity of the DOD-HFACS nanocodes, it is likely that more training time would be required to reach

acceptable levels of reliability using DOD-HFACS as compared to the HFACS framework.

Therefore, although a common U.S. DOD human factors mishap classification tool is desirable, it may need to be adapted for use in other domains, and even with the changes suggested in the discussion of internal and external validity, extensive training would be required to use DOD-HFACS reliably.

### *Methodological considerations*

The classification of the human factors causes of a written mishap summary by raters working as individuals is not an accurate reflection of how a real U.S. Naval aviation mishap would be investigated. Should a naval squadron have an aviation mishap, an aviation mishap board (AMB) will be formed to investigate. At a minimum, the AMB will consist of: an ASO, 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 (3). The AMB will draw upon many resources to identify the causes of a mishap. Therefore, it is unclear how group discussion would effect the reliability and accuracy of the DOD-HFACS coding of a mishap.

The methodology used in this paper represents a departure from the reliability studies that are more frequently reported by researchers evaluating the effectiveness of error classification systems. Researchers have tended to compare the inter-rater reliability of pairs of specially trained raters evaluating a large number of written mishap reports (10,21,26,27,29,30,32,33). It is suggested that when evaluating the reliability of error classification systems, researchers should give consideration to comparing the responses of multiple potential end-users of the system, rather than just pairs of specially trained raters.

## **CONCLUSION**

For DOD-HFACS to be useful and efficient in allowing mishap trends to be tracked over time, it is crucial that the system can be used with acceptable levels of reliability and accuracy. This study has identified that more parsimony, increased mutual exclusivity, and training are required to utilize DOD-HFACS effectively. If a mishap classification system fails to accurately capture the causes of mishaps, this can lead to the misdirection of manpower and funds, representing a missed opportunity to prevent the next mishap.

## **ENDNOTES**

All opinions stated in this paper are those of the author and do not necessarily represent the opinion or position of the U.S. Navy, Naval Aviation School Command, or the Navy/Marine Corps School of Aviation Safety.

## REFERENCES

1. Baker DP, Mulqueen C, Dismukes KR. Training pilot instructors to assess CRM: The utility of frame-of-reference (FOR) training. In: Proceedings of the International Aviation Training symposium. Oklahoma City, Oklahoma, 1999: 291-300.
2. Beaubien JM, Baker DP. A review of selected aviation human factors taxonomies, accident/incident reporting systems and data collection tools. *International Journal of Applied Aviation Studies* 2002; 2: 11-36.
3. Chief of Naval Operations. Naval aviation safety program, OPNAVINST 3750.6R. 2001; Retrieved 21 September 2007 from <http://www.safetycenter.navy.mil/instructions/aviation/opnav3750/default.htm>
4. Dismukes K, Berman B, Loukopoulous L. *The Limits of Expertise. Rethinking Pilot Error and the Causes of Airline Accidents*. Aldershot, UK: Ashgate; 2007.
5. DOD HFACS: A mishap investigation and data analysis tool. 2005; Retrieved on 21 September 2007 from <http://www.safetycenter.navy.mil/hfacs/downloads/hfacs.pdf>.
6. Fleishman EA, Mumford MD. Evaluating classifications of job behavior: A construct validation of the ability requirements scales. *Personnel Psychology* 1991; 44: 523-575.
7. Fleishman EA, Quaintance MK. *Taxonomies of Human Performance: The Description of Human Tasks*. Orlando, FL: Academic Press; 1984.
8. Flin R, O'Connor P, Mearns K. Crew resource management: Improving safety in high reliability industries. *Team Performance Management* 2002; 8: 68-78.
9. Gordon R, Flin R, Mearns K. Designing and evaluating a human factors investigation tool (HFIT) for accident analysis. *Safety Science* 2005; 43: 147-171.
10. Graur D. Human Factors Analysis and Classification System applied to civil aviation accidents in India. *Aviation, Space, & Environmental Medicine* 2005; 76: 501-505.

11. Hamman WR, Beaubien, MJ, Holt, RW. Evaluating instructor/evaluator inter-rater reliability from performance database information. In: Jensen R (Ed.) Proceedings of the 10th International Symposium on Aviation Psychology. Columbus, Ohio, 1999: 1214-1219.
12. Health and Safety Executive. Strategies to promote safe behavior as part of a health and safety management system. London, UK: HSE; 2002.
13. Helmreich RL. On error management: Lessons from aviation. *British Medical Journal* 2000; 320: 781–785.
14. Hollnagel E. Human reliability analysis: context and control. London, UK: Harcourt Brace; 1993.
15. James LR, Demaree RG, Wolf G. Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology* 1984; 69: 85-98.
16. James LR, Demaree RG, Wolf G.  $R_{wg}$ : An assessment of within-group interrater agreement. *Journal of Applied Psychology* 1993; 78: 306-309.
17. Joint Services Safety Chiefs. Policy in the collection and analysis of mishap human factors data. May 2005; Retrieved on 21 September 2007 from <http://www.safetycenter.navy.mil/hfacs/downloads/hfacsMOA.pdf>
18. Kayten P. The accident investigator's perspective. In: Wiener E, Kanki B, Helmreich R (Eds). *Cockpit Resource Management*. San Diego, CA: Academic Press, 1993: 283-310.
19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33(1): 159-174.
20. Law RJ, Sherman PJ. Do raters agree? Assessing inter-rater agreement in the evaluation of air crew resource management skills. In: Jensen R (Ed.) Proceedings of the 8th International Symposium on Aviation Psychology. Columbus, Ohio, 1995: 608-612.

21. Li W-C, Harris D. Pilot error and its relationship with higher organizational levels: HFACS analysis of 523 accidents. *Aviation, Space, and Environmental Medicine* 2006; 77: 1056–1061.
22. O'Connor P, Hörmann H-J, Flin R, Lodge M, Goeters KM, the JARTEL group. Developing a method for evaluating CRM skills: A European perspective. *International Journal of Aviation Psychology* 2002; 12(3): 263-286.
23. O'Connor P, O'Dea A, Melton J. A methodology for identifying human error in U.S. Navy diving accidents, *Human Factors* 2007; 49(2): 214-226.
24. Reason J. *Human Error*. New York: Cambridge University Press; 1990.
25. Reason J. *Managing the risks of organisational accidents*. Aldershot, UK: Ashgate, 1997.
26. Shappell S, Detwiler C, Holcomb K, Hackworth C, Boquet A, Wiegmann DA. Human error and commercial aviation accidents: An analysis using the human factors analysis and classification system. *Human Factors* 2007; 49: 227-242.
27. Shappell SA, Wiegmann DA. *A human error analysis of general aviation controlled flight into terrain accidents occurring between 1990-1998*. Washington DC: Federal Aviation Administration, 2003.
28. 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.
29. Tvaranyas AP, Thompson WT, Constable SH. Human factors in remotely piloted aircraft operations: HFACS analysis of 221 mishaps over 10 years. *Aviation, Space, and Environmental Medicine* 2006; 77(2): 724-734.

30. Van Dyck C, Frese M, Baer M, Sonnentag S. Organizational error management culture and its impact on performance: A two-study replication. *Journal of Applied Psychology* 2005; 90: 1228-1240.
31. Wiegmann DA, Boquet A, Detwiler C, Holcomb K, Shappell S. Human error and general aviation accidents: A comprehensive, fine-grained analysis using HFACS. Oklahoma City: Federal Aviation Authority; 2005.
32. Wiegmann DA, Shappell SA. Human error and crew resource management failures in Naval safety center data, 1990-96 *Aviation, space and environmental* 1999; 70: 1147-1151.
33. Wiegmann DA, Shappell SA. *A Human Error Approach to Aviation Accident Analysis*. Aldershot, UK: Ashgate; 2003.
34. Wiegmann DA, Shappell, SA. Human error analysis of commercial aviation accidents: Application of the human factors analysis and classification system (HFACS). *Aviation, Space, and Environmental Medicine* 2001; 72: 1006–1017.