2015

A Comprehensive Effort to Arrive at an Optimally Reliable Human Factors Taxonomy

Raymond E. King

A COMPREHENSIVE EFFORT TO ARRIVE AT AN OPTIMALLY RELIABLE HUMAN FACTORS TAXONOMY

Raymond E. King, Headquarters, Air Force Safety Center, Human Factors Division, Kirtland Air Force Base, NM

Department of Defense (DoD) members sought to improve the inter-rater reliability of the DoD Human Factors Analysis and Classification System, (DoDHFACS). DoDHFACS differs from the original system developed by Wiegmann and Shappell (2003), based on the work of Reason (1990), by further analyzing mishaps and hazards to a more granular level – arriving at specific "nanocodes." The steps involved in the effort included determining which of the 147 "nanocodes" were rarely/never used and collapsing nanocodes and rewriting definitions to arrive at 109 nanocodes. Next, a stepwise checklist to guide investigators through consideration of nanocodes was created. Student investigators were guided to continually test checklists and generate results to gauge inter-rater reliability. They were asked to offer constructive criticism to hone checklist questions. While inter-rater reliability results are encouraging (Fleiss' Kappa of .847 at the broadest level), additional work is necessary to realize the goal of an optimally reliable human factors taxonomy.

As will be demonstrated, "Human factors" are causal or contributory in a majority of military aviation mishaps. This paper reports on a Department of Defense (DoD) effort to improve the system used to categorize causal and contributing human factors. Specifically, recent attempts to improve coding methods with the goal of achieving better inter-rater reliability and ultimately more actionable recommendations to improve safety will be described.

Hartmann (1977), in a widely read and highly regard article, asserted that reliability is a necessary but not a sufficient basis for validity. Hartmann went on to specify that there are two methods that can be employed to determine reliability: percentage agreement reliability and reliability coefficient. Hartmann advocated for the latter over the former as percentage agreement may produce inflated estimates of reliability. Another issue to bear in mind when considering reliability is that categories must be mutually exclusive and exhaustive (that is, contain no overlapping elements and be complete) to achieve the highest reliability. Overlapping elements may result in observers using different categories for the same observation and thus finding fewer distinctions between entities being compared.

The roots of the Human Factors Analysis and Classification System (HFACS), are described in "Taxonomy of Unsafe Operations" (Shappell & Wiegmann, 1997) and catalogued in a Federal Aviation Administration (FAA) technical report (Shappell & Wiegmann, 2000) and a book "A Human Error Approach to Aviation Accident Analysis" (Wiegmann & Shappell, 2003). Their system is built upon the "Swiss cheese" model of Reason (1990). Reason recommended that a mishap investigation start with the *unsafe act(s)*, which represent(s) active failure. The investigation does not stop there, however, as latent failures and conditions are examined next. Latent conditions may exist undetected and unexpressed for years and include: *preconditions*, *unsafe supervision*, and *organizational influences*.

According to Reason (1990), unsafe acts include both errors and violations. Errors may be skill-based or may be due to decisional or perceptual factors. Violations may be routine (such as cutting the same corners that many others cut) or exceptional. Preconditions for unsafe acts include environmental (physical or technological) factors, conditions of operators (adverse mental states or adverse physiological states or physical/mental limitations) or personnel factors (crew resource management or personal readiness). Reason (1990) argued that it is also essential to investigate at the supervisory and/or organizational level because such factors have direct impact on preconditions. Addressing preconditions is likely to reveal opportunities to improve safety.

Unsafe supervision includes: inadequate supervision, supervisors planning inappropriate operations, a supervisor failing to correct a known problem, and supervisory violations. Finally, there are organization influences, to include resource management, organizational climate, and organizational process. One way to conceptualize these categories is to consider them "bins" containing the smaller units.

Results of a query of the Air Force Automated System (AFSAS) database, which is accessed via a secure website, for fiscal years 2010 through 2013 (1 October 2009 through 30 September 2013) to assess overall human factors involvement in aviation mishaps is depicted in Table 1. These numbers empirically demonstrate that human factors do, in fact, comprise a major concern for aviation safety.

Table 1.
Aviation Mishaps for FY 2010 – 2013

Class	Total Number Aviation Mishaps	Aviation Mishaps with at Least 1 Human Factors Code	Percentage of Aviation Mishaps with at Least 1 Human Factors Code	Total Number of Human Factors Codes
A*	129	113	87.60%	1,452
B*	218	113	51.83%	754
C*	2,518	895	35.54%	2,586
D*	3,142	737	23.46%	1,179
E*	28,803	1,094	3.8%	3,185
Grand Total	34,810	2,952	59.59%	9.156

*As defined in AFI 91-204, 12 February 2014.

It should be noted that DoD HFACS has not been required to be used for Classes C, D, and E mishaps (hi-lited). The reader is thus cautioned not to be misled by the lower percentages and the deflating impact on the grand total. The involvement of human factors is therefore likely heavily underestimated in USAF mishaps, particularly Class C, D, and E mishaps, as a result.

Beaubien and Baker (2002) while generally favorable in their review of HFACS, note that HFACS is a bit coarse, as it does not delineate *reasons* for the conditions it identifies. Beaubien and Baker also note that latent failures are difficult to identify in mishap analysis. The context of their review must be appreciated as they were examining coding schemes that were used with data already collected. Their final point is important: HFACS categories are nominal and not sequential and thus do not reveal a chain of events. Therefore, they do not differentiate causes from effects. That issue, however, is relatively easy to remedy in the overall scheme of an investigation. For example, the USAF constructs a mishap sequence of contributory and causal findings, and embeds DoD HFACS within it. O'Connor (2008) noted the above criticisms and detailed the efforts to address them, to include the formation of a Department of Defense (DoD) Working Group in 2003, which created DoD HFACS. DoD HFACS introduced increased granularity, an additional level of classification: "nanocodes." The original DoD HFACS included 147 nanocodes, organized under the categories (bins) delineated above (*unsafe acts, preconditions, unsafe supervision, organization influences*). O'Connor (2008) examined the reliability of DoD HFACS, version 6.2. He found that U.S. Navy and Marine aviators undergoing mishap investigator training were unable to achieve acceptable reliability, but noted that they had received only minimal training. Although the raters were able to agree on the nanocodes *not* used, they were unable to achieve consistent agreement concerning which nanocodes applied ("there were only seven nanocodes in which 50% or greater of the participants agreed to select the nanocode," *p.* 602). O'Connor noted that raters were confused by the number (147) of available nanocodes and that the nanocodes contained overlapping concepts. O'Connor found that collapsing codes improved inter-rater reliability. O'Connor therefore argued for nanocodes that are exhaustive, parsimonious, and mutually exclusive. O'Connor also noted that his research participants may not have been reading and considering the nanocodes' one-paragraph definitions, relying instead on the names of the nanocodes.

O'Connor called for subject matter experts to review the nanocodes to determine if some could be removed or combined with other nanocodes. O'Connor even went as far as to suggest that the nanocode level be abandoned if acceptable reliability could not be achieved without extensive training. A 2011 Aerospace Medical Association presentation, *DoD Human Factors Analysis and Classification System X*, prepared by human factors practitioners (Brian T. Musselman, Jeffrey D. Alton, Thomas G. Hughes, Patricia LeDuc, Richard J. Farley, & Antonio B. Carvalhais) from the three service safety centers had four expert raters code 54 USAF Class A mishaps with DoD HFACS version 6.2. They found a Kappa coefficient of .5494 with 76 out of 147 (52%) nanocodes having reliability greater than or equal to .60. The authors recommended: "Improve code definition," and development of an "organized training curriculum." Subsequent studies used DoD HFACS X, which contained fewer nanocodes (102, rather than 147). The average Kappa coefficient increased to an impressive 0.84 with expert coders, but novice coders continued to struggle, achieving Kappa coefficients of .2453 and .3239. The authors urged the development of a decision-tree algorithm, redesign of DoD HFACS into larger buckets (even if granularity would be sacrificed), and limiting coding at the nanocode level to experts only.

The steps in the current effort to improve DoD HFACS included determining the frequency that each of the nanocodes was used and considering retiring those nanocodes that were very infrequently used. Nanocodes that

were similar in the phenomena they described, as evidenced by having overlapping definitions, were merged and the definitions reworked. The goal was to reduce the number of nanocodes and to improve the mutual exclusivity of the remaining nanocodes. Specifically, AFSAS was further queried for fiscal years 2010 through 2013 for aviation and ground mishaps to determine the frequency of use of each of version 6.2's 147 nanocodes. It should be noted that AFSAS was queried to arrive at two totals: the first count tallied a specific HFACS nanocode cited once per mishap. Otherwise, a given nanocode assigned against multiple members of a crew would inflate the total. The other tally counted the grand total of HFACS nanocodes used, with no restriction on how many times a nanocode was used in any given mishap. The US Army and Navy, as members of the DoD HFACS Working Group, performed similar tallies. In the USAF, for example, PC 201 used only once for all classes of aviation and ground mishaps. Finally the DoD HFACS Working Group ensured that nanocodes were aligned in the correct bins. Nanocodes that were relocated to other bins were reassigned an alphanumeric to be consistent with the bin the new bin. Ultimately, the 147 nanocodes in version 6.2 were collapsed to109 nanocodes in version 7.0. The Working Group then developed a checklist, colloquially known as "Turbo HFACS," that uses a decision tree to guide investigators. A response of "yes" guides the investigator to the correct "bin" and suggests a list of defined nanocodes. This paper delineates the motivation to change DoD HFACS 6.2 and to document the changes made in DoD HFACS 7.0. This paper also examines the inter-rater reliability of DoD HFACS.

Method

Participants

Three hundred and forty students attending USAF aircraft mishap investigation courses served as participants. Most of the participants were pilots and maintenance personnel attending the Aircraft Mishap Investigation Course (AMIC) at the Headquarters, Air Force Safety Center (HQ AFSEC), Kirtland AFB, NM. Additional data was collected from aerospace medical personnel (flight surgeons, aerospace physiologists, and clinical psychologists) who attended the Aircraft Mishap Investigation and Prevention (AMIP) course, held at the at the USAF School of Aerospace Medicine (USAFSAM), Wright-Patterson AFB, OH.

Procedures

Participants were given approximately 45 minutes to read and code sanitized (basic identifying information had been removed) synopses of mishap reports that had been investigated by Safety Investigation Boards (SIBs). The synopses were approximately two typed, single-spaced, pages in length, using a 10-point font. To protect privilege, all mishap reports were immediately collected at the conclusion of the exercises. These mishap synopses are not published here because the degree of additional sanitizing that would have been necessary to publish them in this report would have rendered them virtually incomprehensible. The research design for this project was not strictly pre-planned, but rather evolved and capitalized on opportunities that presented themselves (see Table 2).

Table 2.
Summary of the Evolution of DoD HFACS version 7.0 Research Activities.

First Trials	Second Trials	Third Trials
Student investigator teams provided Checklist only.	Student investigator teams given answer sheets along with Checklist and required to submit responses on it,	Student investigator teams given answer sheets along with Checklist and required to submit responses on it.
Student investigator teams were directed to use only DoD HFACS version 7.0 for exercise.	Student investigator teams were directed to use DoD HFACS version 6.2 and then introduced to version 7.0 for exercise.	Student investigator teams were taught to use DoD HFACS version 6.2 and then introduced to version 7.0 for exercise.
18-question version of Checklist used.	8-question (with sub questions) version of Checklist used.	8-question (with sub questions) version of Checklist used.
	Student investigator teams asked to list the three to five (and then the five) most important HFACS nanocodes.	Student investigator teams asked to list the five most important HFACS nanocodes.

First Trials.

In the first data collections, 31 student investigator teams used an 18-question version of the DoD HFACS 7.0 Checklist to code three aircraft mishap scenarios. The author presented a brief introduction (approximately 10 minutes) to the DoD HFACS 7.0 Checklist. The participants were directed to use the questions of the Checklist and work together in teams of two or three members. Following the advice of O'Connor and Walker (2011), participants were organized into small teams rather than working alone to better simulate the conditions of a safety investigation board. Participants were instructed to not speak to any member of *another* team about the mishap during the exercise.

Second Trials.

The next series of data collection aimed to directly compare DoD HFACS 6.2 to DoD HFACS 7.0. Student investigator teams were given a mishap scenario and directed to first use version 6.2 as outlined in AFI 91-204. The student investigative teams conclusions were compared to the outcomes as determined by the actual Safety Investigation Board (SIB) and reviewed by the Memorandum of Final Evaluation (MOFE). After the student rater teams' responses using version 6.2 were collected, the teams were trained to use version 7.0 (using basically the same introduction described above) and directed to again code the scenario, without regard to what they coded using version 6.2. To encourage student investigator teams to read nanocode definitions, they were required to record their answers on sheets that only contained the alphanumeric codes, so that they would not base their decisions merely on the names of the nanocodes, without reading and considering the full definition. Moreover, rater groups were asked to list the three to five nanocodes that were the most important in the mishap, of course starting with those that they deemed causal. Following the input from the epidemiologists identified in the Acknowledgements, the participants were ultimately directed to list the five most important DoD HFACS nanocodes. The actual SIB and the MOFE found 12 DoD HFACS nanocodes to be applicable.

Third Trials.

Another data collection was held using a mishap that had been coded with fewer nanocodes by the SIB and which included only nanocodes that transitioned to version 7.0. Because the purpose of AMIC is to train investigators and not serve as a research laboratory, this AMIC class received more detailed instruction on a strategy to use DoD HFACS 6.2. The student investigators needed to be prepared to investigate mishaps immediately upon the completion of their training and there was no start date yet established for the operational transition to version 7.0. In applying version 6.2, student investigators were urged to read and consider definitions rather than just rely on the one-page wire diagram. After these student investigator teams completed their coding with version 6.2, their answer sheets were collected. These student investigator teams were then introduced to version 7.0, using basically the same instruction used in the first two trials.

Qualitative Feedback.

The feedback received from students led the authors and the rest of the Working Group to continually refine questions, eventually arriving at a solution of eight questions with sub-questions. Students were subsequently asked to provide written feedback on their opinions of the changes made in HFACS 7.0.

Results

First Trials.

During the first series of data collection, 18 of 31 (58%) rater teams selected the identical "yes" pattern when coding Scenario One using DoD HFACS version 7.0. Four of the 31 (13%) rater teams selected an identical but alternate pattern. Twenty-four (77%) rater teams selected the same nanocodes as the top three (out of 109) overall codes. The Fleiss' Kappa in considering the responses to the 18 questions was .847. The Fleiss' Kappa for the 109 nanocodes was .545 and the average Pairwise Cohen's Kappa was .543.

The 18-queston version of DoD HFACS, version7.0 did not fare as well with two other scenarios. In 25 rater teams coding Scenario Two, only four rater teams selected an identical "yes" pattern. There were two other common patterns with each being selected by two rater teams. Twenty (80%) rater teams selected the same top three codes. The Fleiss' Kappa in considering the responses to the 18 questions was .498. The Fleiss' Kappa for the 109 nanocodes was .415 and the average Pairwise Cohen's Kappa was .400.

Scenario Three had two of fifteen rater teams selecting an identical yes pattern. Five codes were selected 10 or more times by rater teams. Fifteen rater teams selected the top three overall codes. The Fleiss' Kappa in

considering the responses to the 18 questions was .550. The Fleiss' Kappa for the 109 nanocodes was .487 and the average Pairwise Cohen's Kappa was .512.

Second Trials.

As seen in Table 3, during the exercise using Mishap #1, when the student rater teams used DoD HFACS 6.2, nine out of 14 rater teams (64%) matched at least one of the above findings as being among their most important three to five DoD HFACS nanocodes. One rater team of 14 (7%) matched three nanocodes; eight rater teams (57%) had one match, and five rater teams (36%) had no matches of their top three to five DoD HFACS nanocodes to those of the SIB. Using DoD HFACS 7.0, two student rater teams (14%) had three matches; three student rater teams (21%) had two matches, nine student rater teams (64%) had one match, and zero student rater teams had no matches. Making this contrast even more stark (and more favorable to version 7.0) is the fact that two of the nanocodes identified in Mishap #1 using DoD HFAC version 6.2 did not transition to version 7.0 and thus were not available to the raters during the version 7.0 portion of the exercise.

Table 3.
Comparing DoD HFACS 6.2 to 7.0 Anchored Against Actual SIB Results, Mishap #1

Number of Matches to Actual SIB	DoD HFACS 6.2	DoD HFACS 7.0
3 Matches	1 student rater team matched SIB	2 student rater teams matched SIB
2 Matches		3 student rater teams matched SIB
1 Match	8 student rater teams matched SIB	9 student Rater teams matched SIB
0 Matches	5 student rater teams	

Third Trial.

The results of the exercise using Mishap #2 are presented in Table 4. Two rater teams elected to list only four codes as the "most significant" during the version 7.0 portion of the exercise and could not be persuaded to list more. By doing so, they lessened the opportunity to maximize matching what the SIB found.

Table 4.
Comparing DoD HFACS 6.2 to 7.0, Anchored Against Actual SIB Results, Mishap #2

Number of Matches to Actual SIB	DoD HFACS 6.2	DoD HFACS 7.0
4 Matches	3 student rater teams matched SIB	1 student rater team matched SIB
3 Matches	7 student rater teams matched SIB	5 student rater teams matched SIB
2 Matches	4 student rater teams matched SIB	5 student rater teams matched SIB
1 Match		2 student rater teams matched SIB
0 Matches	1 student rater team matched SIB	

Finally, student raters were asked to provide written feedback on their perception of the relative value of version 7.0 over version 6.2. Initially, the comments were mostly neutral as they are criticisms of both versions. The comments became much more positive. (nine out of 14, with no negative comments). Previously, comments from students were collected in a more informal fashion, but were still useful in the evolution of the checklist. Some typical themes from the student investigators included that the new version is "less intimidating" and "less subjective" and gives investigations structure. Suggestions for improvement included the observation that some of the questions are too broad and the sub questions need to be read and considered even if the instructions advise users to skip over the sub questions.

DISCUSSION

In a series of comparison using a variety of mishap scenarios, the checklist for DoD HFACS version 7.0 performed well. These encouraging results can be explained as follows: A taxonomy that has fewer nanocodes and nanocodes that have distinct meaning improves user satisfaction and may, itself, increase inter-rater reliability. While a systematic approach to considering the larger categories as well as the nanocodes likely is a key component to the improvement of inter-rater reliability, simply encouraging student investigators to consult the definitions of the nanocodes also likely improved inter-rater reliability. Systematically guiding investigators to consider all

nanocodes will increase the likelihood that the definitions of the nanocodes will be read and considered. As pointed out by previous researchers as noted in this report, requiring coding at a finer degree of granularity requires training and providing investigators with the proper resources, such as a checklist.

As noted by the participants, another issue is the correct structure of the checklist questions. Too many questions are likely to try the patience of investigators, while fewer questions with sub-questions run the risk of investigators missing significant areas that could benefit from further inquiry. The feedback gleaned from the students who graciously participated in this research suggest that investigators would be wise to not skim over sub-questions after answering "no" to the major question. While the DoD HFACS Working Group should consider honing the questions and sub-questions, DoD HFACS 7.0 is a step in the right direction according to student feedback and the results obtained in this study. A future revision should revise the questions and elevate some of the sub questions to free standing questions. Above all, any strategy that gets investigators to read and consider the definitions of the nanocodes will result in a better investigative outcome. The "yes/no" format of the questions in version 7.0 results in a clear binning (getting in the ballpark of applicable nanocodes). Such binning of causes and contributing factors represents an advancement in investigations with actionable results as it allows leaders to more accurately allocate resources to reduce future mishaps. Even if there is some disagreement as to which exact nanocode within a bin is the cause, at least the correct bin is identified and proper attention is paid to mitigation of a major cause or contributing factor of mishaps.

Future efforts should include a continued refinement of the questions, as noted above, as well as the creation of a small set of questions to assist Aviation Safety Action Program (ASAP) reporters submit reports that more clearly highlight human factors issues. ASAP reports are considered "safety without the mishap," and thus could better use of DoD HFACS actually help improve safety. Above all, investigators in training in all services must be given ample opportunity to practice investigating and coding mishaps during the "organized training curriculum" advocated by Musselman, et al.

REFERENCES

Beaubien, J. M. & Baker, D.P. (2002). A review of selected aviation human factors taxonomies accident/incident reporting systems and data collection tools. *The International Journal of Applied Aviation Studies, 2,* 11-36.

Hartmann, D.P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis, 10,* 103-116.

O'Connor, P. (2008). HFACS with an additional layer of granularity: Validity and utility in accident analysis. *Aviation, Space and Environmental Medicine, 79,* 599 – 606.

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.

Reason, J. (1990). *Human Error*. New York: Cambridge University Press.

Shappell, S.A. & Wiegmann, D.A. (1997). A human error approach to accident investigation: The taxonomy of unsafe operations. The International Journal of Aviation Psychology, 7(4), 269-291.

U.S. Air Force. AFI 91-204, 12 February 2014, Washington, DC: HQ SEC/SEF.

Wiegmann, D. A. & Shappell, S. A. (2003). A human error approach to aviation accident analysis. Burlington, VT: Ashgate.

Interested readers are directed to a more comprehensive treatment in an upcoming United States Air Force School of Aerospace Medicine (USAFSAM) technical report, *The Development and Inter-rater Reliability of DoD HFACS, Version 7.0* (King, Strongin, Lawson, & Kuhlmann, In Press).