2018

An Attribute Agreement Method for HFACS Inter-Rater Reliability Assessment

Teddy Steven Cotter
Old Dominion University

Veysel Yesilbas
Old Dominion University

Original Publication Citation

Cotter, T. S., & Yesilbas, V. (2018). An attribute agreement analysis method for HFACS inter-rater reliability assessment. In E-H. Ng, B. Nepal, E. Schott, & H. Keathley (Eds.), *Proceedings of the American Society for Engineering Management 2018 International Annual Conference* (pp. 1-13). American Society for Engineering Management (ASEM).

Proceedings of the American Society for Engineering Management 2018 International Annual Conference
E-H. Ng, B. Nepal, E. Schott, and H. Keathley eds.

AN ATTRIBUTE AGREEMENT ANALYSIS METHOD FOR HFACS INTER-RATER RELIABILITY ASSESSMENT

T. Steven Cotter
Old Dominion University
tcotter@odu.edu

Veysel Yesilbas, Ph.D.
Vyesi001@odu.edu

Abstract
Inter-rater reliability can be regarded as the degree of agreement among raters on a given item or a circumstance. Multiple approaches have been taken to estimate and improve inter-rater reliability of the United States Department of Defense Human Factors Analysis and Classification System used by trained accident investigators. In this study, three trained instructor pilots used the DoD-HFACS to classify 347 U.S. Air Force Accident Investigation Board (AIB) Class-A reports between the years of 2000 and 2013. The overall method consisted of four steps: (1) train on HFACS definitions, (2) verify rating reliability, (3) rate HFACS reports, and (4) random sample to validate ratings reliability. Attribute agreement analysis was used as the method to assess inter-rater reliability. In the final training verification round, within appraiser agreement ranged 85.28% to 93.25%, each appraiser versus the standard ranged 77.91% to 82.82%, between appraisers 72.39%, and all appraisers versus the standard was 67.48%. Corresponding agreement for the random sample of HFACS rated summaries were within appraiser 78.89% to 92.78% and between appraisers 53.33%, which is consistent with prior studies. This pilot study indicates that the train-verify-rate-validate attribute agreement analysis approach has the potential to aid in improving HFACS ratings reliability and contributing to accurately capturing human factors contributions to aircraft mishaps. Additional full-scale studies will be required to verify and fully develop the proposed methodology.

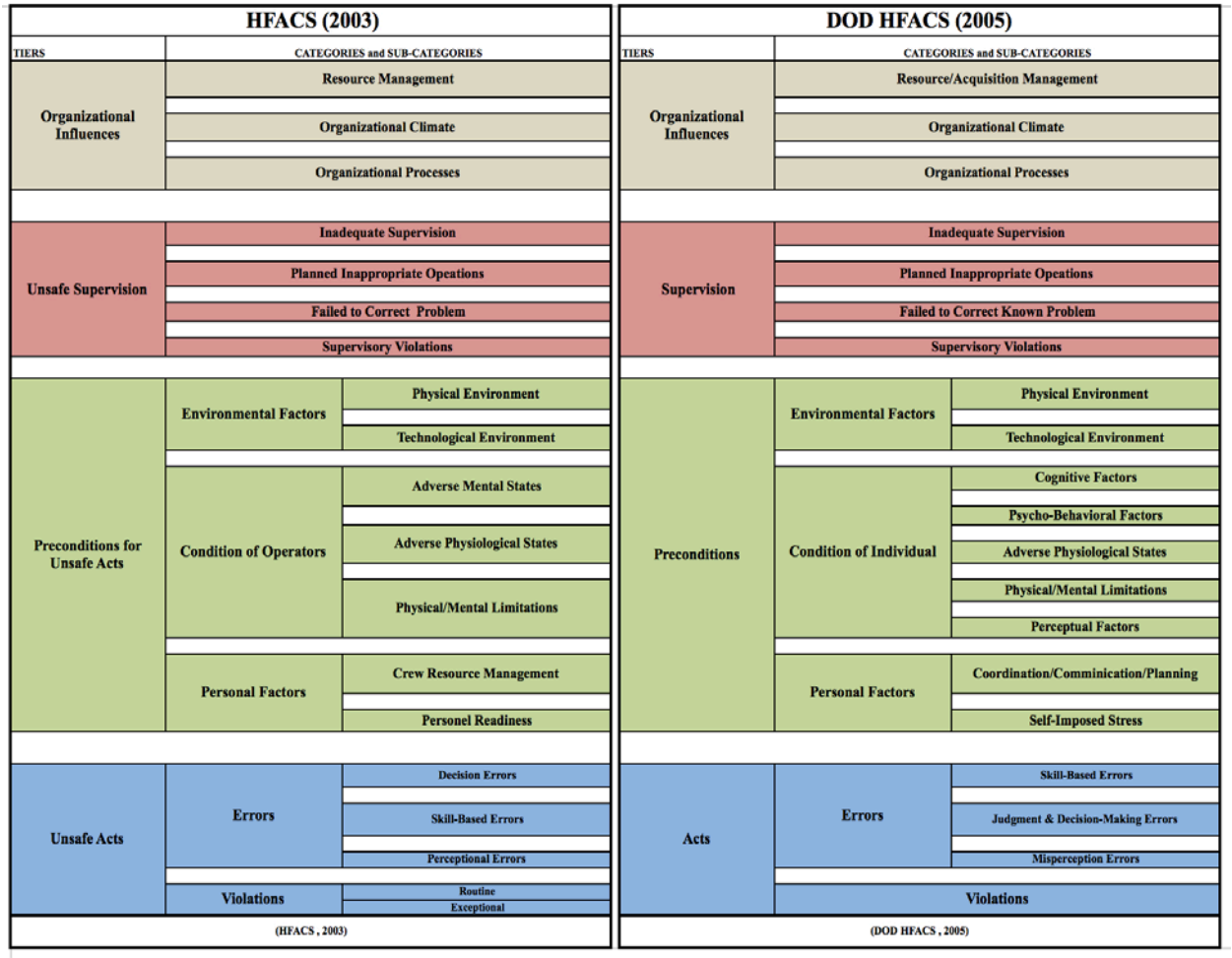
Keywords
Accident Investigation, HFACS, Inner-rater Reliability

Introduction
Reason (1990) Accident Causation Model, also known as the Swiss Cheese Model, is a theoretical model that seeks to explain how accidents manifest across organizational levels. The model's main assumption is that accidents occur in such a way that the causes have relationships across organizational levels. A second assumption is that, at minimum, organizational levels need to function together to prevent accidents. From these assumptions, Reason theorizes that most accidents can be traced to active and latent human failures resulting from prior latent human failures at higher organizational levels. The Human Factors Analysis and Classification System (HFACS), originally adapted from Reason's model to aviation by Wiegmann and Shappell (2003), identifies four tier levels within an organization at which human errors can occur: Organizational Influences, Unsafe Supervision, Preconditions for Unsafe Acts, and Unsafe Acts. The HFACS has been used by the United States Department of Defense (DoD) since 2005, (DOD, 2005) as DOD HFACS with some changes especially at the levels of Preconditions for Unsafe Acts and Unsafe Acts. The DOD HFACS (2005) is composed of the 4 main tier, 14 sub-categories (referred to as category in the Wiegmann and Shappell study), and 147 nanocodes for detailed classification of organizational human errors contributing to aircraft accidents. .

There have been many studies toward the development of accident causation models and frameworks due to desire for decreasing human errors in aviation accidents that result in fatalities and cost a great amount of resources in terms of investigation time, loss of aircraft assets, and litigation. (Yesilbas & Cotter, 2014) Among these studies, no particular or a notable method has been found for evaluation or testing the HFACS taxonomy for validity and rater reliability. Given the HFACS's central role in classifying human errors that contribute to aviation accidents, its validity and the raters' reliability constitutes a substantial function that empower to comprehend the real cause of the

accidents and thus enable proper intervention strategies in terms of organizational safety. For the HFACS to be a useful tool in tracking human causes of aircraft mishaps and providing feedback on the effectiveness of corrective measures, investigators must be able to rate the reports accurately and reliably. Figure 1 depicts the HFACS (2003) and its adopted form DOD HFACS (2005) including tiers, categories and related sub-categories. This study used the DOD HFACS (2005) form as it was the latest version when the research had started.

Exhibit 1. Schematic comparison of HFACS (Shappell & Wiegmann, 2003) and DOD HFACS (DOD, 2005)
(Each of the boxes breakdown to respected nanocodes of human error)



A recent review by Cohen, Wiegmann, and Shappell (2015) examined 111 HFACS manuscripts of which 14 peer-reviewed articles reported rater reliability of HFACS. Notably however: only six of these 14 articles were specifically designed to test reliability. Among the six studies that were designed to assess inter- and/ or intra-rater reliability, three of them demonstrated acceptable levels of reliability and the others showed less reliability. Another recent study by Ergai et al. (2016), examined the inter- and intra-rater reliability of the HFACS data classification process. Results revealed the HFACS taxonomy to be reliable in terms of inter- and intra-rater reliability, with the latter producing slightly higher Alpha values. The study by Ergai et al. assessed the reliability of the HFACS framework as a general accident analysis tool using a large number of trained coders and multiple real-world accident causal factors from a variety of industries. According to Cohen et al. (2015) the reason for these variable results were inadequate training of coders, the use of a small number of accident cases/causal factors, unnecessary modifications made to the HFACS framework, and an inconsistency in the methods used to assess both inter- and intra-rater reliabilities. Another but also substantial reason for this is the lack of a particular method for inter- and intra-rater reliability that needs to include training, testing and evaluation processes. Such a comprehensive methodology can

increase the reliability and establish a desired level of consensus between the inter-rater reliability studies of HFACS that can be compared and improved.

The studies for rater reliability of HFACS depict the need for utilizing HFACS in a more reliable way for further studies in aviation and a variety of other areas as well. This paper reports a pilot study in applying attribute agreement analysis as the measurement tool used to provide feedback during the training of aviators and their subsequent validation in the classification of aircraft mishaps within the HFACS framework.

Methods

Attribute agreement analysis is used to evaluate agreement of assignment to nominal categories or ordinal ratings by multiple appraisers and to determine how likely the appraiser measurement system is to misclassify a part. Attribute agreement analysis provides information on:

- The proportion of cases that each appraiser agrees with himself or herself over all trials (precision).
- The proportion of cases in which each appraiser agrees with a known standard over all trials (individual bias).
- The proportion of cases in which all appraisers agree with themselves (within appraiser repeatability) and with other appraisers (between appraisers reproducibility) over all trials.
- The proportion of cases that all appraisers agree with themselves, with other appraisers, and with the standard over all trials (joint appraiser bias).

There are two primary assessments of attribute agreement:

- The proportion of cases in which the appraisers agree with the standard.
- The proportion of cases in which the appraisers agree with the standard adjusted for the proportion of agreement by chance (various kappa statistics).

Substantial agreement among the raters indicates rating accuracy and reliability. Accuracy is measured individually as the proportion of cases in which each appraiser agrees with the standard and systemically as the proportion of cases in which all appraisers agree with the standard. Inter-rater reliability is regarded as the degree of agreement among raters on a given item or a circumstance. When performing the actual attribute agreement analysis with ordinal data, in addition to the percentage of agreement, Fleiss' kappa statistics and corresponding p-values (reliability estimates) can be used to jointly support the argument of meeting stated minimum agreement requirements. Fleiss' kappa statistic (1971) measures how well the appraisers agree for each standard response with 1.0 indicating perfect agreement, 0 indicating agreement not different from pure chance, and < 0 indicating weaker than expected by pure chance.

The data for this study came from United States Air Force Legal Operations Agency web site (USAF Accident Investigation Boards, 2012). The United States Air Force Legal Operations Agency web site database presents summary and detailed accident reports based on the investigation findings including human factors. This database contained a list of Class A aerospace and ground mishaps and their corresponding summaries and full narratives from the Accident Investigation Board (AIB) of USAF reports between the years of 2000 and 2013. These accidents involved aircraft, remotely piloted aircraft, space systems, and missiles. The majority of the reports in the database include only the executive summaries of the accidents, which may be due to the information being classified and not intended to be shared with the public. An accident report is listed on this site after approval of the USAF Accident Investigation Board. Class A accident reports were used, because they present the most comprehensive information and are prepared with a high level of expertise. This study acquired manned and unmanned aircraft HFACS accident classification data from 347 reports of which 75 detailed accident reports were available for the years between 2010 and 2013. **Exhibit 2** summarizes the assignment scheme by HFACS main and category levels and the code assignments to each category.

Rater Training and Reliability Analysis Design

This research's objective was to test the effectiveness of applying attribute agreement analysis as a feedback tool during training on applying HFACS nanocodes to aircraft incidents and as a measurement tool of inter-rater reliability upon completion of training. The general research method was:

1. Train pilot participants on a random sample of the USAF detailed accident reports.
2. Have the participants rate a random sample of detailed accident reports based only on the summaries and measure inter-rater reliability relative to the nanocodes assigned originally by the expert investigators using attribute agreement analysis.
3. Repeat step 1 and 2 until the participant raters reach a pre-specified agreement rate.
4. Have one trainee assign HFACS nanocodes to the USAF accident summary reports.

Have all pilot participants rate a random sample of the assigned USAF accident summary reports from step 4 and determine the inter-rater agreement rates by attribute agreement analysis.

Exhibit 2. Levels, Categories, Respective Number Of Nanocodes And Abbreviations Used In The Analysis

LEVELS	CATEGORY	CODE	Number of HFACS Nanocodes
Organizational Influences (O)	Resource/Acquisition Management	OR	9
	Organizational Climate	OC	5
	Organizational Processes	OP	6
	Dummy Variable	OD	1
Total Number of Nanocodes in Organizational Influences			**20+1**
Unsafe Supervision (S)	Inadequate Supervision	SI	6
	Planned Inappropriate Operations	SP	7
	Failed to Correct Known Problem	SF	2
	Supervisory Violations	SV	4
	Dummy Variable	SD	1
Total Number of Nanocodes in Unsafe Supervision			**19+1**
Preconditions for Unsafe Acts (P)	Environmental Factors	PE	19
	Condition of Individuals	PC	55
	Personal Factors	PP	18
	Dummy Variable	PD	1
Total Number of Nanocodes in Preconditions for Unsafe Acts			**92+1**
Acts (A)	Skill-Based Errors	AE1	6
	Judgment & Decision-Making Errors	AE2	6
	Misperception Errors	AE3	1
	Violations	AV	3
	Dummy Variable	AD	1
Total Number of Nanocodes in Acts			**16+1**
Total number of DOD HFACS Nanocodes and Dummy Variables			**147+4**

Using this research method, the fundamental sampling question addressed was the accuracy and repeatability with which the one participant classified each of the remaining 272 accident summaries within the HFACS system relative to the known classification by the panels of "experts" in the 75 detailed accident reports.
The sampling design to establish and verify rater reliability was as follows:

1. The sample of $n = 30$ detailed accident reports were randomly selected from the population of N = 75 detailed reports by the independent researcher. The remaining 45 detailed reports were randomly assigned to two categories: 10 to training and 20 to testing by the independent researcher with 15 reports unassigned.
2. One pilot participant was assigned as the test subject and the remaining two were assigned as reference subjects.
3. The pilot participants jointly established classification criteria from the 10 training detailed accident reports. Initial training included the joint review of the DOD HFACS taxonomy including some sample detailed accident reports. Next, the pilot participants reviewed the 10 training detailed accident reports jointly. While some reports included "causal", "contributory", "non-contributory" classification, most of the detailed reports provided all relative causes with respective nanocode(s). As the executive summaries of the reports did not include the "non-contributory" factors, it would not be possible to infer any cause. To this end, the raters decided to classify the all human errors found as causal factors without making any further sorting as "causal" or "contributory." The presence of any cause was assigned an HFACS nanocode within its respective category. For the reports in which a nanocode was not assigned within a category a letter D was entered to the respective level as assignment to the category's dummy variable.
4. The pilot participants independently classified accident causes from the summaries of 10 testing accident reports in accordance with the established HFACS nanocodes classification criteria in two randomly ordered replicates. The independent researcher classified the assigned nanocodes into the HFACS category codes specified in **Exhibit 2**.
5. Attribute agreement analysis was conducted on the classifications. If the measurement metrics Each Appraiser versus Expert Standard > 50%, All Appraisers versus Expert Standard > 50%, and Between Appraiser agreement > 50%, the test subject pilot participant would proceed to Step 6. If any one of the measurement metrics < 50%, the remaining 45 detailed reports would be randomly re-assigned to two categories: 10 to training and 20 to testing. Step 3 would be repeated updating the joint classification criteria to include new information. Step 4 would be repeated on the new set of 10 testing reports. Attribute agreement analysis in this step would be conducted evaluating for all measurement metrics > 50%.
6. The pilot participants independently classified accident causes of the summaries of the n = 30 detailed accident reports in accordance with the established HFACS classification criteria in two randomly ordered replicates. The independent researcher classified the assigned nanocodes into the HFACS category codes specified in **Exhibit 2**. Attribute agreement analysis was conducted evaluating for Each Appraiser versus Expert Standard > 50%, All Appraisers versus Expert Standard > 50%, and Between Appraiser agreement > 50%. If this set of criteria was not met, the process would return to Step 1 and the remaining 45 detailed reports would be randomly re-assigned to two categories: 10 to training and 20 to testing. Steps 3 to 6 were iterated until the set of criteria was met. Once all agreement criteria were met, the test subject pilot participant proceeded to classification in Step 7.
7. The test subject pilot participant classified accident causes of the remaining 272 summary reports in accordance with established HFACS criteria.
8. Upon completion of the classification, a random sample of n = 30 was selected from the 272 summary reports classified by the test subject pilot participant. Using the established classification criteria, the n = 30 summary reports were submitted in random order to the test subject pilot participant for re-classification and to the two reference pilot participants who independently classified accident causes in accordance with the established HFACS classification criteria in two randomly ordered replicates. The independent researcher classified the assigned nanocodes into the HFACS category codes specified in **Exhibit 2**. Attribute agreement analysis was conducted and meeting the set of criteria in Step 6 indicated acceptable training and classification capacity of the test subject pilot participant.

This sampling design modeled training a new accident investigation rater and validating his or her reliability against that of at least two "certified" investigators. Minitab® (version 16.2.1) statistical analysis software was used for all attribute agreement analyses.

Since human subject information was not part of the crash data and the pilot participants provided rating information only about the crash data and did not include any human subject data about themselves, the study was judged to be exempt from review by the Old Dominion University Institutional Review Board (IRB). For attribute

agreement analysis, an "N" code was added to represent disagreement by a researcher with himself between replicates or with other raters. For example, at the Organizational Influences level a code sequence of

USAF	R1-1	R1-2	R2-1	R2-2	R3-1	R3-2
OR	OR	N	N	N	OR	N
OR	N	OP	OP	OP	N	N
OR	N	N	N	N	N	OD

for a given accident summary indicated that the USAF investigators assigned the Organization level cause as OR = Organizational Resource Management, rater R1 assigned OR on the first replicate but changed to OP = Organizational Process on the second replicate, rater R2 assigned OP on replicate one and two, and rate R3 assigned OR on the first replicate and OD = Dummy or no assignment on the second replicate.

Results

Round One Attribute Agreement Analysis

At the DOD HFACS category level, the preliminary percentage of agreement results of round one classification of the summaries of 10 testing accident reports in step 4 above showed acceptable Within Appraisers repeatability of 96.15%, 82.69%, and 73.08% respectively and acceptable between appraisers agreement of 53.85%. However, for Each Appraiser versus Standard, raters one and two exhibited acceptable agreement at 71.15% and 63.46% respectively. Rater three agreed with the standard only 46.15%, which was less than the specified 50% average. After these results, the raters reviewed the same accident reports to identify the differences in code assignments, agree on the correct assignment per report, and the criteria for each assignment. The results of Round One analysis are presented in **Exhibit 3**. Two factors were identified as the causes for this low level of agreement. First, it was the initial part of independent study, and the raters did not think that they had sufficient understanding of the HFACS classification code definitions. Second, they thought that including as many nanocodes as possible would contribute in finding the causes of the accidents. However, it appears that including more nanocodes than required decreased the level of agreement.

Round Two Attribute Agreement Analysis

The pilot participants performed round two attribute agreement analysis on an additional 10 randomly selected accident summaries classifying two replicates with approximately a one-week interval between replicates. The Assessment Agreement results of Round Two are shown in Exhibit 3. The Within Appraisers, Each Appraiser versus Standard, Between Appraisers, and All Appraisers versus Standard agreement percentages were all above the specified 50% average.

Round Three Attribute Agreement Analysis

The research proceeded to step 6 in which the pilot participants were to independently classify accident causes of the summaries of the $n = 30$ detailed accident in two randomly ordered replicates. However, two detailed accident reports were found to be insufficient quality in the detail of their descriptions to admit them for classification. Rather than replace these two reports with two randomly selected from the 15 remainder of the second group of 45, the researchers decided to discard these reports rather than risk compromising the original random assignment.

The remaining $n = 28$ executive summaries of detailed accident reports were randomly assigned and rated in two replicates by the pilot participants with approximately a one-week interval between replicates. The Assessment Agreement results of Round Three are shown in Exhibit 3. The raters' Within Appraisers, Each Appraiser versus Standard, Between Appraisers, and All Appraisers versus Standard agreement percentages were all above specified 50% average. The results from Round Three were assessed to be sufficient to continue to step 7 in which the test subject pilot participant classified accident causes of the remaining 272 summary reports.

Evaluation of the Remaining Reports

The remaining 272 summary reports with no detailed accident information were classified by the test subject pilot participant in accordance with established HFACS criteria. In accordance with step 8, after all reports were classified, n = 30 executive summaries were randomly selected and classified in two replicates by all participant pilots with approximately a one-week interval between replicates. The Round Four inter-rater attribute agreement analysis results are shown in Exhibit 3. The raters' Within Appraisers and Between Appraisers agreement percentages were all above specified 50% average.

Exhibit 3. Attribute Agreement Analysis of HFACS Category Nanocode Assignments

Assessment Agreement	Appraiser	# Inspected	# Matched	Percent	95 % CI
Round 1					
	Rater1	52	50	96.15	(86.79, 99.53)
Within Appraisers	Rater2	52	43	82.69	(69.67, 91.77)
	Rater3	52	38	73.08	(58.98, 84.43)
	Rater1	52	37	71.15	(56.92, 82.87)
Each Appraiser vs. Std.	Rater2	52	33	63.46	(48.96, 76.38)
	Rater3	53	24	46.15	(32.23, 60.53)
Between Appraisers		52	28	53.85	(39.47, 67.77)
All Appraisers vs. Std.		52	23	44.23	(30.47, 58.67)
Round 2					
	Rater1	57	54	94.74	(85.38, 98.90)
Within Appraisers	Rater2	57	53	92.98	(83.00, 98.05)
	Rater3	57	48	84.21	(72.13, 92.52)
	Rater1	57	50	87.72	(76.32, 94.92)
Each Appraiser vs. Std.	Rater2	57	51	89.47	(78.48, 96.04)
	Rater3	57	47	82.46	(70.09, 91.25)
Between Appraisers		57	44	77.19	(64.16, 87.26)
All Appraisers vs. Standard		57	43	75.44	(62.24, 85.87)
Round 3					
	Rater1	163	144	88.34	(82.40, 92.83)
Within Appraisers	Rater2	163	152	93.25	(88.25, 96.58)
	Rater3	163	139	85.28	(78.89, 90.33)
	Rater1	163	133	81.60	(74.78, 87.22)
Each Appraiser vs. Std.	Rater2	163	135	82.82	(76.14, 88.27)
	Rater3	163	127	77.91	(70.76, 84.03)
Between Appraisers		163	118	72.39	(64.86, 79.10)
All Appraisers vs. Std.		163	110	67.48	(59.72, 74.60)
Round 4					
	Rater1	180	143	79.44	(72.80, 85.09)
Within Appraisers	Rater2	180	167	92.78	(87.97, 96.10)
	Rater3	180	142	78.89	(72.19, 84.61)
Between Appraisers		180	96	53.33	(45.76, 60.79)

Evaluation of the Overall Attribute Agreement Analysis Approach

Individual and joint bias is reported in the Each Appraiser versus Standard and All Appraisers versus Standard agreement. Round Two Each Appraiser versus Standard agreement ranged 82.46% to 89.47%, and the Round Three agreement ranged 77.91% to 82.82%. Correspondingly, the Round Two All Appraisers versus Standard agreement was 75.44%, and the Round Three was 67.48%. Repeatability is reported in the Within Appraisers agreement. Testing Round Three Within Appraisers agreement ranged 85.28% to 93.25%, and Round Four ranged 79.44% to 92.78%. Reproducibility is reported in the Between Appraisers agreement. Round Three was 72.39%, and Round Four was 53.33%.

Exhibit 4 reports Fleiss' kappa statistics for Round Three and Four Between Appraisers item agreement and for Round Three All Appraisers versus Standard item agreement. The HFACS category level, Round Three Between Appraisers item agreement ranged 70.84% for SI to 97.48% for 97.19% for OP. The dummy variable assignments ranged 95.94% for PD to 97.76% for AD. The "N" no assignment was 60.64%, or, given the standard errors, statistically less than the category level and dummy variable assignments. The corresponding Round Four Between Appraisers item agreement ranged 55.35% for AE2 to 88.45% for SI within the HFACS categories and 76.23% for PD to 98.06% for SD within the dummy variable assignments. The "N" no assignment was 47.01%, again statistically less than the category level and dummy variable assignments. Exhibit 4 also reports the Round Three All Appraisers versus Standard Fleiss' kappa statistics. The HFACS category item agreement ranged 81.79% for PP to 98.66% for OP. Category SI appears to have been an outlier at the category level with agreement of 39.28%. The dummy variable assignments ranged 88.52% for SD to 98.84% for AD. The "N" no assignment was 48.89%, again statistically less than the category level and dummy variable assignments.

Exhibit 5 presents the summary of the Attribute Agreement Analysis method applied to establish and evaluate HFACS inter-rating reliability of the raters.

Conclusions

The overall attribute agreement analysis approach was found to be potentially viable as a method for providing feedback during the training of aviators and validating rater reliability in the classification of aircraft mishaps within the HFACS framework. Examination of the Round Two to Round Four percent agreement suggests that forgetting set in rather quickly after training. Within Appraisers average percent agreements from Round Two to Round Four declined from 90.64% to 88.96% to 83.70%. The Each Appraiser versus Standard declined from 86.55% to 80.78% from Round Two to Round Three, and the All Appraisers versus Standard declined from 75.44% to 67.48%. The Between Appraisers average percent agreement declined from 77.19% to 72.39% to 53.33% from Round Two to Round Four, and the Round Four agreement of 53.33% was below the Round Two and Three 95% lower confidence limits of 64.16% and 64.86% respectively, indicating that Round Four Between Appraisers average percent agreement was statistically different from Rounds Two and Three. All Round Two through Four average agreement percentages were above the required minimum 50% agreement established from the literature review.

Forgetting was evident also in the Round Three and Round Four Between Appraisers inter-rater reliability summaries using Fleiss's kappa statistics. The HFACS category average agreement Between Appraisers kappa declined from 86.67% for Round Three to 75.60% for Round Four. Likewise, the dummy variable kappa declined from 96.91% to 88.35%. The Round Three HFACS category average agreement All Appraisers versus Standard was 95.37%, and the dummy variable All Appraisers versus Standard was 94.31%. Thus, reproducibility and bias relative to the standard were relatively equivalent.

Exhibit 4. Inter-rater Reliability Summary Data Using Fleiss's Kappa Statistic**.**

Response	Kappa	Std Err	Z	P(vs> 0)	Response	Kappa	Std Err	Z	P(vs> 0)
Round 3 Between Appraisers					**Round 4 Between Appraisers**				
OR	0.81624	0.0202237	40.361	0.0000	OR	0.827586	0.0192450	43.003	0.0000
OP	0.97191	0.0202237	48.058	0.0000	OP	0.827586	0.0192450	43.003	0.0000
OD	0.97475	0.0202237	48.199	0.0000	OD	0.903898	0.0192450	46.968	0.0000
SI	0.70842	0.0202237	35.030	0.0000	SI	0.884497	0.0192450	45.960	0.0000
SP	0.88680	0.0202237	43.850	0.0000					
					SV	0.664804	0.0192450	34.544	0.0000
SD	0.96451	0.0202237	47.692	0.0000	SD	0.980623	0.0192450	50.955	0.0000
PE	0.87284	0.0202237	43.160	0.0000	PE	0.706681	0.0192450	36.720	0.0000
PC	0.92742	0.0202237	45.858	0.0000	PC	0.789679	0.0192450	41.033	0.0000
PP	0.78096	0.0202237	38.616	0.0000	PP	0.820659	0.0192450	42.643	0.0000
PD	0.95937	0.0202237	47.438	0.0000	PD	0.762322	0.0192450	39.611	0.0000
AE1	0.95325	0.0202237	47.135	0.0000	AE1	0.917719	0.0192450	47.686	0.0000
AE2	0.91236	0.0202237	45.114	0.0000	AE2	0.553529	0.0192450	28.762	0.0000
AE3	0.90699	0.0202237	44.848	0.0000	AE3	0.567702	0.0192450	29.499	0.0000
AV	0.79709	0.0202237	39.414	0.0000					
AD	0.97764	0.0202237	48.342	0.0000	AD	0.887003	0.0192450	46.090	0.0000
N	0.60640	0.0202237	29.985	0.0000	N	0.470165	0.0192450	24.430	0.0000
Overall	0.87944	0.0064759	135.80	0.0000	Overall	0.765485	0.0068589	111.60	0.0000

The main contribution of this study to inter-rater reliability analysis of the assignment of HFACS codes to aircraft accident reports was the application of attribute agreement analysis methodology in the **Methods** section. As noted, given that there are three "Organizational Influences" categories, four "Unsafe Supervision" categories, three "Preconditions for Unsafe Acts" categories, and four "Unsafe Acts" categories plus one dummy variable for each category level, there are 18 categories and $4 \times 5 \times 4 \times 5 = 400$ path classifications for each aircraft accident under the HFACS. This number of path classifications can be multiplied further, since USAF experts assign category codes that create partial paths and multiple paths within the same accident report. Thus, assignment of an accident report to a discrete path classification is not always possible. The attribute agreement analysis inter-rater reliability method developed as part of this work overcame this need for discrete path classification by:

1. Treating each HFACS categorical level as an independent assignment. This decomposed each path by Reason's Swiss Cheese model to four independent classification problems.
2. Adding a dummy variable to each HFACS categorical level as a pass-through category for accidents in which USAF investigators did not make code assignment for the given level.
3. Normalizing the data into a Poisson counting process within a respective category.

These modifications allowed each path to be treated as arising from a multiplicative process of independent variables.

Exhibit 4 (continued)**.** Inter-rater Reliability Summary Data Using Fleiss's Kappa Statistic**.**

Response	Kappa	Std Err	Z	P(vs> 0)	Response	Kappa	Std Err	Z	P(vs> 0)
Round 3 All Appraisers vs Standard									
OR	0.88680	0.0319765	27.733	0.0000					
OP	0.98664	0.0319765	30.855	0.0000					
OD	0.98564	0.0319765	30.824	0.0000					
SI	0.39276	0.0319765	12.282	0.0000					
SP	0.84607	0.0319765	26.459	0.0000					
SD	0.88517	0.0319765	27.682	0.0000					
PE	0.91694	0.0319765	28.675	0.0000					
PC	0.94316	0.0319765	29.495	0.0000					
PP	0.81787	0.0319765	25.577	0.0000					
PD	0.91334	0.0319765	28.563	0.0000					
AE1	0.95327	0.0319765	29.811	0.0000					
AE2	0.84193	0.0319765	26.329	0.0000					
AE3	0.95133	0.0319765	29.751	0.0000					
AD	0.98835	0.0319765	30.908	0.0000					
N	0.48894	0.0319765	15.290	0.0000					
Overall	0.84895	0.0099721	85.133	0.0000					

The inter-rater reliability procedure developed in this work is designed to verify the individual rater's reliability before and after training and continuing into practice. The first step was to establish the measurement standard for acceptable inter-rater agreement. To this end, this work relied on a prior study by O'Connor, et al. (2010) indicating only a 55% agreement among raters of aircraft accident reports. This study set the standard for between rater agreement and all raters' agreement to experts' classification at greater than or equal to 50% average or 50/50 odds of random assignment classification. A similar procedure can be applied in actual practice to continually identify and revise the state of "certified" accident investigators' joint inter-rater capability.

The second contribution was the tradeoff analysis between confidence in the difference to detect and the sampling resolution over a range of sample sizes to select a sample size that provided ≥ 90% confidence in detecting differences between any two raters from the $p = 0.50$ base random assignment case. Again, the difference to detect and sampling resolution can, and should be, incorporated into certifying accident investigators' joint inter-rater capability.

The third contribution was the development of the eight-step rater reliability method in the Methods section. The pre-classification and inter-rater testing attribute agreement analyses can continue for multiple rounds until the between raters and all raters' agreement to experts' classification achieve some agreed upon average agreement rating standard. The pre-classification and inter-rater testing attribute agreement analysis establishes a trainee rater's reliability *a priori* to rating a body of accident summaries. After rating the accident summaries, the post inter-rater testing by classifying a random sample of the trainee classified summaries, in this study the 30 random samples from the 272 classified accident summaries, by the trainee and reference raters should demonstrate inter-rater reliability greater than or equal to the established standard for "certification" to be awarded. Procedures should be developed to periodically re-certify accident investigators inter-reliability agreement as continuing to meet the established standard.

Exhibit 5. Summary of the Attribute Agreement Analysis method for inter rater reliability.

PHASE		**ACTIONS**	**NUMBER of the REPORTS**	**Standard Used**
TRAINING		Didactic training, discussing and reviewing detailed reports	10 reports from 75 detailed HFACS coded reports	Already coded by AIB panel experts in the detailed reports
TESTING	Round1	2 replicates with 1-week interval by 3 pilots	10 reports from 75 detailed HFACS coded reports	
	Round2	2 replicates with 1-week interval by 3 pilots	10 reports from 75 detailed HFACS coded reports	
	Round3	2 replicates with 1-week interval by 3 pilots	30 reports from 75 detailed HFACS coded reports (2 of the 30 reports were insufficient in data)	
EVALUATION	Round4	Rating by the test subject pilot	272 reports – executive summaries- which don't have the detailed HFACS coding	HFACS taxonomy
		2 replicates with 1-week interval by 3 pilots	Random 30 reports from 272 reports without coding	Test Subject pilot's codings

Finally, the categorical level classification scheme developed for this work transformed the HFACS classification data into a format suitable for attribute agreement analysis. Each categorical level was assigned multiple rows, one for each category assigned within the level. This allowed for multiple category assignments within a category level. In addition to the category codes and the dummy variable code, a code of "N" was assigned to show disagreement between raters within a categorical level or of a rater with himself between replicates.

Future research is planned to extend this attribute agreement analysis approach to HFACS inter-rater reliability at the nanocode level and then to validate the approach through research using USAF accident investigators and pilot trainees. Originally, the researchers gathered nanocode level causal assignment data and tried to apply attribute agreement analysis at the nanocode level. Attribute agreement analysis using nanocodes was possible at the $n = 10$ accident reports used in the Round One and Two training phase. However, the number of nanocode categories exceeded the allowed 50 categories in Minitab for the Round Three and Four $n = 28$ and $n = 30$ required to maintain the 90% confidence to detect at least a 10% difference from the minimum average agreement of 50% used in this study. Extension of this methodology to the USAF accident investigators and pilot trainees will require securing support from the United States Air Force.

The distinguishing feature of this study is that the attribute agreement analysis is used for the inter-rater reliability of DOD HFACS coding by including training, testing and evaluation of the rater. Thus, the study provided that the subject test pilot can rate the DOD HFACS based accidents.

The study by Ergai et al. (2016) suggests that training might also incorporate the use of a classification tool or flowchart, with the intention of increasing the reliability of HFACS for future training. In fact, a well-established nanocode in the taxonomy under the causal categories can be used instead of a flowchart; the raters use the nanocodes as an explanatory feature for the accident reports. Using only the causal levels for accident classification may not be sufficient to rate the causes precisely. In this study, the raters used the nanocodes under the categories in DOD HFACS that helped them to rate the causes of the accidents better.

Recommendations

The inter-rater reliability study methodology developed in this work can be conducted to establish and improve assessment reliability for any aviation organization applying the HFACS directly or any organization in another sector

adapting the HFACS system to its sector. Other sectors will have to develop their own respective accident categorical level classification schemes and adapt the methodology for assessment and possibly certification of raters.

References

Cohen, T. N., Wiegmann, D. A., & Shappell, S. A. (2015). Evaluating the Reliability of the Human Factors Analysis and Classification System. *Aerospace Medicine and Human Performance*, 86(8), 728-735. doi:10.3357/AMHP.4218.2015

DOD. (2005). *Department of Defense Human Factors Analysis and Classification System*.

Ergai, A., Cohen, T., Sharp, J., Wiegmann, D., Gramopadhye, A., & Shappell, S. (2016). Assessment of the Human Factors Analysis and Classification System (HFACS): Intra-rater and inter-rater reliability. *Safety Science*, 82, 393-398. doi:http://dx.doi.org/10.1016/j.ssci.2015.09.028

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76, 378-382. doi:10.1037/h0031619

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

Wiegmann, D. A., & Shappell, S. A. (2003). *A human error approach to aviation accident analysis : the human factors analysis and classification system.* Aldershot, Hants, England ; Burlington, VT: Ashgate.

Yesilbas, V., & Cotter, T. S. (2014). STRUCTURAL ANALYSIS OF HFACS IN UNMANNED AND MANNED AIR VEHICLES. *Proceedings of the American Society for Engineering Management 2014 International Annual Conference*, Virginia Beach, VA.

Acknowledgment

The views expressed in this article are those of the authors and do not reflect the official policy or position of any nation's Armed Forces, Department of Defense, or Government.

T. Steven Cotter is a Lecturer with the Engineering Management and Systems Engineering department at Old Dominion University. He earned a Ph.D. in Engineering Management and Systems Engineering from Old Dominion University, a Master of Science in Engineering Management with a concentration in quality/reliability engineering from the University of Massachusetts at Amherst, a Master of Business Administration with a concentration in finance and a Bachelor of Science both from the University of South Carolina, and a diploma in Electronic Technology from Graff Area Vocational and Technical School (now Ozarks Technical Community College). He is a certified Quality Engineer and Reliability Engineer with the American Society for Quality. His research interests are in engineering design analytics, human-intelligence/machine-intelligence decision governance, human-intelligence/machine-intelligence quality systems design, and systems statistical engineering.

Veysel Yesilbas is an independent academic researcher and software tester. He earned a Ph.D. in Engineering Management and Systems Engineering from Old Dominion University, a Master of Science in Security with a concentration in future technological improvements from the Turkish Air War College, and a Bachelor of Science in Electronic Engineering from the Turkish Air Force Academy at Istanbul, Turkey. His research interests are human factors in aviation accidents with a specialty at Human Factors Analyses and Classification System.