Applied Ergonomics 44 (2013) 261–273

Contents lists available at SciVerse ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo

Predictive models of safety based on audit findings: Part 1: Model development and reliability

Yu-Lin Hsiao [a,*], Colin Drury [b], Changxu Wu [b], Victor Paquet [b]

[a] Department of Industrial and Systems Engineering, Chung Yuan Christian University, Chung Li 32023, Taiwan
[b] Department of Industrial and Systems Engineering, State University of New York (SUNY) at Buffalo, Buffalo, NY 14260, USA

ARTICLE INFO

Article history:
Received 23 September 2011
Accepted 16 July 2012

Keywords:
Safety audit
Human error
Aviation maintenance
HFACS
Measurement reliability

ABSTRACT

This consecutive study was aimed at the quantitative validation of safety audit tools as predictors of safety performance, as we were unable to find prior studies that tested audit validity against safety outcomes. An aviation maintenance domain was chosen for this work as both audits and safety outcomes are currently prescribed and regulated. In Part 1, we developed a Human Factors/Ergonomics classification framework based on HFACS model (Shappell and Wiegmann, 2001a,b), for the human errors detected by audits, because merely counting audit findings did not predict future safety. The framework was tested for measurement reliability using four participants, two of whom classified errors on 1238 audit reports. Kappa values leveled out after about 200 audits at between 0.5 and 0.8 for different tiers of errors categories. This showed sufficient reliability to proceed with prediction validity testing in Part 2.
© 2012 Elsevier Ltd and The Ergonomics Society. All rights reserved.

1. Introduction

Safety audit of work systems has become an important way to measure the potential for system errors without waiting for the consequences of these errors to manifest themselves. Such proactive tools as safety audits could have much value if validated against safety performance. In a survey of Human Factors/Ergonomics (HF/E) audit tools (Drury and Dempsey, 2012), both reliability and validity of many audit tools were assessed. Often no reliability or validity measures were given for these audit tools: where either was measured, inter-rater reliability was assessed more often than validity. Where validity was measured (Koli et al., 1998), this typically used expert opinion as the validation criterion, e.g. the Koli et al.'s ERgoNomics Audit Program (ERNAP) integrates a variety of ergonomic audit tools to carry out an ergonomic evaluation of maintenance and inspection operations. It was validated against six HF/E professionals viewing videos of jobs which were audited by ERNAP. We can find no validations of audit tools, specifically those meant to evaluate HF/E concerns or safety, against future safety performance, i.e. prediction validity.

The study reports here as Part 1 and Part 2 uses data from an existing audit system in the domain of civil aviation maintenance as the basis for a prediction model of future safety performance. Independent data sets of both audit records and maintenance safety performance were supplied by a civil aviation regulatory authority. Part 1 presents the justification for using aviation maintenance as a domain, the development of the model that derived and classified human errors found in the audit records, and the measurement reliability study necessary to future testing of model validity. Part 2 performs the validation of the model by predicting future safety performance from past audit findings.

1.1. The aviation maintenance domain

According to the International Air Transportation Association (IATA, 2003), about 70 percent of the root causes of flight accidents are derived from issues of human factors, and maintenance factors started the accident chain in 24 accidents over a total of 92 (26%). Human error is natural, especially in the complicated environment of airliner maintenance. Over the past 20 years, the aviation industry has established many different systems and procedures to ensure maintenance safety despite human error, e.g. reviewed by Gramopadhye and Drury (2000). On-Job-Training (OJT) programs, Quality Assurance (QA) programs, and Standard Operating Procedures (SOPs) have all become obligatory requirements for any airline. From the perspective of a civil aviation regulatory authority, it is important to ensure the correct implementation of these systems. Consequently, safety audit has now become one vital and proactive method for regulators to detect potential failures in aircraft maintenance system.

* Corresponding author.
E-mail address: yhsiao@cycu.edu.tw (Y.-L. Hsiao).

0003-6870/$ – see front matter © 2012 Elsevier Ltd and The Ergonomics Society. All rights reserved.
http://dx.doi.org/10.1016/j.apergo.2012.07.010

It is assumed by the industry that proactive techniques such as audits would contribute to the risk mitigation of erroneous activities and eventually benefit accident or incident prevention. However, a causal relationship between proactive measures and safety performance such as accident/incident rates must be shown to be valid to support the above assumption. Amazingly, this question of validity does not appear to have been answered for aviation maintenance field or indeed any other system where proactive measures are in use.

If aviation maintenance accidents are a legitimate cause of concern, and we are trying to validate safety audit as a predictor of accident/incident in this domain, then the simplest procedure would be to correlate audit findings with accident/incident reports. This simple model was attempted by the first author for an aviation regulatory authority by correlating the number of significant audit findings from prior months with the incident rates for each month. However, no correlations were statistically significant. Some other variable may intervene between audit findings and safety performance, which might have impeded the direct connection between them, so that the predictive validity of safety audits in the aviation industry remains unidentified. Therefore, the current study started by postulating existing human errors in audit findings as the predictors rather than just counting amounts of audit finding, and the purpose is to examine the prediction validity of safety audits from an HF/E perspective.

It is the existence of on-going safety audit procedures that prompts their examination as possible predictors of future safety performance. A positive finding would allow airline, or third-party repair organizations (Drury et al., 2010) to focus interventions on future safety predictions using already-collected data. To find a more rational basis for turning the raw audit findings into potential predictors, we developed a model of human error specific to aviation maintenance, but readily usable in other domains. There were no items in our classification model that could apply only to aviation maintenance, although we did not explicitly consider this expandability in our error classification development.

1.2. Human factors analysis and classification system − maintenance audit (HFACS-MA)

The concepts of human error have not been comprehensively considered and adopted in recent audit systems. Although most of the issues found by audit reports could be attributed to some varieties of human error, the detailed study of root cause or error association is impractical without an appropriate classification scheme for human error. The goals of active human error detection and prevention can only be partially impacted by current simple descriptive statistical analysis of audit records.

Reason (1990) proposed that "accuracy of error prediction depends very largely on the extent to which the factors giving rise to the errors are understood". Expanding on his idea, it is necessary to develop a practical human error classification framework for a maintenance audit system to analyze the failures found in daily reports, and to accordingly assist understanding the status of human failures. In this study, we integrated the concepts of human error and safety models to develop a comprehensive framework for maintenance audit.

The demands of maintenance audit systems should include a specific designed analysis tool, a comprehensive taxonomy of human error related to maintenance activities, and a clear distinction between different error categories to illustrate the interrelationship between different failures. Except for MEDA (Rankin, 2000), the literature cites in Table 1 either represents conceptual models: SHELL model (ICAO, 1998) and PEAR (Johnson, 2001), or are designed as accident investigation and analysis tools: Reason model (Reason, 1990), HFACS (Shappell and Wiegmann, 2001a,b), HFACS-ME (Schmidt et al., 2000) and Wheel of Misfortune (O'Hare, 2000). Though the design purpose of MEDA meets the demands of an aviation maintenance audit, its contributing factors lack hierarchical classification to further distinguish causal relationships between factors and detailed explanation to cover their scopes. Regarding to the comprehensiveness of human error, HFACS comprises the most detailed depiction of error categories, but pays more attention to flight operations. MEDA and all other models or tools appear similarly lacking in management or organizational factors.

Table 1
Purpose and shortcomings of human error taxonomies to meet the demands of aviation maintenance audit systems.

	Design purpose	Shortcoming
Reason model	Accident investigation and analysis tool	Detailed explanation of each error category
HFACS	Accident investigation and analysis tool	Focused on flight operation field
HFACS-ME	Accident investigation and analysis tool	Organizational factors
Wheel of misfortune	Accident investigation and analysis tool	Management factors
SHELL	Conceptual model	Management and organizational factors
PEAR	Conceptual model	Management factors
MEDA	Incident/event investigation and analysis tool	Distinction of the contributing factors Details of management and organizational factors

Therefore, aviation maintenance audit systems need a human error classification framework to practically analyze audit reports. This need directly drove the design of our taxonomy. From the perspective of audit activities of an aviation authority, we assumed the maintenance organization (e.g., repair station or airline) to be the complete system to study, i.e. the analytic entity of the classification framework.

The fundamental construct of our framework is based on the ideas of HFACS, which is developed for investigating and analyzing the human causes of aviation accidents, both military and civil (Shappell et al., 1999; Shappell and Wiegmann, 2001a,b). HFACS has two noteworthy advantages: first, it provides more details of the management and organizational aspects; second, it presents a comprehensive hierarchical level of human error and a detailed definition of each error condition. The broad utilization of HFACS has demonstrated that it is an effective tool to identify both active and latent failures presented in accidents. Both qualitative and quantitative studies of HFACS also address the latent influences of supervisory and organization on the unsafe acts of first line workers (Shappell and Wiegmann, 2001a,b; Dambier and Hinkelbein, 2006; Li et al., 2007). HFACS splits various human errors into four main failure tiers: Unsafe Act, Precondition for Unsafe Act, Unsafe Supervision, and Organizational Influence (Fig. 1). We retained the same failure tiers of HFACS in our framework. As with other extensions of HFACS (Schmidt et al., 2000; O'Connor, 2008; Olsen and Shorrock, 2010), we named our modified taxonomy as HFACS-Maintenance Audit (HFACS-MA) (see Fig. 2).

In addition to this theoretical perspective, we also conducted informal interviews with experienced inspectors in one civil aviation authority (which has requested anonymity) to discuss and evaluate the applicability of the framework. Accordingly, HFACS-MA integrates other well accepted concepts from the field of human factors, management, and safety culture/climate during the development period. In the following section, we address the differences between HFACS-MA and HFACS.

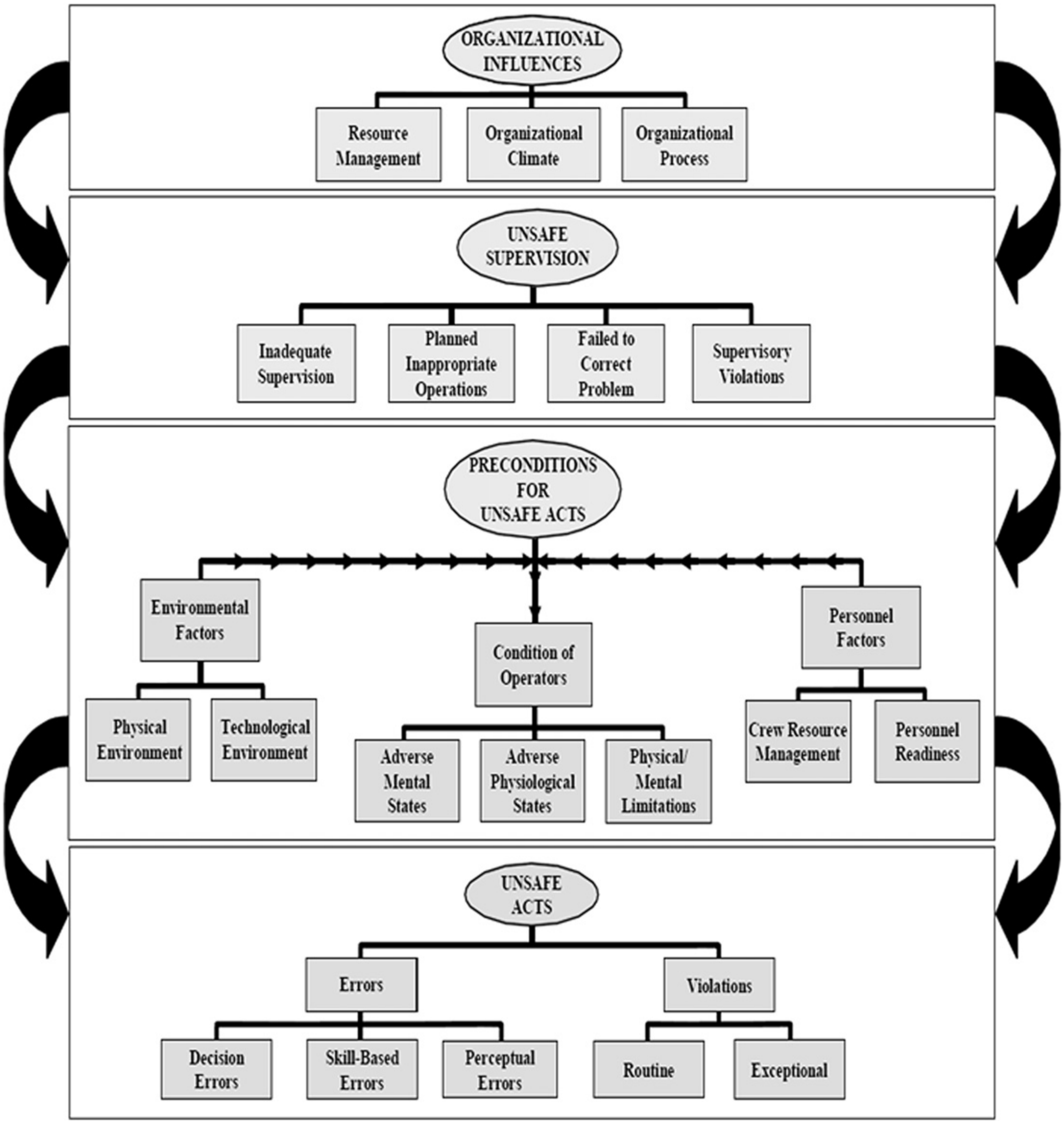


Fig. 1. The HFACS framework (Shappell and Wiegmann, 2004).

1.2.1. *Unsafe act*

HFACS-MA classifies the Unsafe Act tier into Error and Disobedience through personal intention. The definition of Error, and its subordinate failures: Skill-based and Decision Error remain unchanged. The classification of Perceptual Error, listed in HFACS, was combined with Skill-based Error in HFACS-MA because of their similar behavior descriptions and the few occurrences of Perceptual Error in maintenance audit reports.

In HFACS, Perceptual Error represents decisions made based on faulty perceived information when individual sensory input is either degraded or “unusual”. Perceptual Errors such as spatial disorientation or visual illusion are more likely to relate to specific flight operations (pilot) than to the maintenance environment, although there is scope for such errors during maintenance, e.g. when inspecting a large array of structure such as fuselage frames. In maintenance audit reports, we found narratives related to Perceptual Error were usually due to attention degradation or insufficiencies. They were similar to behavior descriptions of Skill-based Error: action due to failures of attention, memory and/or skill deficiency without conscious thought. In our previous interviews, interviewees sometimes faced the dilemma of choosing between Perceptual Error and Skill-based Error because of the similarities of contents and the lack of adequate information to distinguish them in audit records. Moreover, we considered that an integrated error grouping would improve the practicability of quantitative analysis in Part 2 (i.e. the small number of occurrences of Perceptual Error classified from audit reports would increase the difficulty of analysis and examination of predictive validity). Thus, we decided to combine the maintenance-related details of Perceptual Errors within Skill-based Error.

On the other hand, Disobedience is assumed to be intended behavior in contravention of existing proceduresor rules. According to the feedback of the informal interviews with experienced auditors, they actually use “Violation” when something against the law is found and they need to process the legislative enforcement actions. It should also be noted that English is not the native language for the interviewees. In other words, “Violation” has a legally-specific meaning for these government officials. Therefore, the original “Violation” category of HFACS was replaced by “Disobedience” in our study. (Note: in our newest version, “Disobedience” will be replaced by “Noncompliance” to allay misconceptions in the industry. However, its definition remains the same.)

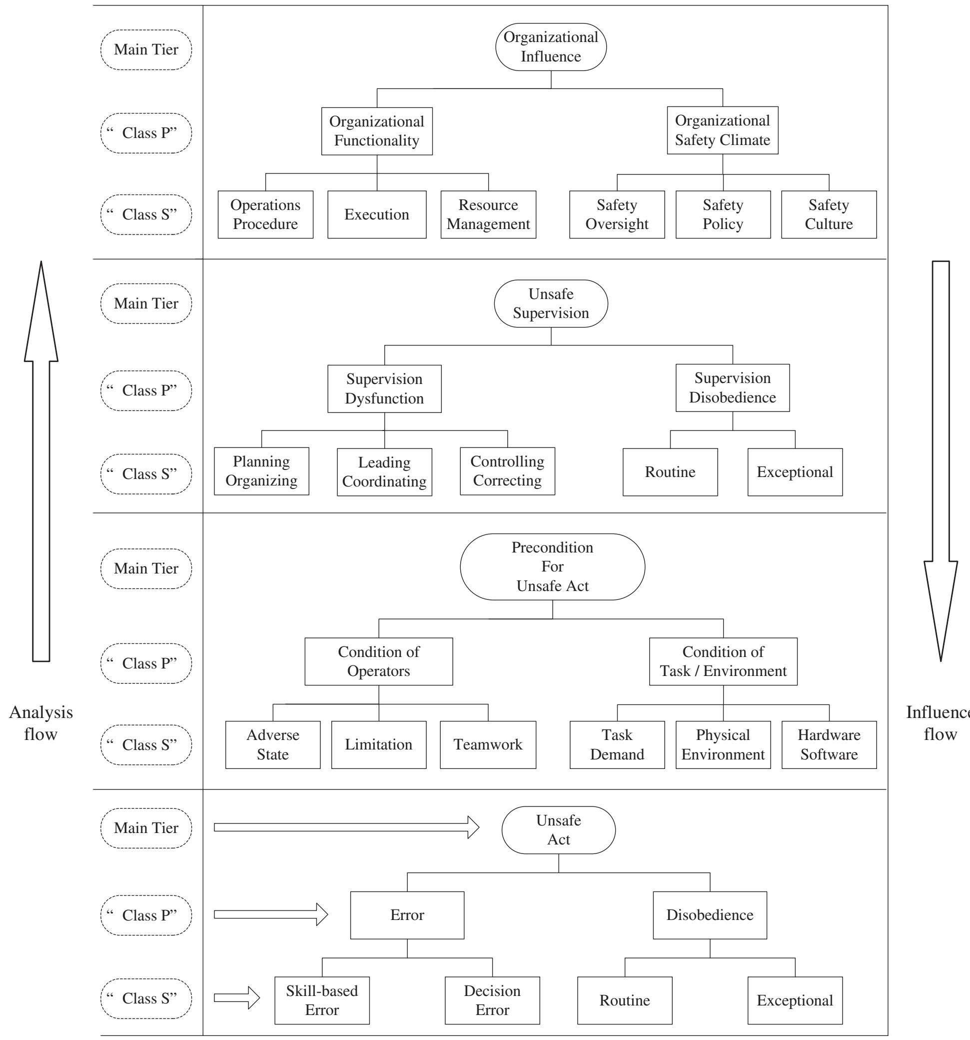


Fig. 2. The complete framework of HFACS-MA.

1.2.2. Precondition of unsafe act

In contrast to HFACS, the Precondition category of HFACS-MA is also informed by SHELL model, Wheel of Misfortune, PEAR, and MEDA. Table 2 shows specifically how each of the classifications in the framework related to the prior taxonomies. (Note that Precondition factors related to management activities are considered separately later.) Compared with HFACS, we classified the latent factors of Precondition into two main categories: Condition of Operators and Condition of Task/Environment.

Table 2
Classification involved in precondition of unsafe acts.

	Operators	Task/environment
HFACS	Personal factors Condition of operators	Environment factors
Wheel of misfortune	Operator resources	Task demands; Interfaces
SHELL	Liveware (L, L–L)	Software (S); Hardware (H); Environment (E)
PEAR	People (P); Action (A)	Action (A); Environment (E); Resource (R)
MEDA	Technical knowledge/skills Individual factors Communication	Information (e.g. work cards, maintenance manuals) Equipment/tools/safety equipment Job/task Environment/facilities

The personal factors and the condition of operators of HFACS were incorporated, and became the Adverse State, Limitation, and Teamwork in HFACS-MA. Adverse State and Limitation consider the effects of physical, mental, and competence status of first-line personnel in either temporary or permanent ways respectively. Teamwork, which is similar to the Crew Resource Management (CRM) or Maintenance Resource Management (MRM) of HFACS, addresses the defects of interactions and cooperation between technicians or the rest of the maintenance team.

On the other hand, the environmental factors of HFACS were further expanded based on the concepts of tasks in the high reliability organizations (HRO) (Roberts, 1990) and the Wheel of Misfortune: complexity, dynamics/tight coupling, and importance/priority of tasks. Thus, the subordinate failure type: Task Demand is added under the Condition of Task/Environment.

1.2.3. Unsafe supervision

During the informal interviews with experienced inspectors, we found the classification of Unsafe Supervision of HFACS was confusing to these officials who use English as second language. (Note: We did not translate HFACS into a non-English version, due to potential inaccuracies in the translation process; see Ma et al.,

2009) Some interviewees brought up this question immediately; they could not tell the difference between “Inadequate Supervision”, “Inappropriate Operations” and “Failed to Correct Problem” because they thought these error categories at least partially overlapped. The brief framework introduction helped in understanding these differences but some interviewees still had difficulty in appreciating the differences between these error categories at the end of the interview. (Note: Beyond any possible language barrier, we believe that limited exposure to HFACS before the interview was also one of the reasons for any definition confusion.)

On the other hand, in this study, the major design purpose of HFACS-MA is to conduct quantitative analysis to examine the predictive validity of audits. Because we anticipated a potential deficiency of data in some specific error categories (which would increase the difficulty of examining the prediction validity in Part 2 of this study), we were keen to establish a three layer hierarchy especially for Unsafe Supervision and Organizational Influence (in HFACS, only Unsafe Act and Precondition have a three layers hierarchy) to facilitate the flexibility and integration ability for future analysis (by integrating similar subordinate errors using their higher parent category).

We tried to modify the original taxonomy to conciliate the above concerns. As a result, we utilized the same classifying idea as in Unsafe Act to categorize inappropriate behaviors of supervisors: Dysfunction (unintended activity) and Disobedience (purposeful action). While the classification of Supervision Disobedience retained the same taxonomy as HFACS, we used the theories of management functions to help classify activities of Supervision Dysfunction. The concepts of management functions have reached convincible consensus among professionals. In Table 3, we list four authors to separate the functions appropriate to the mid-level managers with the potential for Unsafe Supervision. Since the functions of budget and staff are the authority of higher level managers, we considered them at the organization level. For the middle level managers, we placed Planning/Organizing into the first category, Leading/Coordinating in the second, and Controlling/Correcting into the third. Unsafe Planning/Organizing includes inappropriate work plans, poor resource allocation, and lack of technical support before maintenance tasks start. Leading/Coordinating error is failure to provide sufficient guidance, commands, or communication during the maintenance process. Finally, Controlling/Correcting error represents failed oversight, detection of potential problems, or allowing “known” deficiencies to continue unchecked.

1.2.4. *Organizational influence*

From the viewpoint of personnel in an organization, the factors of organizational level include the fallible decisions or activities of upper-level managers, interactions between departments, and personal perceptions of safety. Therefore, the factors classified as organizational aspects were incorporated into two categories: Organizational Functionality and Safety Climate in HFACS-MA.

We integrated the organizational process and resource management of HFACS into Organizational Functionality, whose definition is based on the concepts of Wheel of Misfortune, defined as the corporate decisions, rules, and supervisor activities that govern daily processes of the organization. Again, according to the theories of management functions, we classified the inappropriate decisions and activities of top managers into four separate groups: Operations/Procedure (as in Planning/Organizing), Execution (as in Leading/Coordinating), Resource Management (as in Budget/Staff), and Safety Oversight (as in Controlling/Correcting). Since Safety Oversight is also a dominant factor affecting Organizational Safety Climate (Cooper, 1998; Flin et al., 2000; Varonen and Mattila, 2000), we decided to classify it under the category of Safety Climate.

Table 3
Proposed management functions across four researchers.

Researchers	Management functions			
Fayol (1987)	Plan, Organize	Command, Coordinate	Control, Correct	Resource
Gulick (1937)	Plan, Organize	Direct, Coordinate	Report	Budget, Staff
Koontz et al. (1986)	Plan, Organize	Lead	Control	Staff
Robbins and Coulter (2005)	Plan, Organize	Lead	Control	

Schneider (1975) defined organizational climate as “molar perception people have of their work settings” (p. 473). Currently, organizational climate has been described as a subjective attribute of employees, and their shared perception of their organization, e.g. observable procedures, practices, and events (Denison, 1996; Patterson et al., 2005). In this study, we define Safety Climate as the subjective working atmosphere within the organization. It is affected by the treatment and practices of an organization directed toward internal individuals. We attributed Safety Climate to three factors in HFACS-MA: Safety Oversight, Safety Policy, and Safety Culture.

At the organizational level of aviation maintenance, Safety Oversight includes safety improvement programs, self-audit programs and accident/incident investigations. The oversight system is designed to help the whole organization to detect and correct existing problems: the more the emphasis upon the oversight system, the greater the likely safety consciousness of the personnel. Overall, Safety Oversight is a manifestation of the importance of safety at the organizational level.

Safety Policy provides the official guidelines that direct the daily decisions of first-line staff and managers (e.g., the safety commitment of top managers, drugs and alcohol treatment). These policies are the obvious face of safety presented within the organization, but effectively they can easily become a pretense if the managers do not put them into practice.

In contrast to policy, safety culture is considered as the unofficial or unspoken attitudes, values, beliefs, and customs that employees share related to safety (Shappell and Wiegmann, 2003). It is also what people often characterize as “the way things get done around here” (Cooper, 1998). The confusion between Safety Climate and Safety Culture has always existed (Hale, 2000). In this study, we consider Safety Culture as a broad multidimensional concept, and Safety Climate as an indication of Safety Culture (Cox and Flin, 1998), placing Safety Culture under Safety Climate in our framework.

In summary, despite these concepts and models (except perhaps management function theories) have being broadly applied to investigation and causal analysis of aviation accidents, interviews and discussion with officials and inspectors showed the necessity to develop a framework tailored specifically to an aviation maintenance audit system. HFACS-MA is designed specifically to be utilized in analysis of daily audit reports, rather than in accident investigation, and future users are assumed to be the official inspectors and human factors experts of a regulatory authority even though the framework has obvious application in self-audit of airlines. The detailed comparison of taxonomy between HFACS and HFACS-MA could be reviewed in Table 4.

In conclusion, the complete framework of HFACS-MA is similar to a fault-tree analysis structure. The influence direction between different failure tiers is a downward flow where the upper levels can affect the lower ones. Conversely, actual audit use is likely to be in the upward direction during the analysis process where analysts are expected to address each case from the tier of Unsafe Act, then Precondition, Unsafe Supervision, and eventually Organizational Influence. Similar analysis flow of human error was utilized in HFACS extensions (Rashid et al., 2010).

Table 4
Comparison between HFACS and HFACS-MA.

Human error tier	HFACS		HFACS-MA "Class P"	"Class S"	Comments
Unsafe acts	Errors	Decision errors Skill-based errors Perceptual errors	Error	Decision error Skill-based error	- "Perceptual errors" were integrated into the category of "Skill-based error"
	Violations	Routine Exceptional	Disobedience	Routine Exceptional	- "Violations" was re-named as "Disobedience"
Preconditions for unsafe acts	Condition of operators	Adverse mental states Adverse physiological states Physical/mental limitations	Condition of operators	Adverse state Limitation Teamwork	- "Personnel readiness", "Adverse mental states" and "Adverse physiological states" were combined into "Adverse state."
	Personnel factors	Crew resource management Personnel readiness			- "Crew resource management" was similar to "Teamwork"
	Environmental factors	Physical environment Technological environment	Condition of task/environment	Task demand Physical environment Hardware/software	
Unsafe supervision	Inadequate supervision Planned inappropriate Operations Failed to correct problem		Supervision dysfunction	Planning/organizing Leading/coordinating Controlling/correcting	- We used the concepts of management functions to modify "Inadequate supervision", "Planned inappropriate operations" and "Failed to correct problems"
	Supervisory violations		Supervision disobedience	Routine Exceptional	
Organizational influences	Organizational process		Organizational functionality	Operations procedure Execution	- "Organizational process" was related to the category of "Operation procedure"
	Resource management			Resource management	
	Organizational climate		Organizational safety climate	Safety oversight Safety policy Safety culture	- "Organizational climate" was expanded to "Safety oversight", "Safety policy" and "Safety culture"

The transition of audit records from qualitative documents to quantitative data should help identify any systematic trends of human failures and facilitate active hazard prevention with the development of appropriate classification framework as the first step. Although the original authors measured high reliability of HFACS (Shappell and Wiegmann, 2001a,b), recent independent reliability assessments have been less encouraging. O'Connor (2008) used the military version (HFACS-DOD) finding inadequate reliability in the nano-codes. Olsen and Shorrock (2010) performed reliability analyses of the Australian Defense Force version (HFACS-ADF) finding low inter-rater agreement when actual incident records were coded. Thus any new variant of HFACS needs to have assured reliability. In this study, we need to confirm that HFACS-MA can be utilized consistently by different users to support the long-term analysis of human failures in maintenance systems. Accordingly, the purpose of this study in Part 1 is to verify the measurement reliability of HFACS-MA.

2. Data collection

To put HFACS-MA into practice, we have to make certain that future users can identify and analyze audit failures into the same or similar causal factors, i.e. reach an acceptable level of agreement among themselves, so that the subsequent analysis of audit reports can represent a meaningful status of human failures. In this study, we developed a two-phase procedure to measure the inter-rater agreement of HFACS-MA between independent participants and collect consensus results of human failures sourced from audit reports for the purpose to conduct quantitative analysis to examine prediction validity in our next study, Part 2.

2.1. The design of the procedure

2.1.1. Materials

The classification stimuli (subject) which were analyzed were historical audit reports obtained from one civil aviation regulatory authority. All records were independent and conducted in native language of the authority; Fig. 3 is a sample of English translation to demonstrate the general content of audit reports. Translation of the original records was not considered in this study to avoid the possible errors in translation (Ma et al., 2009). Contents of the audit records were all remained unchanged to avoid possible bias of data arrangement. The number of the stimulus sets in phase I was 79 based on power analysis while for phase II, it totaled 1238, divided into twelve trial blocks. These stimuli were randomly chosen from the database, and ordered randomly.

2.1.2. Participants

Because HFACS-MA incorporates many human factors concepts and is designed to be utilized by human factors experts, we decided to recruit graduate students in human factors field, who already possess sufficient background knowledge to ensure rapid learning. Since the stimulus materials were from a non-English-speaking authority, the entire procedure was conducted in the native language of that authority. Thus participants have to be fluent in both that language and in human factors background. Each participant received 2 h of training from the principal investigator at the beginning of the procedure to become familiar enough with the framework that representative results could be achieved and also that we could study the state of learning performance during the analysis process. Four raters were needed for phase I, and two repeated raters from phase I were randomly recruited again in phase II. The time between phases I and II was two months.

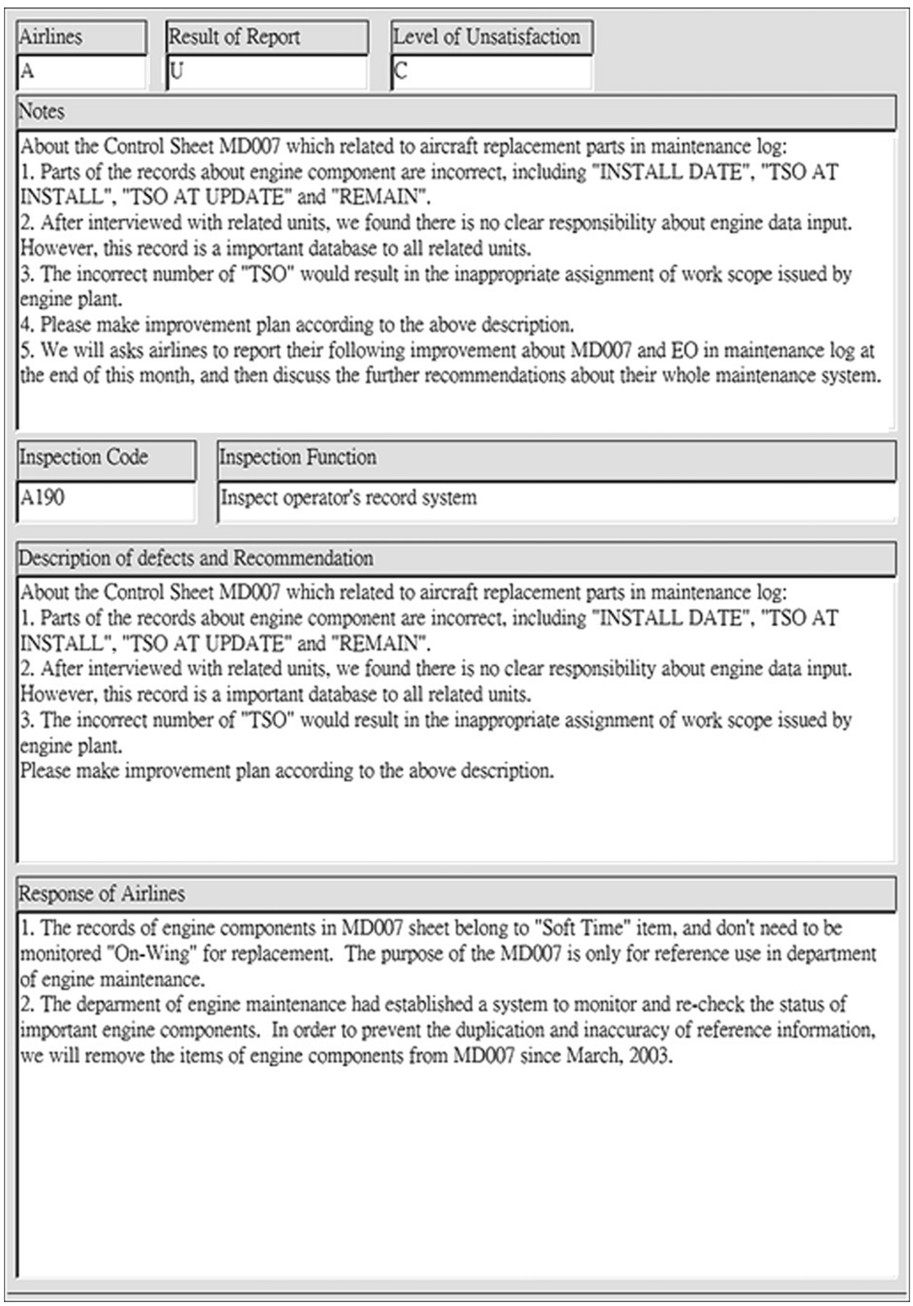
2.1.3. Procedure and instruction

Because the audit records were all remained original plain text without any data arrangement, the participants (raters) were asked to perform two successive tasks: read audit records to find existing flaws (first task), and diagnose each flaw which they found in each record into specific human error types based on the framework of HFACS-MA (second task).

After the raters finished the first task, if they had different opinions about determining the existing flaws, the participants needed to discuss the case to decide if the flaws exist or not before they conduct the second task. This step was required for the purpose of measuring inter-rater agreement. Once the raters finished the second task, which was to analyze the errors found in each report independently, their individual results were recorded by themselves separately on paper. Immediately after the analysis, they were required to announce their results to each other. If the results of the two error analyses were different, the participants needed to discuss the case and work as a team to attain a consensus conclusion at the end of each case for further analysis purpose. The consensus results would be utilized later in the examination of prediction validity in Part 2 of this study.

2.1.4. Analysis class of HFACS-MA

Because some specific errors were rare in the audit reports, the frequencies of occurrence of these unusual failures were mostly zero in any given monthly period. For instance, the monthly frequencies of zero occurrences of Decision Error and Exceptional Disobedience, were 61 and 73 times respectively ($N = 78$) (see Fig. 4a). This phenomenon would certainly increase the investigation difficulty of predictive validation in Part 2 of this study if the input values of some specific variables were mostly zero. Where this constraint existed, we tried to alleviate the problem by integrating similar errors at the next higher level in the classification framework. In HFACS-MA, we set up “Class P” factors as the ‘parent’ category of its ‘subordinate’ errors, “Class S” (see Fig. 1 and Table 1 for details). These integrated “Class P” factors would accumulate the frequencies of their subsidiary errors (“Class S”) to increase the quantities of non-zero monthly frequencies. To continue the above example, combining the zero occurrences of Decision Error and Exceptional Disobedience’s into their parent factors, Error and


Airlines: A
Result of Report: U
Level of Unsatisfaction: C

Notes

About the Control Sheet MD007 which related to aircraft replacement parts in maintenance log:
1. Parts of the records about engine component are incorrect, including "INSTALL DATE", "TSO AT INSTALL", "TSO AT UPDATE" and "REMAIN".
2. After interviewed with related units, we found there is no clear responsibility about engine data input. However, this record is a important database to all related units.
3. The incorrect number of "TSO" would result in the inappropriate assignment of work scope issued by engine plant.
4. Please make improvement plan according to the above description.
5. We will asks airlines to report their following improvement about MD007 and EO in maintenance log at the end of this month, and then discuss the further recommendations about their whole maintenance system.

Inspection Code: A190
Inspection Function: Inspect operator's record system

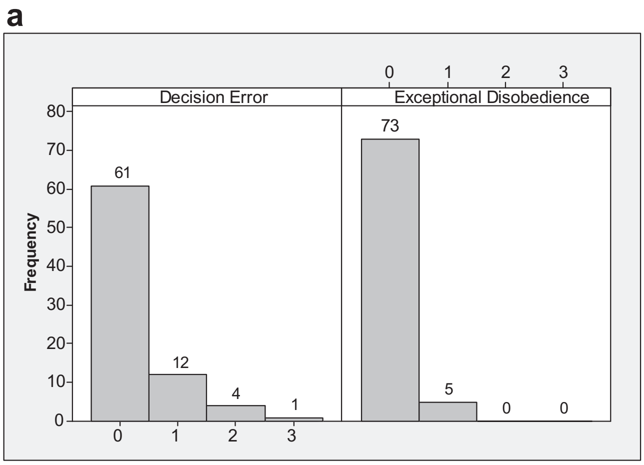
Description of defects and Recommendation

About the Control Sheet MD007 which related to aircraft replacement parts in maintenance log:
1. Parts of the records about engine component are incorrect, including "INSTALL DATE", "TSO AT INSTALL", "TSO AT UPDATE" and "REMAIN".
2. After interviewed with related units, we found there is no clear responsibility about engine data input. However, this record is a important database to all related units.
3. The incorrect number of "TSO" would result in the inappropriate assignment of work scope issued by engine plant.
Please make improvement plan according to the above description.

Response of Airlines

1. The records of engine components in MD007 sheet belong to "Soft Time" item, and don't need to be monitored "On-Wing" for replacement. The purpose of the MD007 is only for reference use in department of engine maintenance.
2. The deparment of engine maintenance had established a system to monitor and re-check the status of important engine components. In order to prevent the duplication and inaccuracy of reference information, we will remove the items of engine components from MD007 since March, 2003.

Fig. 3. Sample of the audit records in this study.



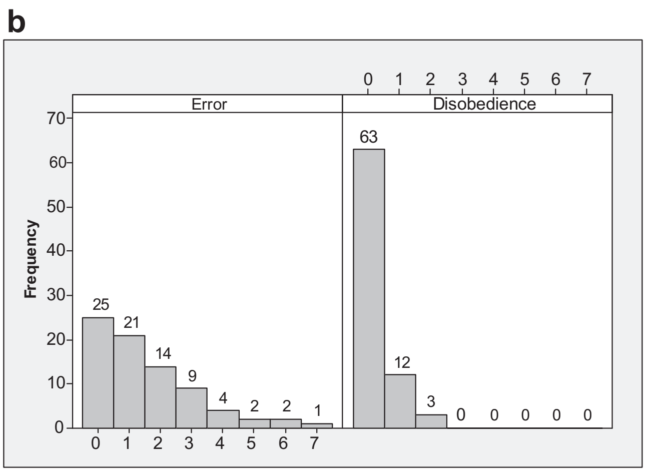


Fig. 4. a. Histogram of decision error and exceptional disobedience ("Class S"). b. Histogram of error and disobedience ("Class P").

Disobedience (as "Class P"), reduced them to 25 and 63 times respectively (see Fig. 4b).

"Class S" includes totally 21 error types such as Skill-Based Error, Teamwork, Planning/Organizing, and Operation Procedure; these belong to eight parent categories in "Class P", e.g. Error, Condition of Operators, Supervision Dysfunction and Organizational Functionality respectively. During the reliability survey, all failures found in audit records were classified into the 21 "Class S" errors. But because we eventually used those "Class P" categories to predict future safety performance in Part 2, the reliability analysis results of both "Class S" and "Class P" have to be presented in this paper to assure their corresponding measurement reliability.

2.2. Measurement of reliability

In order to eliminate the effects of chance agreement, we utilized the Kappa coefficient (Cohen, 1960), the standard for measuring inter-rater agreement in many different academic fields (e.g., education, disease diagnosis, epidemiology, and psychological applications), as the reliability indicator. Many studies using the Kappa method were related to dual raters (Devitt et al., 1997; Wing et al., 2002; Smits et al., 2009). The Kappa coefficient can provide valuable information on the reliability of diagnostic and examination procedures (Sim and Wright, 2005), which meets the requirement of this study. It indicates as the proportion of agreement between raters rating n subjects after chance agreement has been removed (Soeken and Prescott, 1986), and takes the form as Equation (1):

$$\text{Kappa} = \frac{\text{observed agreement} - \text{chance agreement}}{1 - \text{chance agreement}}$$

Symbolically,

$$K = \frac{P_o - P_c}{1 - P_c} \quad (1)$$

where P_o = the overall proportion of observed agreement, $(1/n)\sum_i n_{ii}$; P_c = the overall proportion of chance-expected agreement, $1/n^2 \sum_i n_{i\cdot} n_{\cdot i}$; n_{ij} = the number of subjects assigned rating i by Rater 1 and rating j by Rater 2; the observed frequency in the i, jth cell; $n_{i\cdot}$ = marginal row frequencies, $\sum_j n_{ij}$; $n_{\cdot j}$ = marginal column frequencies, $\sum_i n_{ij}$.

The value of Kappa ranges from 1.0 (complete agreement) to 0.0 (chance only). Although the choice of the benchmarks of Kappa is inevitably arbitrary and problematic (Brenner and Kliebsch, 1996; Hripcsak and Heitjan, 2002), various authors had proposed measurement thresholds of Kappa value based on their experience to provide criteria for Kappa interpretation. Landis and Koch (1977) suggested the values in Table 5 as standards for strength of agreement for Kappa. Banerjee et al. (1999) believed that values between 0.40 and 0.75 represent fair to good agreement beyond chance for most study purposes.

In general, a Kappa value above 0.6 is considered as "good". However, as the number of rating categories grows, the potential for disagreement will increase correspondingly (Brenner and Kliebsch, 1996). The number of choice categories in most literature reviews ranged from two to five. In our study, we ran the Kappa analysis for the four main failure tiers of HFACS-MA, i.e. Unsafe Act, Precondition, Unsafe Supervision and Organizational Influence (see Fig. 2), where the number of available failure options ranged from five to seven. During the analysis procedure, the participants could only select from each tier of HFACS-MA, e.g. for Unsafe Act, only five options could be chosen (including N/A); for Organizational Influence, only seven. This inherent numerosity of HFACS-MA increased the difficulty of reaching an acceptable Kappa value in each tier.

To examine any learning effects of HFACS-MA, the procedure was divided into two parts: an introductory phase ("I"), and an experienced phase ("II"). Since Kappa is frequently assessed for statistical significance through a hypothesis test (Sim and Wright, 2005), for phase I, the Kappa value is expected to above 0.4 (i.e. null hypothesis $K < 0.4$); and 0.6 for phase II (i.e. null hypothesis $K < 0.6$).

3. Results

3.1. Reliability measurement of HFACS-MA

In this section, the overall inter-rater agreement analysis of HFACS-MA was presented under both Class P and Class S to assure their measurement reliability and future application in Part 2 of this study.

3.1.1. The Kappa value of each failure tier using "Class S" factors

In this section, we analyzed and distinguished the Kappa value based on each HFACS-MA tier. In other words, all Kappa values were represented by Unsafe Act, Precondition, Unsafe Supervision, and

Table 5
Standards for Kappa coefficient (Landis and Koch, 1977).

Kappa value	<0	0.01–0.20	0.21–0.40	0.41–0.60	0.61–0.80	0.81–1
Strength of agreement	Poor	Slight	Fair	Moderate	Substantial	Almost perfect

Organizational Influence (see Fig. 5a and b). The major measuring difference between "Class S" and "Class P" is the number of available options, for instance, the measurement of Organizational Influence in Table 6a ("Class S") includes totally seven options and forms a 7 * 7 kappa matrix. In Table 6b ("Class P"), there are only three options to form a 3 * 3 matrix.

In phase I, four participants were recruited in the analysis procedure. The Kappa value of Unsafe Act was 0.47, and overall agreement was 73% in phase I. [Note: we will provide overall agreement data as well as the Kappa values: Olsen and Shorrock (2010) quote Ross et al. (2004) as preferring agreement over Kappa.] The z value against the null hypothesis of $K < 0.4$ was 2.41 rejecting the null hypothesis with 95% confidence, i.e. the classification reliability of Unsafe Act was better than anticipated. However the Kappa value of Unsafe Act was the only one of the four error tiers to reach significance in phase I. Although the overall agreement of Precondition, Unsafe Supervision, and Organization Influence was 63%, 63% and 54% respectively, their Kappa coefficients were all lower than 0.4. The overall reliability in phase I could only be rated at a "fair" level based in Table 5. Consequently, the results of phase I did not establish the measurement reliability of HFACS-MA.

Fig. 5a summarizes the Kappa values of the two repeated participants in phase I and phases II. There were total 1238 analyzed records in phase II, and we divided them into twelve trial blocks which each block contains 100 records except that the last block had 138 records. Trail number 0 in X axis indicates the Kappa value of phase I, and numbers 1–12 denote the trail blocks of phase II. For phase II, both the overall agreement and Kappa values increased beyond phase I. The Kappa values of both Unsafe Act and Precondition all attained the assumed criterion ($K > 0.6$) in every section. The z values against the null hypothesis were 8.4 for Unsafe Act and 10.43 for Precondition. The z value of Supervision was 5.83 in phase II. Therefore, these three failure tiers all showed with 95% confidence that their classification reliability was better than anticipated. However, while the average result of Unsafe Supervision reached the 0.6 standard, records 1–200 and 501–600 (trial numbers 1, 2 and 6) were slightly lower than 0.6. The measurement reliability of Unsafe Act, Precondition, and Unsafe Supervision all met at least the "good" standard.

The Kappa values of Organizational Influence were all poorer than expected. Although the average of the overall agreement of the organization tier was about 0.81 similar to the other tiers, the

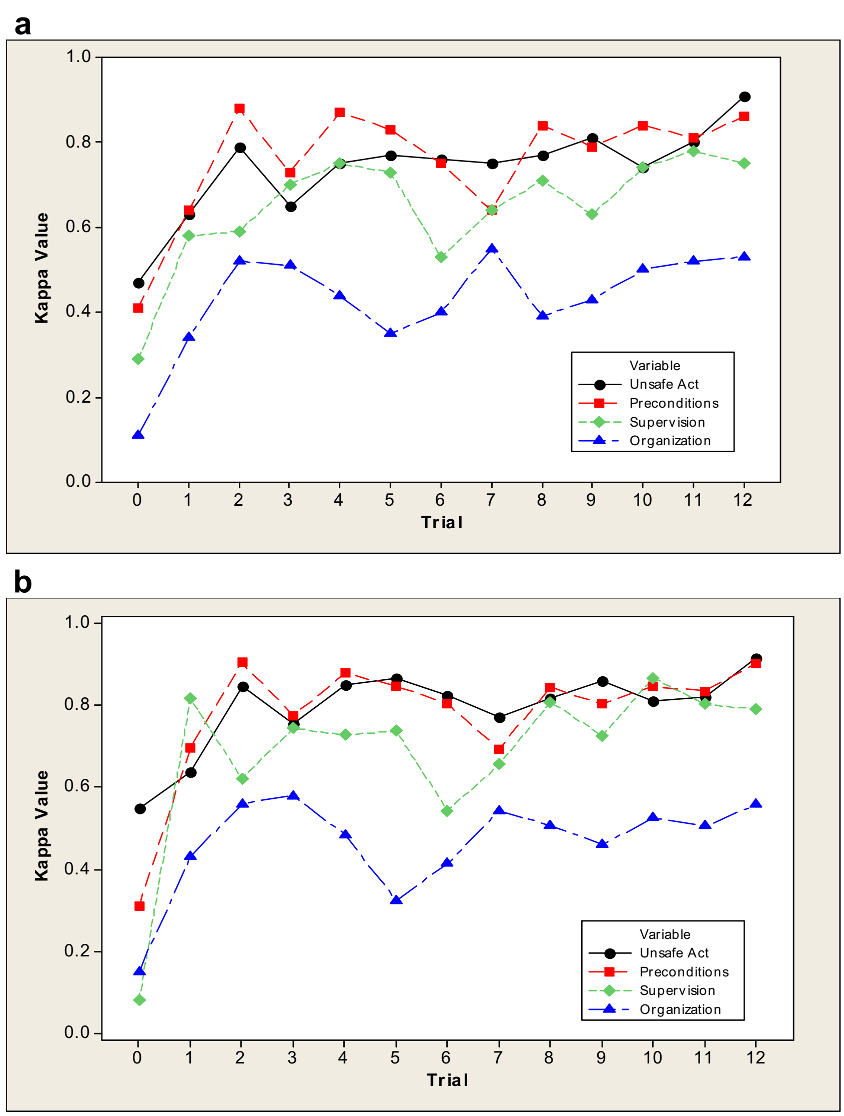


Fig. 5. Kappa values of four failure tiers using a. "Class S" and b. "Class P" factors by two repeated raters (The Phase I of the reliability procedure was coded "0" and Phase II was coded from "1" to "12").

Table 6a
The Kappa table of organizational influence ("Class S").

Rater # 1 \ Rater # 2	Procedure	Execute	Resource	Oversight	Policy	Culture	NA	Total
Procedure	21	2	2	2	1	1	6	35
Execute	1	2	1	4	0	1	4	13
Resource	0	0	43	1	0	3	16	63
Oversight	1	4	5	17	1	6	22	56
Policy	0	0	2	2	4	4	10	22
Culture	3	5	3	6	5	32	25	79
NA	13	8	18	24	4	48	974	1089
Total	39	21	74	56	15	95	1057	1357

average Kappa value of Organizational Influence only achieved 0.47 ("moderate" level), and the results of trial numbers 1, 5, and 8 were between 0.3 and 0.4. The significant number of "not applicable" (NA) responses, totaling 1172, was considered as one possible reason for moderate Kappa value (see Table 6a). Therefore, the Kappa values were re-calculated based only on the six failure categories of Organizational Influence, removing the NA option. The sample size of Kappa values totaled 185, still sufficient to reach an appropriate statistical power for Kappa measurement. The corresponding Kappa value was 0.547, which still failed to reach our assumed criterion ($K > 0.6$). As a result, the reliability measurement of Organizational Influence simply attained the "moderate" level based in Table 5. It should be noted here that although the number of audit records in phase II was 1238, because some records revealed more than one flaw in single report, the sum of analyzed data in Tables 6a and b was 1357.

In summary, after combining the results of phases I and II, the overall results of the Kappa analysis of four failure tiers using "Class S" factors nearly meet the original assumption, and reach an acceptable level of reliability measurement. Unsafe Act, Precondition, and Unsafe Supervision were fit for the "substantial" levels, and Organizational Influence was in the "moderate" level.

3.1.2. The Kappa value of each failure tier using "Class P" factors

Because of the rare appearance of some failures, we repeated the Kappa analysis by using the parent categories of "Class S", as "Class P" (see Table 5). This was done because the later quantitative analysis for prediction validity purpose in Part 2 of this study would be carried out at "Class P". Therefore, the reliability of failure categories at "Class P" needs to be established before HFACS-MA could be further utilized.

The Kappa values of four main failure tiers which were computed by "Class P" errors are illustrated in Fig. 5b. Generally speaking, the Kappa values of "Class P" were similar to "Class S", without large variations. In phase I of the procedure, Unsafe Act ($K = 0.54$) was still the only error category to meet the anticipated value ($K > 0.4$). In phase II, the Kappa values of both Unsafe Act and Precondition all attained the assumed criterion, $K > 0.6$, in every trial. The z values against the null hypothesis were 14.32 for Unsafe Act ($K = 0.82$) and 16.31 for Precondition ($K = 0.82$). The overall Kappa value of Unsafe Supervision was 0.75, and z value was 8.13. Therefore, these three failure tiers all proved with 95% confidence that their classification reliability was better than anticipated. The measurement reliability of Unsafe Act, Precondition, and Unsafe Supervision all reached the "substantial" standard in "Class P", and had similar analysis results as "Class S". However, the Kappa values of Organizational Influence still remained poorer than expectation while using "Class P" factors for calculation. The average Kappa of Organizational Influence was 0.51 ("moderate" level), which slightly improved upon "Class S" ($K = 0.47$), but still failed to reach the assumptive criterion.

Table 6b
The Kappa table of organizational influence ("Class P").

Rater # 1 \ Rater # 2	Function	Climate	NA	Total
Function	72	13	26	111
Climate	23	77	57	157
NA	39	76	974	1089
Total	134	166	1057	1357

In conclusion, combined the Kappa analysis of "Class S" and "Class P" (Fig. 5a and b), Unsafe Act, Precondition, and Unsafe Supervision all gave "substantial" levels of inter-rater agreement, and Organizational Influence was at the "moderate" level. Therefore, the human error taxonomy of HFACS-MA did reach an acceptable level of the reliability measurement, and should be able to be utilized in the further quantitative analysis in Part 2 of this study.

3.2. ANOVA and learning curves analysis of human failures

The learning effect of using HFACS-MA was examined in this section. As seen in Fig. 6, the performance starts with a rapid improvement (from Trial 0–200) follows by reduced improvements with further practice. Such ubiquitous learning curves are usually described well by power functions, and are often said to follow the "power law of practice" (Newell and Rosenbloom, 1981). The power law of practice can be explained by the cognitive "chunking" processing of the input and output of the task (Johnson et al., 2003). Participants learn the frequently occurring input–output patterns and develop their own decision schema in the initial few trials, but the rarer input–output patterns might require hundreds of trials for participants to memorize as chunks. The mathematical power equation of the learning curve can be expressed as Kappa value = (constant a) * (Trial)^(constant b) in this study. Ritter and Schooler (2004) mentioned that when using this equation, the constants can be easily computed by taking the log of the Kappa value and trial number to compute a linear regression. In other words, the power equation fits the regression line by using a log–log transformation. This function was fitted to the Kappa values shown in Fig. 6, giving a significant fit for all data, with $r^2 > 0.8$ and $p < 0.001$ for all four curves.

When the Kappa values were subjected to a two-way ANOVA, both failure tier and trial block were significant ($F(3, 36) = 80.03$, $p < 0.001$ and $F(12, 36) = 13.37$, $p < 0.001$ respectively). Comparing the means of each trial block, only Phase I and the first trial of Phase II were different from the other trials. Indeed, when trial blocks were re-coded as Phase I ("0"), Phase II trial 1–100 ("1") and Phase II trials 101–1238 ("2"), the error type effect was lower than the previous analysis with $F(3, 40) = 19.24$, $p < 0.001$, while for trial block an even higher significance ($F(2, 40) = 52.83$, $p < 0.001$) was

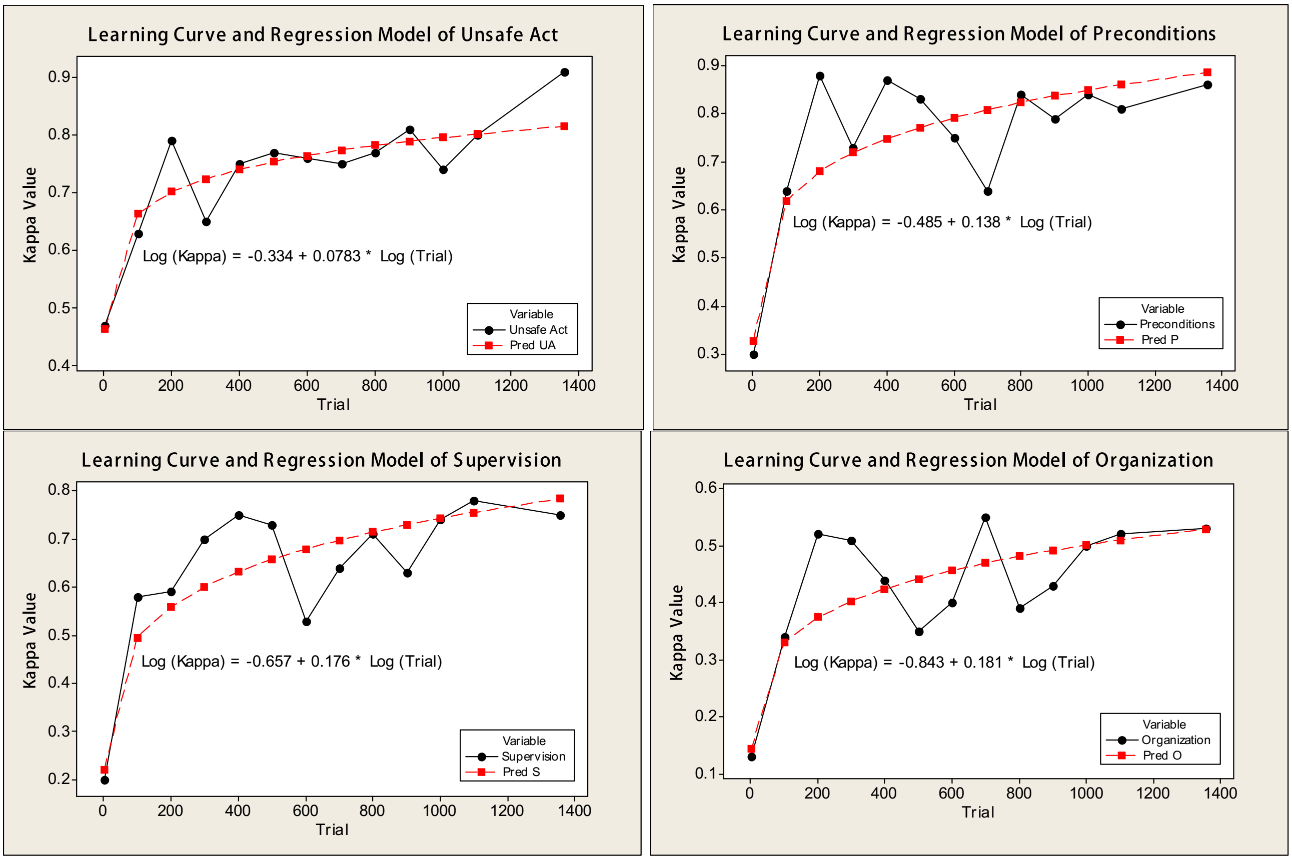


Fig. 6. Learning curve and regression model of four failure types.

found. Based on the results of Tukey, Bonferroni, and Sidak Method with 95% Confidence, the trial blocks were also re-grouped as Phase I ("0"), Phase II trials 1–1000 ("1") and Phase II trials 1001–1357 ("2"), and both re-grouped trial blocks and error type remained significant ($F(2, 40) = 48.25$, $p < 0.001$ and $F(3, 40) = 24.80$, $p < 0.001$ respectively). Tukey, Bonferroni, and Sidak post hoc comparison among the three re-coded classes showed differences significant at $p < 0.01$. The observation of a rapid plateau in the learning effect was confirmed by these analyses.

On the other hand, since there was only one observation value for each error tier and trial block [Note: Kappa value is measured across multiple records, e.g. 100 audit records], the degrees of freedom of error for the interaction between failure tier and original trial block was zero, and ANOVA was unable to further analyze the interaction. The re-coded trail blocks were also used for interaction measurement, and did not find significant result among them ($F(6, 40) = 0.28$, $p = 0.943$).

In conclusion, based on the results of learning curve analysis and the ANOVA, the four failure tiers all illustrated a clear learning effect. The participants appeared to become more acquainted with the analysis process from the initial phase but soon reached a plateau, as confirmed by discussions with the raters after the procedure. They became increasingly familiar with the contents of audit reports and developed their own decision schema for the classification procedure.

4. Discussion

4.1. Evaluation of the Kappa values

According to discussions with participants after the procedure, it was easier for participants to diagnose the failure categories in Unsafe Act than other failure tiers. This was because the contents of the audit record described most details about the inappropriate behavior of first-line workers, which is more straightforward for inspectors to discover and explain. In fact, we found that many audit reports lacked clear details of root cause explanations beyond the first-line employee, especially for organizational issues. This caused an extremely high percentage of Kappa analysis results to be the N/A option (see Tables 6a and b). Some reports even lacked the complete description of unsafe acts. Compared with the reliability study of HFACS which found 95% agreement or even higher at the Unsafe Act tier, our study only reached 88% overall agreement in Unsafe Act (Compared with the usual HFACS analysis of accident investigation reports), the incompleteness of daily audit reports is one of the reasons affecting the magnitude of the reliability measurement.

This phenomenon is resulted from the inherent limitations common to most contemporary civil aviation audit systems. Inspectors of regulatory authorities have neither sufficient resources (e.g., time and budget), nor adequate training in human factors concepts to conduct further investigation during their daily inspections. In addition, the description of inappropriate activities is solid enough for inspectors to proceed with enforcement (e.g. a violation) or close the case, which is how they see their primary responsibility. Thus, the raters may have to imply the possible causal failures contributing to Unsafe Act based on somewhat deficient information. This phenomenon was quantified by the Kappa values in our study. In the ANOVA given earlier, the Kappa values varied significantly among four main failure tiers. In the post hoc Tukey comparison tests, the Kappa values for Unsafe Act was not significantly different ($p = 0.95$) from that for Precondition, but all other comparisons were significant at $p < 0.01$. Therefore, insufficient description or information about latent factors in the audit reports was probably a considerable cause of low Kappa values especially at the tiers of Unsafe Supervision and Organizational Influence.

In order to examine the learning effects of using HFACS-MA, the training course was 2 h in phase I, which was similar to the training time of DOD-HFACS (O'Connor, 2008). Compared with HFACS itself, which provided at least one week of training for participants, the training time of our procedure was apparently inadequate to achieve ultimate reliability. However, Olsen and Shorrock (2010) chose air traffic controllers who were familiar with HFACS-ADF as raters and still found low reliability. In this study, the initial unfamiliarity of the tasks to the participants required more practice for them to become well acquainted with the concepts of the framework, and develop their own decision schema of case patterns. The analysis process was filled with aviation jargon, descriptions of aircraft maintenance and airlines operations initially unfamiliar to the human factors trained raters. Furthermore, because the participants came from the HF/E field, they were more familiar with the analysis of Unsafe Act and Precondition than Unsafe Supervision and Organizational Influence. These all increased the challenge for the participants to become competent raters in the reliability measurement procedure. In phase II, because two participants were recruited from phase I, the two raters had become more proficient in analysis. This can be recognized from Fig. 5a and b, which showed that the Kappa value of phase II reached a higher level than phase I. Based on the fitted regressions in Fig. 6, we tried to estimate the possible Kappa values with trial number reached 2000 and 3000. The estimated Kappa values of Unsafe Act, Precondition, and Unsafe Supervision in 2000 trials were all above 0.85 while Organizational Influence reached only 0.57. The Kappa of Organizational Influence would need more than 3000 trials to attain 0.6 which was the original set assumption of this study.

Moreover, compared with common diagnosis tasks, e.g., disease judgment: positive or negative (Hripcsak and Heitjan, 2002), the analysis task of the procedure was more complicated for raters. Because the contents of the audit records were narrated in a flat style of writing without summarized or numbered key points, the reliability procedure task was in effect two sub-tasks: first identify possible causes from the plain text description in the raw data, and then classify existing or potential causes into specific categories of error. Most other classification tasks only require the second of these. This difficulty increased the workload of the participants and thus may contribute to the lower Kappa values in this study. Ritter and Schooler (2004) considered as the difficulty of the tasks varied, the resulting line of performance would not appear as a smooth curve, but bounce around. And the elevated difficulty of the analysis task of the reliability procedure did illustrate in Fig. 6.

From the perspective of mathematical theories, the high percentage of chance agreement in the study was the noticeable reason of low Kappa in phase I. The rates of chance agreement of Precondition and Unsafe Supervision in phase I were obviously higher than in phase II. For Organizational Influence which has a total six error categories plus the N/A option to choose from in "Class S" (see Table 6a), the agreement responses of the N/A option was 974 times out of 1357 totally because both raters couldn't find any latent cues or possible failure descriptions related to the organizational tier due to insufficient information in many audit reports. The chance agreement of Organizational Influence tier steadily remained around 50–80% in trial blocks. Consequently, the Kappa value of the organizational tier had a difficulty in reaching a substantial level even when its overall agreement is about 80%.

According to Ludbrook (2002), Sim and Wright (2005), prevalence and bias effects are the key factors that can influence the magnitude of Kappa. A prevalence effect is represented by the inequality of representations of different categories. The bias effect means that one rater chooses some categories with higher (or lower) probability than the other raters when disagreements between raters happen. According to the results of the reliability procedure, the prevalence effect was more obvious than bias; in Tables 6a and b, the differences between agreement proportions were more noticeable than between disagreement. When a large prevalence exists, Kappa is lower than when the prevalence effect is low (Sim and Wright, 2005). It should be noted that in many laboratory studies, the stimulus material can be closely controlled (e.g. to ensure equality of categories), rather than selected from existing field reports. However, this research was closer to a field study, in which the sequence and allocation of the analysis samples, i.e. audit reports, could not be controlled. Accordingly, a prevalence effect became an inevitable consequence of the study and the low Kappa value of Organizational Influence was partly attributed to the existence of a prevalence effect.

4.2. Generalizability of the model

This is the first study to utilize an HFACS-like framework to analyze safety audit reports. Although we made several modifications to the original framework to fit its use in the maintenance audit field, the core concepts behind HFACS-MA and HFACS about active and latent human error are still the same. The most obvious difference of HFACS-MA is in the Unsafe Supervision and Organizational Influence tiers where we renamed the supervision failures due to language concerns and reorganized the original two levels hierarchy to three levels for analysis purposes (see Table 4). However, the meaning and descriptions of most failure categories were remained unchanged. Most failure categories in HFACS could be easily mapped to similar errors in HFACS-MA. Therefore, we believe the analysis results of HFACS-MA are still comparable to other HFACS-like models while discussing the context of human failures.

The utilization of maintenance terminology in analysis tool such as MEDA (Rankin, 2000) is intended to describe the specific results and behavior of human failures in the aviation maintenance domain. This kind of operational definition would definitely increase the usability to people in our chosen domain. However, based on the case studies of accident and incident reports, although the behavior descriptions or events may vary (e.g. incorrect installation of components or tools left in the airplane), we can still find similar contributing factors such as attention or memory failure across different cases. This is the reason why we utilized human factors categories instead of maintenance-specific behaviors to concentrate on the causes behind the observable human error descriptions in this study. The purpose is to categorize diverse maintenance failures by using well-accepted error categories. Generalization of HFACS-MA to other fields (e.g., nuclear power plant or chemical industry) is one of the future purposes of this research. Also, the expected users of HFACS-MA will have some expertise in HF/E. Therefore, this study used failure categories based on the human factors and management fields to represent the causes of various maintenance errors.

5. Conclusions

Since an audit program is considered a proactive method for accident prevention, we believe that a study based on a current audit system, e.g. a quantitative analysis of the states of human error, can benefit the accuracy of error detection and consequently the improvement of flight safety. The use of an HF/E model as the basis for our classification came about through a prior unsuccessful attempt to predict safety performance from raw numbers of audit findings, and from the knowledge that humans are involved in the causal chain of most aviation accidents. Therefore, it was important to develop a wide-ranging human error classification framework to detect and quantify the prospective risk. In this study, HFACS-MA was developed to fit the requirement of maintenance audit systems. After conducting the reliability study and considering the factors influencing the Kappa values, HFACS-MA can be regarded as a reliable classification tool to analyze daily audit records of maintenance organizations.

We believe that aviation accidents are preventable if people can detect and take actions to eliminate the potential hazards before accidents happen. The future purpose of this study is to examine the validity of HFACS-MA's safety prediction model based on the status of human error. In Part 2 of this study, the rates of human error will be utilized to develop a forecasting model to predict the safety performance of maintenance systems. The ultimate goal of our research is to establish a comprehensive methodology of risk management, and eventually facilitate preventing tragic accidents from happening.

References

Banerjee, M., Capozzoli, M., et al., 1999. Beyond kappa: a review of interrater agreement measures. The Canadian Journal of Statistics 27 (1), 3–23.

Brenner, H., Kliebsch, U., 1996. Dependence of weighted kappa coefficients on the number of categories. Epidemiology 7 (2), 199–202.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46.

Cooper, D., 1998. Improving Safety Culture: a Practical Guide. John Wiley & Sons Ltd.

Cox, S., Flin, R., 1998. Safety culture: philosopher's stone or man of straw? Work & Stress 12 (3), 189–201.

Dambier, M., Hinkelbein, J., 2006. Analysis of 2004 German general aviation aircraft accidents according to the HFACS model. Air Medical Journal 25 (6), 265–269.

Denison, D.R., 1996. What is the difference between organizational culture and organizational climate? A native's point of view on a decade of paradigm wars. The Academy of Management Review 21 (3), 619–654.

Devitt, J.H., Kurrek, M.M., et al., 1997. Testing the raters: inter-rater reliability of standardized anaesthesia simulator performance. Canadian Journal of Anaesthesia 44 (9), 924–928.

Drury, C.G., Dempsey, P.G., 2012. Human factors and ergonomics audits. In: Salvendy, G. (Ed.), Handbook of Human Factors and Ergonomics, fourth ed. John Wiley & Sons.

Drury, C.G., Guy, K.P., et al., 2010. Outsourcing aviation maintenance: human factors implications, specifically for communications. International Journal of Aviation Psychology 20 (2), 124–143.

Fayol, H., 1987. General and Industrial Management: Henri Fayol's Classic Revised by Irwin Gray. Lake Publishers, Belmont, CA.

Flin, R., Mearns, K., et al., 2000. Measuring safety climate: identifying the common features. Safety Science 34, 177–192.

Gramopadhye, A.K., Drury, C.G., 2000. Human factors in aviation maintenance: how we got to where we are. International Journal of Industrial Ergonomics 26, 125–131.

Gulick, L., 1937. Notes on the Theory of Organization.

Hale, A.R., 2000. Culture's confusions. Safety Science 34 (1–3), 1–14.

Hripcsak, G., Heitjan, D.F., 2002. Measuring agreement in medical informatics reliability studies. Journal of Biomedical Informatics 35, 99–110.

IATA, 2003. Annual Safety Report.

ICAO, 1998. Human Factors Training Manual. pp. 11–15.

Johnson, E.J., Bellman, S., et al., 2003. Cognitive lock-in and the power law of practice. Journal of Marketing 67, 62–75.

Johnson, W.B., 2001. Industry experience: implementing technology. In: The 15th Symposium on Human Factors in Maintenance and Inspection, London, UK.

Koli, S., Chervak, S., et al., 1998. Human factors audit programs for nonrepetitive tasks. Human Factors and Ergonomics in Manufacturing 8 (3), 215–231.

Koontz, H., O'Donnell, C., et al., 1986. Essentials of Management. McGraw-Hill, New York.

Landis, J., Koch, G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Li, W., Harris, D., et al., 2007. Routes to failure: analysis of 41 civil aviation accidents from the Republic of China using the human factors analysis and classification system. Accident Analysis & Prevention, 1538.

Ludbrook, J., 2002. Statistical techniques for comparing measures and methods of measurement: a critical review. Clinical and Experimental Pharmacology and Physiology 29, 527–536.

Ma, J., Drury, C.G., et al., 2009. Language error in aviation maintenance: quantifying the issues and interventions in four world regions. International Journal of Aviation Psychology 20 (1), 25–47.

Newell, A., Rosenbloom, P.S., 1981. Mechanisms of skill acquisition and the law of practice. In: Anderson, J.R. (Ed.), Cognitive Skills and their Acquisition. L. Erlbaum Associates, Hillsdale, NJ, pp. 1–55.

O'Connor, P., 2008. HFACS with an additional layer of granularity: validity and utility in accident analysis. Aviation, Space and Environmental Medicine 79 (6), 599–606.

O'Hare, D., 2000. The 'wheel of misfortune': a taxonomic approach to human factors in accident investigation and analysis in aviation and other complex systems. Ergonomics 43 (12), 2001–2019.

Olsen, N.S., Shorrock, S.T., 2010. Evaluation of the HFACS-ADF safety classification system: inter-coder consensus and intra-coder consistency. Accident Analysis & Prevention 42, 437–444.

Patterson, M.G., West, M.A., et al., 2005. Validating the organizational climate measure: links to managerial practices, productivity and innovation. Journal of Organizational Behavior 26 (4), 379–408.

Rankin, W.L., 2000. The Maintenance Error Decision Aid (MEDA) Process. IEA 2000/HFES 2000 Congress.

Rashid, H.S.J., Place, C.S., et al., 2010. Helicopter maintenance error analysis: beyond the third order of the HFACS-ME. International Journal of Industrial Ergonomics 40, 636–647.

Reason, J., 1990. Human Error. Cambridge University Press, New York.

Ritter, F.E., Schooler, L.J., 2004. The learning curve. International Encyclopedia of the Social & Behavioral Sciences, 8602–8605.

Robbins, S.P., Coulter, M., 2005. Management. Prentice Hall.

Roberts, K., 1990. Some characteristics of one type of high reliability organizations. Organization Science 1 (2), 160–176.

Ross, A.J., Wallace, B., et al., 2004. Technical note: measurement issues in taxonomic reliability. Safety Science 42, 771–778.

Schmidt, J., Schmorrow, D., et al., 2000. Human factors analysis of naval aviation maintenance related mishaps. In: Proceedings of the IEA 2000/HFES 2000 Congress.

Schneider, B., 1975. Organizational climates: an essay. Personnel Psychology 28, 447–479.

Shappell, S., Wiegmann, D., 2001a. Applying the human factors analysis and classification system (HFACS) to the analysis of commercial aviation accident data. In: 11th International Symposium on Aviation Psychology, The Ohio State University, Columbus, OH.

Shappell, S., Wiegmann, D., 2001b. A Human Error Analysis of Commercial Aviation Accidents Using the Human Factors Analysis and Classification System (HFACS). Federal Aviation Administration.

Shappell, S., Wiegmann, D., 2003. A Human Error Approach to Aviation Accident Analysis. Ashgate Publishing Limited, Hants.

Shappell, S., Wiegmann, D., 2004. HFACS Analysis of Military and Civilian Aviation Accidents: a North American Comparison. ISASI.

Shappell, S., Wiegmann, D., et al., 1999. Beyond mishap rates: a human factors analysis of U.S. Navy/Marine Corps TACAIR and rotary wing mishaps using HFACS. Aviation, Space and Environmental Medicine 70, 416–417.

Sim, J., Wright, C.C., 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical Therapy 85 (3), 257–268.

Smits, M., Janssen, J., et al., 2009. Analysis of unintended events in hospitals: inter-rater reliability of constructing causal trees and classifying root causes. International Journal for Quality in Health Care 21 (4), 292–300.

Soeken, K.L., Prescott, P.A., 1986. Issues in the use of kappa to estimate reliability. Medical Care 24 (8), 733–741.

Varonen, U., Mattila, M., 2000. The safety climate and its relationship to safety practices, safety of the work environment and occupational accidents in eight wood-processing companies. Accident Analysis & Prevention 32, 761–769.

Wing, L., Leekam, S., et al., 2002. The diagnostic interview for social and communication disorders: background, inter-rater reliability and clinical use. Journal of Child Psychology and Psychiatry 43 (3), 307–325.