On Classification in the Study of Failure, and a Challenge to Classifiers

Kimberly S. Wasson; University of Virginia; Charlottesville, Virginia, U.S.A.

Keywords: classification, failure, investigation

Abstract

Classification schemes are abundant in the literature of failure. They serve a number of purposes, some more successfully than others. We examine several classification schemes constructed for various purposes relating to failure and its investigation, and discuss their values and limits. The analysis results in a continuum of uses for classification schemes, that suggests that the value of certain properties of these schemes is dependent on the goals a classification is designed to forward. The contrast in the value of different properties for different uses highlights a particular shortcoming: we argue that while humans are good at developing one kind of scheme: dynamic, flexible classifications used for exploratory purposes, we are not so good at developing another: static, rigid classifications used to trap and organize data for specific analytic goals. Our lack of strong foundation in developing valid instantiations of the latter impedes progress toward a number of investigative goals. This shortcoming and its consequences pose a challenge to researchers in the study of failure: to develop new methods for constructing and validating static classification schemes of demonstrable value in promoting the goals of investigations. We note current productive activity in this area, and outline foundations for more.

Introduction

The study of failure and the development and practice of investigation activities rely in part on a wealth of classification schemes. These schemes serve a number of goals and purposes, some of them more successfully than others. Common purposes include providing a springboard for consideration of ideas from many angles, through the filter of a classification scheme that facilitates such exploration, as well as providing a mechanism to group and organize low-level data for specific analytic purposes and to direct responsive action. These purposes suggest certain properties that allow classifications to be more successful at accomplishing their intended goals, and classifications that are useful for disparate purposes will embody disparate properties. For example, flexibility is desirable in some circumstances, while rigidity is desirable in others. Consistent interpretability is desired of all schemes.

In this paper, we first survey a number of classification schemes developed for various purposes relating to failure and its investigation, and abstract from this survey a continuum of goal types that such classifications are intended to promote. We then discuss classification properties that are useful or valuable in promoting these goals, as well as those that inhibit them. We argue that current practice is generally insufficient to achieve a particular set of goals. In particular, we find that humans are more successful at creating and productively using one type than another, and that our lack of strong foundation for development of the second type negatively impacts the value of data generated via the use of some schemes. We examine this issue to better understand its mechanics, and suggest how significant improvements can be made to the state of affairs. In particular, we encourage the systematic exploitation of relevant knowledge from related disciplines, and provide two models of how it might be done, in the form of examples of current productive activity in this area.

Survey of Classification Schemes

In this section, we examine several well-known and less-well-known classification schemes constructed for various purposes relating to failure and its investigation, and discuss their values and limits.

(1) Petroski's Design Paradigms: Overview: In [11], Petroski presents a collection of design paradigms that exemplify error and judgment in engineering. His goal is to highlight the role of judgment and experience in achieving good design, and through the presentation of case histories, he hopes to aid the development of judgment in his readers by providing them with years of experience essentially by proxy.

Value: Petroski provides an origin for a wealth of discussion, a scaffold for consideration of ideas from many sides, and a filter by which to draw out commonalities among many events (for example, by providing an example of "tunnel vision" in design, he encourages the reader to generate analogous additional examples, possibly from disparate subfields of engineering, in order to highlight cross-cutting concerns). From this, it is possible to gain insight by generalizing across large amounts of experience and extrapolate from patterns. The case studies also provide accessible cues for anecdotes that drive home messages. This is a non-trivial accomplishment--the lessons are sold and remembered.

Limits: Petroski's paradigms were never intended "...to constitute a unique, distinct, exhaustive, or definitive classification of design errors." Indeed, they can sometimes be almost too flexible, so as to have little meaning (if everything can be everything, what is anything?) Thus this classification is not appropriate for doing any sort of quantitative analysis, but neither is it meant to be. As for the goal of aiding in the development of judgment, the paradigms are light on mechanism. That is, we are provided with a collection of models of good and bad judgment, but it is never explicitly discussed just *what* judgment *is*, at a psychologically low level, and how this classification scheme helps to develop it.

(2) Perrow's Interaction/Coupling Chart: Overview: In [10], Perrow argues that there exists the "possibility of managing high-risk technologies better than we are now," in addition to obvious steps like safer design and better operator training [10]. He argues that even with the most advanced safety mechanisms in place, some kinds of accidents are inevitable. He characterizes systems susceptible to such accidents by their high interactive complexity, that is, a large number of dependencies among elements of the system, and their tight coupling, that is, a lack of flexibility in the structure and timing of the processes that make up a system. High interactive complexity and tight coupling together affect the behavior of systems possessing them in critical ways that make appropriate response exceedingly difficult in critical situations.

Value: Perrow provides a scaffolding for discussing salient characteristics of high-consequence systems. It is somewhat less flexible than Petroski's paradigms, partly because Perrow wants to be able to drive policy decisions based on his classification, and to do so, it must have some integrity. He intends to provide a foundation for decision-making about which kinds of proposed systems should and should not be built, and which kinds of existing systems should be abandoned or modified. His classification does inform such decisions with useful information not previously thus synthesized.

Limits: Perrow only succeeds to the point that one agrees with his rationales for assigning industries to classes. The classes *are* somewhat subjective. While interactive complexity and tight coupling are reasonably well-defined notions and definitely capable of generating insight about

systems, they do not necessarily provoke easy consensus in classifying systems under consideration when other issues are thrown in that can affect policy decisions. In particular, there is disagreement about some aspects of the safety of nuclear power (but such disagreements provoke discussion, an emergent value from this limit).

(3) Reason's Generic Error-Modeling System: Overview: In [13], Reason discusses the psychological characterization of human error and presents a classification scheme by which to organize human error types. It is based on Rasmussen's SRK model [12], and enhanced by Reason's addition of further distinctions. The Generic Error-Modeling System uses research results from psychology about the mechanics of human information processing to inform a breakdown of error types according to the cognitive processing mode with which they are associated. It takes as a substrate the notion of the mind as a General Problem Solver (as per [9]) and first separates error events according to whether they occur before a problem is detected or afterward. Those that occur before map to Rasmussen's Skill-Based level, while those that occur after map to his Rule- and Knowledge-Based Levels. Those occurring before are further divided into slips and lapses, and those that occur afterward separate into Rule-Based mistakes and Knowledge-Based mistakes.

Value: Reason provides an explicit examination of mechanics that is theoretically founded and can be used to motivate preventive and corrective actions. It is not only more objective than the schemes of Petroski or Perrow, but it is more likely to be meaningful to the creation of strategies explicitly intended to take this problem into account when designing systems that better cope with it.

Limits: While it provides the possibility for constructing useful responses, it doesn't actually follow through (though that is reasonably beyond the scope of Reason's work). It lacks functional direction for application and requires others to take up the charge. One such researcher is Busse, whose work will be treated later in this paper [1].

(4) NASA, FAA, AIMS and ESRD Classifications Schemes for Use in Investigation and Monitoring: Overview: We treat these classification schemes together because they have in common certain properties with which we are concerned.[1] The schemes under consideration here are drawn from NASA's Procedures and Guidelines for Mishap Reporting, Investigating, and Recordkeeping [6] (NPG), the FAA's Order on Aircraft Accident and Incident Notification, Investigation and Reporting [16] (FAAO), the Australian Incident Monitoring Scheme [14] (AIMS), and National End Stage Renal Disease Patient Safety Taxonomy [7] (ESRD). The NPG and FAAO each outline policies and procedures governing activities to be undertaken during the investigation of specific incidents and accidents under their respective jurisdictions; the FAAO additionally governs certain activities of the National Transportation Safety Board (NTSB). Included in it are a number of schemes that direct the classification of the large volume of information generated as a result of an investigation. For example, the NPG and FAAO each provide a scheme for classifying undesired events (NASA "mishaps", aircraft incidents and accidents) according to severity; these classification assignments drive organizational response. The AIMS and ESRD Initiative provide direction in monitoring of ongoing activities and events and the inclusion of relevant information in databases for analysis. For example, both systems include schemes for attributing various levels and sources of causality of an adverse event.

[1] The reader might recognize that this is itself an implicit classification. It has its own value of aiding in the organization of this work and drawing the reader's attention to properties common to the classifications under discussion, and its own limit of being ad hoc, that is, useful for the purpose at hand, but in focusing on particular properties, it potentially ignores others by which other useful analyses might be attained.

Value: All of the classification schemes provide guidance in accomplishing activities that have the potential to bring about improvements in the systems with which they are concerned, through correction and prevention of existing faults and other sources of failure. Thus they apply at a lower level than the schemes previously discussed, which lack specific application mechanisms. Further, the results of the classifications associated with individual investigations and monitoring activities can be collected and analyzed together for trends that can provide additional insight.

Limits: The main drawback to these schemes is that inherent in them are semantic ambiguities that impede many of the goals of investigation and monitoring. For example, if an ambiguity allows two different reporters to classify identical events in different categories, then the classification scheme lacks integrity: analyses based on it are likely to find false patterns and miss actual ones. For example, the NPG classification scheme for mishap severity allows an interpretation with a contradiction [4], the FAAO generally classifies events involving loss of life as accidents while classifying those that involve hazardous materials, even if loss of life occurs, as incidents [16], AIMS gives little guidance in teasing apart the vagaries of inattention, fatigue, haste, or stress [1, 14], and the ESRD taxonomy definitions of root cause, proximate cause, and proximal cause are so circular and ungrounded as to leave the user more confused than had he not read them [7]. These ambiguities exist because the classification schemes were not developed with, for example, the rigor of Reason in exploiting a scientific basis (but they do have clear applicability where Reason lacks it). Busse characterizes the AIMS classification as forcing reliance on judgment and lacking in substance or discriminatory power, which can be said of each of the other schemes as well. The FAAO indicates explicitly that in the case where a particular need is not provided for by the document, investigators should use their judgment. But ambiguous classification schemes and over-reliance on judgment cannot promote the goal of integrity in classification, and thus meaningfulness of analyses and validity of responses.

(5) TransportNSW and NTSB Classification Schemes for Use in Investigation and Monitoring:
Overview: We treat these classification schemes together because they have in common certain properties with which we are concerned. The schemes under consideration here are drawn from The New South Wales Department of Transport's monitoring system for Signals Passed at Danger [15] (NSW) and the National Transportation Safety Board's scheme for allocating investigative resources according to distinctions in event severity [2, 8] (NTSB). NSW distinguishes three levels of severity of signals passed at danger: low, medium, and high, using factors such as total distance by which the signal was overrun and whether damage or death resulted. NTSB separates major from serious accidents using similar factors such as amount of damage and number of fatalities.

Value: Each of these schemes is unambiguous, and therefore capable of providing consistency that is missing in the schemes treated in the previous section. Patterns and trends observed are more likely to be actual patterns and not false ones.

Limits: While the disambiguity of these schemes allows more consistent tracking of data, it is not clear that the data being tracked are interesting. This is because the divisions among the classes in the schemes are based on observed outcomes rather than the origins of those events. In order to respond in a useful way, it is necessary to know how an event came to pass and not just its result. Whether a train overruns a signal by 183 meters or 184 meters (two separate categories in the existing scheme) is far less useful than knowing, for example, the distance the train is likely to be carried by its mass and inertia once the brakes are applied. The latter could form the basis of a taxonomy that helps to distinguish whether the brakes were a factor in an undesired event. Likewise in separating aircraft accidents from incidents based on the severity of the loss

sustained; whether a loss was sustained is more often a function of chance or luck than of the origins of contributing faults. [2] describes a near miss that would almost certainly have resulted in a midair collision had the aircraft had GPS installed; as it was, that the two aircraft missed each other was attributed not to any safety measure but rather to random noise in the ability of the aircraft to follow their programmed flight paths [2]. Certainly in observing outcomes there are intuitive apparent differences: multiple deaths seem to warrant more scrutiny than minor mechanical damage, and factors such as public relations encourage this to be so. But this is a false correlation when it comes to strategizing for prevention: in each of these cases, the quantitative measure of a degree (of loss, of damage, of arbitrary distance overrun) masks the problem of determining, through qualitative means such contextualizing an overrun distance in something physically meaningful, the likelihood of recurrence ([2] for aircraft near miss incident).

Analysis

The survey presented above affords the description of a continuum of uses for classification schemes, that suggests that the value of certain properties of classifications is dependent on the goals the classification is designed to forward. For example, flexibility in a classification might be desirable if the scheme is intended to provide a springboard for exploration of ideas, as in directing the consideration of an entity from many sides (as with Petroski's paradigms). On the other hand, rigidity is more desirable if we are concerned with trapping data related to a particular event into a characterization to be analytically processed with the goal of producing specific, actionable results (as with the tracking of error data to be used in informing, for example, system redesign). This continuum can thus be partitioned to reflect a dichotomy whereby non-domain-specific, high-level classifications tend to be dynamic and flexible, based on intuition, and in the service of exploration and generation of insight, while low-level, domain-specific classifications, generally applied to specific events under investigation or monitoring, tend to be static and rigid, and in the service of creation of analyzable organization in data, in a repeatable fashion.[2] One classification type favors flexibility, the other favors consistency and integrity. Among the classification schemes treated in this survey, those presented in survey sections 1 and 2 are more representative of the first type; Petroski and Perrow are concerned with abstracting inductively from large numbers of events in order to intuit patterns worthy of exploration. They might encourage some analysis, but neither provides much in the way specific, low-level results to be acted on in the correction or prevention of domain-specific faults. The schemes presented in the final two sections of the survey are the complement; they are explicitly constructed to trap low-level details of specific events in order to collect and analyze them, to direct corrective action on the systems involved. The remaining scheme, that of Reason, is something of a straddler in this analysis; while his scheme is not domain-specific or in the service of investigation of individual events, it *is* concerned with low-level cognitive mechanics that precipitate human error, and in addressing these mechanics, provides the foundation for specific corrective strategies in systems that suffer as a result of human error. This attention to origination of faults (in this case, human errors), and not just observation of their results, is valuable and precisely the kind of insight lacking in the NTSB and NSW schemes that distinguish classes by more arbitrary or less meaningful factors. However, what Reason lacks is the domain-specificity and application

[2] Repeatable, because we desire consistency not just within the investigation of a single event, but across multiple investigation instances that can be analyzed together in studies of wider scope. Further, repeatability allows analysis of the process itself in order to improve it; one cannot improve on a process that one cannot characterize and document, to know where to start in making the improvements.

mechanism to be able to *use* this low-level, mechanical information to inform corrective strategies in specific systems.[3]

The contrast in the value of different properties of different classification schemes for different uses highlights a shortcoming in achieving a particular purpose in the continuum: we argue that while humans are good at developing one kind of scheme, they are not so good at developing the other. Being furious pattern matchers, we are good at spotting the things common among entities under consideration; this would seem to indicate that we might do well at all aspects of classification. However, while we are able to induce patterns in disparate entities, and do well in exploring ideas and generating insights through dynamic, flexible classifications, and work with (and benefit from) their ambiguities and contradictions, we are far less successful at reaching the goals we intend for rigid, static, domain-specific classifications. We can construct them, that is, propose *some* taxonomy for a given environment, and declare it to be rigid and static, but it often turns out to be the wrong set of divisions--invalid as a rigid system, because we failed to set it up along the best possible lines and with the necessary explicit precision available to users. Without these properties, such schemes cannot meet the goals of integrity, meaningful analyzability of data, repeatability, or ability to motivate valid corrective responses.

Specifically, these deficiencies derive from two sources. As we saw in survey section 4, achieving the necessary explicit precision is one problem. This is a linguistic issue, and derives from the fact that our needs for this kind of precision are not something we are cognitively built to handle naturally. In [4], Hanks, Knight, and Holloway discuss the specific cognitive mechanics that allow for ambiguity and thus provide the environment for assumption to be relied upon in interpreting language. However, while these mechanics provide for high-bandwidth and language efficiency in the common case in which interlocutors share sufficient experience, these same mechanics backfire with severe consequences when the needs for precision are out of the ordinary. This allows for, and more likely, encourages, variation in the interpretation of, for example, guidelines directing the investigation of any disaster within their scope, limiting the degree to which that investigation can achieve its goals.

Survey section 5 provided discussion of the other main problem with developing successful static classification schemes; even if we can achieve the requisite precision to allow all users to arrive at the same interpretation, such interpretations are only useful if they are meaningfully connected to determination of origins of faults and not just their results. Recall, it is of far more value, from a standpoint of correction and prevention, to know if the distance overrun by a train was a factor of the braking system than whether it was 183 meters or 184; likewise is of far more value to know that two aircraft events of the same *potential* severity derived from the same electrical malfunction than that one of the events was accompanied by actual damage and loss of life while the other was not.

Why aren't we good at building valid static classifications? Because we build them on the wrong bases and with insufficient rigor in disambiguation. Our lack of strong foundation in developing useful rationales and methods of accurately increasing precision impedes progress toward a number of goals, like repeatability of process, meaningful analysis, and ability to drive valid corrective action. The problem has occasionally been referred to as an issue of the integrity of the classification, but most existing solutions amount to little more than "be careful." "Be careful" isn't enough. We need foundations. We cannot escape all uncertainty in interpretation, nor can we know every rational path from origin(s) to fault, but to advance our ability to generate useful

[3] We recognize that this is quite reasonably beyond the scope of his work; it is rather the field that lacks the means to apply Reason's work. Busse is making strides in this direction [1].

responses to specific events, and thereby to advance our understanding of failure generally, we should be trying more systematically to use all resources at our disposal to direct ourselves in removing all *unnecessary* uncertainty and misguidance.

Mandate

This shortcoming and its consequences pose a challenge to researchers in the study of failure: to develop new, more rigorously grounded methods for constructing and validating domain-specific static classification schemes. It is not enough to "be careful" in writing precision-oriented guidance documents, nor is it sufficiently productive assign investigative resources or develop corrective actions based on the results of a fault without also accounting for its origin(s) and the potential damage they allow. Even if we attempt to account for these, we are not doing as much as we can unless we are applying available relevant results systematically. We need methods of rigorously analyzing domains to access the structure and organization that mediates the knowledge through which we actually interact with the domains. It may be that one intuitive User Interface failure mode of a device is having its power supply kicked out of the wall, while another is having a coffee spilled on it, but these scenarios do not tell the whole story, do not represent the whole picture of our interaction with this specific device and the organized collection of concepts and understandings that mediate this interaction. Can we do better at capturing this information and driving static classifications off of it, such that the classifications have more integrity and thus the data analyses generated from them are more meaningful and the processes themselves can be made rigorous and repeatable and corrective actions are valid?

Current Work in Advancing this Cause

There are at least two projects taking specifically this approach in developing more valid static classifications. One is the methodology for better management of natural language throughout system lifecycles advocated by Hanks and Knight [3], which provides not only for better organization and contextualization of domain terminology for use in investigation guidelines and report documents, but for virtually any other component of a system lifecycle that relies on the use of natural language. This methodology is founded on results from linguistics and cognitive psychology that characterize specific cognitive mechanics involved in communication, and uses these mechanics to inform well-defined techniques and support tools for reducing the potential for miscommunication embodied in lifecycle artifacts using natural language. In particular, it addresses the related problems of precision and accuracy in communication using domain-specific terminology, to be used in classifications or otherwise. That is, it provides support structures and direction for communicating the correct concept, and at the appropriate level of granularity. Cognitive linguistic research results are thus exploited to shape methods that can be used to drive the construction of classifications that are less ambiguous and more cooperative with the deficiencies of natural human semantic organization--these methods do not just add explication, they add it in the right amounts and in the right places to allow interpretations and therefore dependent decisions of higher integrity.

Another project exploiting existing foundational results from relevant areas is the Cognitive Error Analysis methodology of Busse [1]. This work seeks to use existing results in psychology and cognitive science to inform techniques designed to reduce the incidence of human error in critical environments. It starts from the recognition that Reason's classification scheme, as discussed above, has desirable rigor in examining the origins and mechanics of human error, but lacks sufficient direction in application of its insights to the problem of developing preventive and corrective strategies. Busse addresses the problem of providing that functional direction, and her work has led to examination and improvement of a number of classification schemes used in

critical environments (e.g., a neo-natal intensive care unit). In particular, her work "...shows how error categorisation, when done according to a cognitive level of performance and latent factors, can provide the basis for sound, structured, and theory-based remedial recommendations" [1].

Of note is a further project that appears promising. The Laboratory of Decision Making and Cognition at Columbia University has a project in Human Error in Naturalistic Medical Environments. Among its goals is: "to develop a cognitive framework of medical errors in critical care environments, where decisions are often made under high stress, time pressure, and with incomplete information, which leads to a high degree of uncertainty in diagnosis and management. Our objectives include (1) developing a cognitive taxonomy of errors where each type of medical error is associated with a specific cognitive mechanism (2) a theoretical explanation of why these errors occur and prediction of the circumstances in which a specific error could occur, and (3) a cognitive intervention strategy based on the taxonomy that can prevent or reduce each category of medical error" [5]. However, while this initiative appears well-founded with regard to the priorities discussed in this paper, there are as yet no apparent results from this research group.

These projects, while providing example models, on their own contribute only drops in the proverbial bucket; their value has not yet been exploited, and there is a wealth of other foundations that can be explored for their usefulness in constructing more valid and useful static classification schemes. For example, there is far more available in both psychology and linguistics than either Busse or Hanks and Knight have explored. Among further options in linguistics is discourse analysis, and there are any number of high- and low-level psychological results relevant to human information processing, problem solving, and memory with surely hidden value. Sociology can inform interactions among individuals in modes other than linguistic communication. Biology can inform meaningful classification schemes for analyzing the effects of devices on live tissue. Chemistry and physics can do the same for interaction among any bits of matter or energy. In theory, results in the natural and social sciences could obviate the need to rely on any ambiguous or ungrounded classification scheme, but we as a community must make the commitment of resources to apply them.

Conclusion

There exists a continuum of uses and goals for classification schemes in the study of failure, and thus a continuum of properties that are useful and desirable in these schemes. The continuum of properties can be partitioned into a dichotomy opposing schemes that are dynamic and flexible, used for exploration and discovery, vs. those that are static and rigid, used for trapping data for analysis and creation of specific new results that inform preventive and corrective actions. The first type characterizes domain-independent inquiry, abstractions from many events, collected according to observed patterns that encourage new consideration of new angles. Flexibility is valuable, since it allows the examination of entities from many sides, sometimes simultaneously, and encourages generation of insight. The second type characterizes domain-specific inquiry, and in-depth investigation of individual or closely related events, in which classifications are created and applied, rather than observed and induced. The goal of the second type of classification is the opportunity for meaningful analysis, repeatability of process, integrity of results, and the ability to act on them. While humans are successful in creating and using the first type of classification, we are not as good at meeting the goals of the second type, because even though we can make schemes rigid by fiat, we have difficulty in developing classification schemes that are sufficiently disambiguous as well as sufficiently rationally founded to be useful. The result is an overabundance of invalid static classification schemes that do not support the goals for which they are intended. In this work, we assessed the state of the field with regard to this issue,

characterized the properties that contribute to the construction of more valid static classification schemes, and identified two projects addressing the problem in productive ways. We further suggested other research avenues that have potential to make positive contributions and encourage new work in these areas.

Acknowledgements

Students in the graduate seminar on Forensic Software Engineering led by myself and John Knight at the University of Virginia in the Fall of 2002 provided a valuable environment for discussion of ideas that contributed to this paper. This work was partially funded by NASA grants NAG-1-2290 and NAG-02103, and NSF contract CCR-0205447.

References

1. Busse, D.K. *Cognitive Error Analysis*. Doctoral Dissertation, Department of Computing Science, University of Glasgow, 2002.
2. Greenwell, W.S., J.C. Knight, and E.A.Strunk. *Risk-Based Classification of Incidents.* 2003 Workshop on the Investigation and Reporting of Incidents and Accidents, 2003.
3. Hanks, K.S. and J.C. Knight. *In Search of Best Practices for the Use of Natural Language in the Development of High-Consequence Systems.* Supplement (FastAbstracts) to the Proceedings of the International Conference of Dependable Systems and Networks, 2002.
4. Hanks, K.S., J.C. Knight, and C.M. Holloway. *The Role of Natural Language in Accident Investigation and Reporting Guidelines.* Proceedings of the 2002 Workshop on the Investigation and Reporting of Incidents and Accidents, C. Johnson, ed., 2002.
5. Laboratory of Decision Making and Cognition, Columbia University. http://www.cpmc.columbia.edu/patel/ldmc/LDMCWebpage090802/FrontPage090402.htm, as on March 28, 2003.
6. National Aeronautics and Space Administration QS/Safety and Risk Management Division. *NASA Procedures and Guidelines for Mishap Reporting, Investigating, and Recordkeeping (NPG 8621.1),* 2000.
7. National Patient Safety Foundation. *End Stage Renal Disease Patient Safety Taxonomy.* http://www.esrdpatientsafety.com/taxonomy.html, as on March 28, 2003.
8. National Transportation Safety Board. *Accidents and Accident Rates by NTSB Classification, 1983 through 2002, for U.S. Air Carriers Operating Under 14 CFR 121.* http://www.ntsb.gov/aviation/Table2.htm, as on March 28, 2003.
9. Newell, A. and H. Simon. *Human Problem Solving*. Englewood Cliffs, NJ: Prentice Hall, 1972.
10. Perrow, C. *Normal Accidents: Living with High-Risk Technologies*. Princeton: Princeton University Press, 1999.
11. Petroski, H. *Design Paradigms: Case Histories of Error and Judgment in Engineering.* Cambridge: Cambridge University Press, 1994.
12. Rasmussen, J. *Information Processing and Human Machine Interaction*. Amsterdam: North Holland Press, 1986.
13. Reason, J. *Human Error*. Cambridge: Cambridge University Press, 1990.
14. Runciman, W.B., A. Sellen, R.K. Webb, J.A. Williamson, M. Currie, C. Morgan and W.J. Russell. *Errors, Incidents, and Accidents in Anaesthetic Practice.* Anaesthesia and Intensive Care 21(5): 506-519, 1993.
15. Transport NSW (New South Wales Department of Transport). *Signals Passed at Danger.* http://www.transport.nsw.gov.au/safety_reg/spad.html, as on March 28, 2003.

16. US Department of Transportation Federal Aviation Administration. *Aircraft Accident and Incident Notification, Investigation, and Reporting*. http://www2.faa.gov/avr/aai/TABL8020.htm, as on March 28, 2003.

Biography

Kimberly S. Wasson, University of Virginia; Charlottesville, Virginia, U.S.A.; telephone - +1.434.982.2225; fax - +1.434.982.2214; e-mail - ksh4q@cs.virginia.edu

Ms. Wasson is a doctoral candidate in Computer Science at the University of Virginia. Her primary interests include software requirements, software forensics, and linguistic and psychological issues affecting the quality of activities in both areas.