Challenges of Measuring Performance for FAA's Safety Oversight System

Mark Hansen and Carolyn McAndrews

FAA regulates the safety of the aviation industry through the safety oversight system, which is a system of rulemaking, standard-setting, certification, accident investigation, rule enforcement, and surveillance activities. Federal programs, including those of FAA, use performance indicators to measure the achievement of program goals. As part of a broader program of developing risk management methodologies, FAA is researching performance indicators that can be used to measure the performance of the safety oversight system. One of its goals is to create performance indicators that can describe the safety oversight system's influence on safety outcomes such as fatalities. Creating performance measures that link activities to safety outcomes is challenging because it is difficult to establish the causation between oversight activities and these safety outcomes. This challenge is not unique to FAA, and external reviewers such as the Government Accountability Office have recommended that other high-reliability sectors, including rail, develop such indicators. In addition to safety outcomes, other aspects of safety oversight system performance can be described with meaningful metrics. The background and motivation for oversight evaluation in the aviation industry and in general are discussed, as well as the challenges, some generic and some unique, of evaluating aviation safety oversight activities. Research is also presented on how safety oversight evaluation is conducted outside aviation.

The FAA safety oversight system promotes flight safety by "assuring airworthiness of aircraft, competency of airmen, and adequacy of flight procedures and air operations" (*1*). This mission is performed by setting and enforcing regulations and standards; certifying individuals, organizations, and equipment; surveilling airlines and aviation-related activities; and investigating accidents and incidents. FAA's Flight Standards Service (AFS) carries out these activities with a workforce of 4,600 employees located in FAA headquarters, nine district offices around the United States, and 135 field offices around the United States and the world. The budget of aviation safety oversight was about $870 million in 2004 (*2*).

The philosophy that guides FAA's oversight system has evolved over the last two decades. Until the 1970s, the system focused mainly on developing rules and regulations and implementing them through the processes of certification, inspection, and enforcement. The enforcement includes a wide range of responses, from informal, on-the-spot corrections to formal administrative actions, to legal penalties such as fines and certificate actions (*3*). In short, the goal of the system was to "convert safety standards from concept to reality" and "to protect the many conscientious pilots from the irresponsible few" (*4*).

The mantra of the present safety oversight system, conspicuously displayed on the AFS seal, is "System Safety." This is a broader and somewhat more elusive concept than the pursuit of safety through regulatory compliance. Two published definitions of system safety are "the application of technical and managerial skills to identify, analyze, assess and control hazards and risks" or, more succinctly, "organized common sense" (*5, 6*). Major elements of the system safety philosophy include the allocation of oversight resources on the basis of risk, the need for airlines to have operating systems in place to identify and mitigate hazards and risks, and the promotion of a safety culture in which workers throughout an organization strive to increase safety (*3*).

At FAA this approach to safety oversight is most clearly manifested in the Air Transportation Oversight System (ATOS), which AFS uses to oversee the largest passenger airlines. ATOS was conceived during a 90-day review of the safety oversight system requested by Administrator Hinson in the aftermath of the ValueJet crash in the Everglades in 1996. Under ATOS, teams of inspectors are assigned to each airline. Each team develops a comprehensive surveillance plan that "includes a series of inspection tasks to determine whether the airline has systems in place to ensure safety and a second series of inspections to verify that the airline is actually using those systems" (*5*). The plan is developed in a structured fashion that takes into account inspectors' assessments of the internal safety management systems, safety performance history, operational stability, and operating environment (*5*). Inspection tasks are also heavily structured; inspectors complete detailed checklists indicating whether the airline is satisfying requirements for having systems in place for managing safety, known as safety attribute inspections, and for actually using those systems, termed element performance inspections.

FAA implemented ATOS for the 10 largest passenger air carriers in October 1998. Although external observers were generally supportive of ATOS as a concept, there has been widespread criticism of its implementation. A 2002 report by the U.S. Department of Transportation (DOT) Inspector General identified three areas of weakness: incomplete development of key ATOS processes, in particular inspection data analysis; inspector training and deployment; and lack of national authority leading to inconsistent procedures across regions (*7*). Many inspectors have been dissatisfied with the program because the time required to implement it resulted in curtailed surveillance activity. The National Transportation Safety Board (NTSB) report on the crash of Alaska Airlines Flight 261 cited deficient oversight as a contributing factor, which inspectors attributed to the ATOS transition (*8*).

Despite the difficulties with ATOS, FAA remains committed to the system safety approach and has established several programs to

Department of Civil and Environmental Engineering, National Center of Excellence for Aviation Operations Research, University of California–Berkeley, 107 McLaughlin Hall, Berkeley, CA 94720.

Transportation Research Record: Journal of the Transportation Research Board, No. 1937, Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 31–36.

extend system safety concepts to its oversight of other segments of the aviation system. The Certification, Standardization, and Evaluation Team (CSET) program, for example, applies such concepts to the process of certifying new entrant air carriers, and the System Safety Approach for General Aviation (SAGA) program does the same for general aviation, beginning with agricultural operators (*5*). On a more ambitious scale, AFS has initiated the System Approach to Safety Oversight (SASO) program, whose aim is to "transform the AFS and the aviation industry to a national standard for safety risk management" (*9*). The SASO program will include improved decision support and training for inspectors, incentives for industry to implement safety management programs, and enhanced systems for acquiring, storing, sharing, and analyzing safety data.

To support AFS's system safety initiatives, and in particular the SASO program, FAA's Office of Aviation Research (AAR) has undertaken a research and development program entitled Risk Management Decision Support (RMDS). The goal of the program is "to improve safety by making AFS oversight more systematic, effective, efficient, and targeted to deal with identified risks by developing risk management and decision support capabilities" (*10*). RMDS research activities currently focus on commercial aviation, and they address six requirements including the development of an aviation safety evaluation construct, risk analysis methods, safety performance measures, diagnostic procedures, decision support tools for aviation safety evaluation, and finally the development of a methodology for evaluating the safety oversight system.

The focus of this paper is on AAR-sponsored research in the last of these areas, a methodology for evaluating the safety oversight system. Although the other RMDS tasks are intended to yield concepts, methods, and tools that will be incorporated into the safety oversight system itself, the task of this research is geared to evaluating the health and performance of the oversight system that results from the SASO program—in a sense, a methodology for overseeing oversight. The evaluation system used here is expected to be independent of the safety oversight system itself so that it can track improvement, or degradation, as the safety oversight system evolves. It is expected to identify situations in which more oversight is needed and in which oversight can be curtailed. Above all, the evaluation system must inform the easily posed but difficult question: what does the safety oversight system contribute to aviation safety?

WHY EVALUATE THE OVERSIGHT SYSTEM?

The goal here is to develop a methodology for evaluating the FAA safety oversight system. For convenience, the desired outcome is named the Oversight System Evaluation Methodology (OSEM). An evaluation methodology should provide AFS and FAA management with the ability to analyze the health and performance of the aviation safety oversight system. It would help managers determine whether the oversight system is functioning properly and whether it has the desired effects on aviation safety. A sound evaluation methodology should be a basis for an evaluation system that supports management in its efforts to attain continuous improvement through refining oversight practices and reallocating oversight resources. Further, an evaluation system based on a sound methodology would support AFS and FAA participation in assessments of the oversight system conducted by external agencies such as the DOT Inspector General, the Government Accountability Office (GAO), and the Office of Management and Budget (OMB). Ideally, an internal safety oversight evaluation process may reduce the need for such assessments by enabling internal management to see and address problems before they become serious enough to attract the attention of external reviewers.

The logic for developing an OSEM and an evaluation system is not much different from the logic for other evaluation activities under way throughout the federal government. Many of these evaluation activities have roots that extend back several decades. Patton traces the beginning of modern-day evaluation research to "the massive federal expenditures on an awesome assortment of programs during the 1960s and 1970s" when "accountability began to mean more than assessing staff sincerity or political headcounts of opponents and proponents" (*11*). The need for such research derived from two factors: there was not enough money for government to do everything that the vision of the Great Society might call for, and more than money was required to solve complex human and social problems. These circumstances, combined with Kennedy-era belief in the ability of science to find solutions to social problems, triggered a spate of evaluation studies featuring rigorous experimental design, data collection, and causal modeling.

The federal aviation safety oversight program predated the Great Society by nearly half a century and developed at a time when the questions that spur modern-day evaluations were not widely asked. For the most part, AFS was what Wilson termed a "procedural organization" in which "managers can observe what their subordinates are doing but not the outcome (if any) that results from those efforts" (*12*). Aside from that, evaluation of the safety oversight system rested on the judgments and ideologies of senior managers, presidential administrations, and Congress. In making these judgmental assessments, decision makers were informed by the statistical safety record, results of accident investigations, feedback from the aviation community, and episodic inquiries and studies spurred by catastrophic accidents, occasional scandals, and the continuing stresses and strains created by rapid growth and technological change in commercial aviation.

With the Clinton administration's National Performance Review and contemporaneous passage of the Government Performance and Results Act (Results Act), scientific evaluation received new impetus, extending its reach to virtually every corner of government. In the cause of making government "work better and cost less," these initiatives sought to "refocus managerial attention on outputs rather than inputs" (*13*). The Results Act requires all government agencies to prepare annual performance plans focused on achieving measurable progress toward program goals. In order to implement the Results Act, the OMB developed the Program Assessment Rating Tool to assess "how program evaluation is used to inform program planning and to corroborate program results" (*14*). Furthermore, OMB reorganized so that budgeting and program planning decisions are made in an integrated fashion. Thus, the ability to measure outputs has become an important factor in the ability of programs to obtain budgetary inputs.

In addition to the governmentwide push toward program evaluation, factors more specific to aviation safety oversight and to AFS motivate and shape the effort to develop an OSEM. First, AFS's safety oversight program has been continually criticized by external evaluators, including the GAO, the NTSB, Ralph Nader, and former DOT Inspector General Mary Schiavo. The latter two have written popular books that give substantial attention to the subject, and GAO issued dozens of critical reports on aviation safety oversight beginning in the 1970s (*15, 16*). Second, much of the FAA has recently been transformed into a performance-based organization. In 1997, the National Civil Aviation Review Commission recommended "establishment of a Performance Based Organization (PBO) within the FAA for the development, management, and provision of air traffic services" that "would be held accountable by committing to specific measurable goals with

targets for improved performance. In exchange, the PBO is granted managerial flexibilities" (*17*). This recommendation resulted in the establishment of FAA's Air Traffic Organization (ATO) in 2003. While AFS is not part of ATO, it is no doubt influenced by ATO's commitment to performance management and measurement.

Finally, the same thinking that led AFS to adopt the system safety approach also creates the impulse to measure oversight system performance. Systems safety has its roots in the total quality management philosophy that stresses the minimization of production variation and risk, participative management, continuous improvement, and ongoing performance monitoring. In undertaking the OSEM, AFS is practicing what it preaches to the industry it regulates.

CHALLENGES

It is challenging to create indicators that can reveal what the safety oversight system contributes to aviation safety. One way to determine this contribution is to consider the influence of the safety oversight system on safety outcomes. Safety outcomes could include airplane crashes with fatalities, but outcomes could be extended to include risk precursors and initiating events such as engine failure or insufficient operator training. Or perhaps there are other conceptualizations of safety that may yield meaningful measures that are easier to obtain. In any case, development of the OSEM presents several significant challenges.

Infeasibility of Experimentation

To answer the question about how much safety the oversight system contributes to safety outcomes, one would like to compare two sets of air transport service, one with a safety oversight program and the other without safety oversight. An experiment similar to this has been performed only once. In the early 1920s, before the federal regulation of commercial carriers, the federally operated U.S. Air Mail Service had a safety oversight program that included regular medical exams for pilots, careful aircraft inspection, a 180-item checklist used at the end of each trip, and regular engine and aircraft overhauls. The investment in the safety program was significant: the ratio of mechanics to aircraft was nearly 4 to 1 and 94% of air mail service employees were ground personnel. The safety benefits of this investment were clear: the fatality rate for the U.S. Air Mail Service was one per 789,000 mi flown from 1922 to 1925, whereas the comparable figure for itinerant commercial fliers (for 1924 only) was one per 13,500 mi (*18*).

Nowadays, controlled experimentation with the safety oversight system is unacceptable. The consequences of an accident are too great for experimentation. Besides, in a competitive environment, FAA should regulate equitably. In lieu of a controlled experiment, quasi-experimental methods must be used that estimate the effects of "naturally occurring" variation in oversight activities. For example, an investment analysis of the SASO program prepared by FAA compared accident rates and inspection hours per aircraft for large jet operators, commuter carriers, and general aviation. Not surprisingly, they are inversely related: large jet operators are inspected more intensively and have lower accident rates. But there are many other factors besides safety oversight that contribute to these differences in safety outcomes (*19*). Another quasi-experiment, reported in the SASO Mission Need Statement, is a comparison by the U.S. Navy Naval Safety Center, which reports that two aircraft—the F/A-18 and the A-7—designed by using system safety principles had accident rates 60% and 80% lower than those for equivalent aircraft—the F-14 and F-4—whose design did not use this approach (*20*).

Small Number of Accidents

The outcome that is the most desired—a decline in crashes, fatalities, and injuries—has totaled nearly zero since the early years of the industry. Air travel, particularly travel on large scheduled air carriers, is a pretty safe activity. In 1998, Barnett and Wang estimated the passenger death risk. They found that the death risk per scheduled jet flight from 1987 to 1996 for all first-world airlines, inside and outside the United States, was about one in 8 million. Another way to say this is that an individual could take one flight per day for 21,000 years before becoming a fatality statistic (*21*). Moreover, more recent work by Barnett reveals that variation in 10-year fatal accident rates among major airlines may plausibly be attributed to chance alone (*22*). If the hypothesis that all large airlines are equally safe cannot be rejected, conclusive statistical results about the effect of safety oversight on large airline safety are virtually impossible.

Instead of measuring safety outcomes with accident and fatality rates alone, one could increase the number of observations by including other events and conditions that may be precursors to fatal accidents. Use of this method introduces more uncertainty into the model. One would need to establish how safety oversight affects precursors and also how those precursors affect safety outcomes such as crashes or fatalities. Because precursors do not always lead to accidents, a precursor to an event can actually indicate two different things. First, a precursor may signal a lack of safety because, somewhere in the series of events, the precursor event occurred. However, a precursor that does not lead to an accident represents a successful escape and may indicate that the airline's safety defenses are deep and robust.

Barnett and Wang discuss the use of precursors (mishaps) and wrote that "data analysis fails to support the conjecture that, the greater an airline's involvement in mishaps, the greater its propensity to suffer the disasters that passengers fear" (*21*). They explained their intuition: "Suppose that all airlines suffer emergencies at the same rate, but that some are more adept than others at resolving them without dire consequences. Then the more skillful airlines will have relatively few disasters but relatively more nonfatal mishaps, while the less-adept carriers will exhibit the opposite pattern" (*21*).

Understanding of Safety Production

In contrast to the role of the federal government in the U.S. Air Mail Service in the example mentioned earlier, FAA's current safety oversight system does not carry out maintenance, training, or operations directly. The safety oversight system can influence safety only indirectly through certification, rulemaking, rule enforcement, the threat of enforcement, surveillance, and the collection of information through accident investigation. FAA's oversight system achieves its goals by influencing the activities and production practices of aircraft manufacturers, airlines, labor unions, and other stakeholders. The effectiveness of the safety oversight system depends on the strength and effect of its influence on these other activities.

Instead of focusing all energy on constructing measures that relate oversight and safety outcomes, most of the de facto performance measures of the safety oversight system set that relationship aside and concentrate on more accessible measures of performance such as the

distribution of inspections across airlines over time, whether FAA carried out its plan for inspections, and whether inspectors have the appropriate training. There are a number of functions that the safety oversight system carries out to produce its broader mission of ensuring safety. Some of these functions, such as surveillance, are more accessible in terms of performance measurement. For example, the surveillance function comprises observable and recordable activities such as performing inspections, researching operator information, and training. Performance measures can be formed by using information such as number of inspections, number of inspector hours, number of hours of training, number of satisfactory or unsatisfactory inspections, the regulations violated, enforcement measures (e.g., fines), and so forth.

If the production of safety were understood better, that is, if more were known about how firms in the aviation system adapt to FAA's safety oversight activities, the influence of these observable safety oversight activities on the inputs into safety (e.g., use of pilot checklists, use of maintenance checklists, internal evaluation within air carriers) could be examined rather than focusing on crashes and fatalities.

Difficulty of Interpreting Inspection Results

Relying on data from surveillance activities also presents a complication for analysis. The results of inspections can be difficult to interpret. For example, in the case in which the surveillance system finds many violations, discovering the violations could mean that the inspection program is working very well, that the inspections are targeted at areas with the greatest risk, and that the inspectors are skilled and are reporting findings consistently. But the scenario could also point to problems in the oversight system because somewhere upstream the oversight system allowed underlying problems to persist. Thus, the health of the oversight system could be poor even if the surveillance program is healthy.

Similarly, in a case in which the surveillance system finds few violations of federal aviation regulations, this scenario could mean that the surveillance system is healthy and that the oversight system is healthy. But the situation could also indicate that the surveillance program is not targeting inspections to areas that exhibit weakness.

There is a third explanation of this scenario, one that has been documented by the GAO. A report from 1998 states: ". . . inspectors do not consistently report violations. Many of those we interviewed and surveyed volunteered that they handle many violations informally and, if compliance can be achieved on the spot, may not enter violations into their tracking system" (*3*). Relying on inspection outcome data to measure the health of the oversight system reveals only behaviors that have been recorded. If inspectors sometimes find that reporting all behaviors is a hindrance to achieving safer outcomes, the inspectors may choose to lock in the safer outcome instead of using the inspection procedure to reach the same result.

EXPERIENCE IN OTHER SECTORS

To gain perspective, and with the hope of finding an existing system that may have surmounted the challenges just identified, safety oversight systems and their evaluation in other sectors including rail and nuclear power generation were examined. Both sectors have well-developed safety oversight systems, although the specific oversight processes are different for each sector. Although the processes in each sector are different, safety engineers, managers, and policy makers learn from the experiences of all sectors. Each sector relies on the cumulative experience across sectors in human and organizational factors, ergonomics, defense-in-depth, system safety, and risk assessment to carry out the oversight missions. In a similar way, external reviews of safety oversight systems by GAO and OMB often pose similar questions about whether regulations and surveillance activities are uniformly (or otherwise justifiably) applied, whether plans for oversight are carried out, and whether oversight system managers are improving safety oversight. As a result of the communication across sectors, and the similar training of safety managers, many safety systems and safety oversight systems exhibit similar designs that differ mainly because of differences in regulatory mandates and organizational characteristics.

Although rail and aviation face similar challenges with respect to the interpretation of inspection findings and multiple models of safety production, oversight evaluators in rail do not have the problem with small numbers. The safety record shows consistent improvements in safety, as measured by the numbers of accidents and fatalities. But the railroad industry has more observations with which to measure the effects of its oversight activities. From January to December 2003, for all railroads, the industry experienced 13,958 accidents and 860 fatalities (*23*).

Since 1996, FRA has used a collaborative approach to oversight in which teams comprising regulators and representatives from both railroad management and labor unions assess the safety needs of the railroad and set to work addressing the needs. FRA named this approach the Safety Assurance and Compliance Program (SACP). FRA monitors progress on the implementation of the safety plan and considers the completion of the plan and its achievements as indicators of successful safety oversight. This collaborative approach complements, but does not substitute for, the traditional rulemaking, surveillance, and enforcement method of oversight, which FRA uses as well.

The OMB found, as part of its standardized performance review of programs, that the FRA's safety oversight system needed a formal evaluation process but that one did not exist at the time of the assessment (*24*). But even without a formal process for assessing the health and performance of the safety oversight system, FRA adjusted its oversight system "in gradual shifts" and with a "best practices approach" to problem solving (*25, 26*). FRA's newest approach to oversight, the collaborative approach, may allow for informal evaluation to occur regularly. The participants in the evaluation include both the regulator and representatives from industry and labor, and one of the primary outputs of the evaluation is increased communication among these stakeholders.

Safety oversight in nuclear power production relies on the traditional method of oversight. The Nuclear Regulatory Commission (NRC) collects operating data, in the form of performance indicators, from each nuclear reactor. Utilities self-report these indicators to NRC each quarter and NRC uses its surveillance program to verify some of the measures. NRC also measures activities that are not self-reported such as whether a utility has a safety-conscious work environment.

Just as NRC relies on operations and safety performance indicators to carry out its oversight program, it also uses performance indicators to assess the performance of the oversight system. Data for these performance indicators are collected from oversight operations records (number of inspections, date of inspection, etc.) and through surveys of stakeholders. NRC uses 70 performance measures to characterize the performance of the oversight system. The performance measures assess the oversight system's (*a*) performance indicator program, (*b*) inspection program, (*c*) significance determination process (the process through which NRC determines whether a utility's opera-

tions, as measured through the performance indicators and inspection findings, are safe enough or whether they need more oversight), (*d*) assessment program, and (*e*) communication activities and other program issues. A few illustrative self-assessment performance indicators include the percentage of inspection reports that find the utility is in compliance with their program, the number of significant changes made to the inspection process, the completion of planned oversight activities, the timeliness of oversight activities, the predictability of the significance determination findings, and the accurate communication to the public of oversight findings (*27*).

Although the NRC's oversight evaluation system is formal and the FRA's is relatively informal, both of these evaluation systems share an important characteristic. The evaluation systems in both industries open the lines of communication between the regulator and the regulated industry. In the context of nuclear power production, utilities and advocacy groups are regular and vocal participants in rulemaking and changes in the safety oversight system. For example, NRC's self-assessment from calendar year 2003 described stakeholder survey responses on the topic of the performance indicators:

> [R]esponses . . . indicated that the public and the nuclear industry have varying views on the efficiency and effectiveness of the PI [performance indicator] program. The industry generally believed that the PI program was working well. . . . By contrast, the public has become increasingly concerned that the PIs are being managed by the licensees and have become ineffective as indicators of plant performance. (*27*)

In this example, the utility companies believe that the performance indicators are a good model of how they produce safety. According to them, if the performance indicators indicate safe operations, then the operations are probably safe. However, the representatives of the public interest have a different model of safety production, and they believe that other performance indicators would more accurately describe the relationship between reactor performance and safe outcomes. In this case, the self-assessment process collected information about competing models of safety production.

None of the self-assessment performance indicators in NRC's evaluation system measure the safety oversight system's direct effect on safety outcomes, but they do address the issues over which NRC has direct control, which is an important distinction. As illustrated in the foregoing example, NRC has control over which performance indicators it chooses, but its judgment will be influenced by guidance from stakeholders.

The literature about other industries was approached in hopes of learning about the evaluation methodologies used there: the performance measures, the analytical frameworks, and the key components of oversight that reveal the health of the system. It was found that other industries do use performance measures and that the performance measures are not so different from those used by the external evaluators such as GAO. It was also found that none of the indicators link oversight activities to safety outcomes.

As oversight evaluation in these other industries was reviewed, it became clear how these industries use evaluations in practice. From these studies it can be seen that the dominant characteristics of the industry, such as the positions of stakeholders or the preference for informal or formal communication methods, will be characteristics of an evaluation system. From these other industries it can also be seen that the oversight evaluation systems are not independent from the structure of the oversight system: NRC uses performance indicators to measure the performance of the utilities and its own oversight system; FRA uses a collaborative approach to oversight as well as to its oversight evaluation.

CONCLUSIONS AND FUTURE WORK

It is reasonable to investigate the health and performance of the aviation safety oversight system. There is a strong public interest in maintaining and improving aviation safety; the public and its elected representatives are entitled to know the return on a program with a direct public cost approaching $1 billion and an indirect cost to industry that is considerably higher. In addition, performance measurement is an integral part of performance management.

But evaluation of this system is extremely challenging. Experimentation is impossible, and naturally occurring quasi-experiments are difficult to find. Small numbers of accidents, particularly in commercial aviation, although a blessing to all, pose a further hurdle to the evaluator. The analysis of incidents and precursors, although appealing from a statistical standpoint, is tricky because an incident that does not lead to an accident can signify different things. Inspection results are also difficult to interpret, particularly when inspectors, perhaps for good reason, often do not report violations. Moreover, as AFS reorients its oversight program toward a system safety approach, measures of compliance, even if reliable, will become less relevant.

Underlying the challenges of creating performance measures of system safety is the question, What is a healthy oversight system and what makes it healthy? If the answers to those questions are known, they reveal where to go looking for performance measures. Traditionally, the oversight system is thought to perform well if it has oversight plans and those plans are followed. One can go further and check to see that those plans are sensible: consistent regulation yet targeted to areas of greatest risk. But the system safety literature shows that there is more to safety production than making and following plans and rules; procedural safety is as outmoded as is a procedural organization. Aviation inspectors may have known this all along, giving rise to an inspector culture that encourages voluntary reporting and smooth relations between inspectors and the certificate holder. What are the signs that tell whether the oversight system is asleep at the wheel or simply working quietly, subtly, but effectively? Does NRC's annual tome evaluating its oversight system with its 70 indicators really tell more than FRA's simply reporting that the railroads are carrying out their plans to improve safety?

It is certainly possible to create an OSEM that allows AFS and FAA management more than they now know about how the oversight system is working. A methodology that relates safety oversight activities to safety production must combine a risk analysis framework and an organizational approach. There must be an effort to create evaluation metrics that capture the link between oversight activity and safety outcomes. But pursuing this goal will respect the delicate nature of aviation safety oversight and recognize the well-known fact that evaluation drives behavior. The approach to assessing oversight system health, like that of the physician concerned with human health, must start with the principle "first do no harm."

ACKNOWLEDGMENT

The research discussed in this paper was sponsored by FAA through the National Center of Excellence for Aviation Operations Research.

REFERENCES

1. Kemp, D. The Federal Aviation Administration's Air Safety Program. Presented at Government-Industry System Safety Conference, NASA Safety Office, National Aeronautics and Space Administration, May 1–3, 1968.

2. *Budget of the United States Government, Fiscal Year 2005*. Office of Management and Budget, Appendix, p. 767. www.whitehouse.gov/omb/budget/fy2005/appendix.html. Accessed July 31, 2004.
3. *Aviation Safety: Weaknesses in Inspection and Enforcement Limit FAA in Identifying and Responding to Risks.* GAO/RCED-98-6. General Accounting Office, Feb. 1998.
4. Burkhardt, R. *The Federal Aviation Administration.* Frederick A. Praeger, New York, 1967.
5. *Aviation Safety: FAA's New Inspection System Offers Promise, but Problems Need to Be Addressed.* GAO/RCED-99-183. General Accounting Office, June 1999.
6. Lederer, J. National Aeronautics and Space Administration. Presented at Government-Industry System Safety Conference, NASA Safety Office, National Aeronautics and Space Administration, May 1–3, 1968.
7. *Oversight of Aircraft Maintenance, Continuing Analysis and Surveillance Systems.* AV-2002-066. Office of Inspector General, U.S. Department of Transportation, 2001.
8. *Aircraft Accident Report: Loss of Control and Impact with Pacific Ocean, Alaska Airlines Flight 261, McDonnell Douglas MD-83, N963AS About 2.7 Miles North of Anacapa Island, California, January 31, 2000.* NTSB/AAR-02/01. National Transportation Safety Board, 2002.
9. Niemeier, D. SASO Communication Messages. Presented at Risk Management Decision Support Meeting, Atlantic City, N.J., July 13, 2004.
10. Fazen, K. M. Risk Management Decision Support—RPD 676 Commercial Aviation Requirements Program Review. Presented at Risk Management Decision Support Meeting, Atlantic City, N.J., July 13, 2004.
11. Patton, M. Q. *Utilization-Focused Evaluation.* Sage Publications, Beverly Hills, Calif., 1978.
12. Wilson, J. Q. *Bureaucracy: What Government Agencies Do and Why They Do It.* Basic Books, New York, 1989.
13. Kettl, D. F., and J. J. DiIulio, Jr. *Inside the Reinvention Machine: Appraising Governmental Reform.* Brookings Institution, Washington, D.C., 1995.
14. *Assessing Program Performance for the FY 2005 Budget.* Office of Management and Budget. www.whitehouse.gov/omb/part/. Accessed July 31, 2004.
15. Nader, R., and W. J. Smith. *Collision Course: The Truth About Airline Safety*. Tab Books, Blue Ridge Summit, Pa., 1994.
16. Schiavo, M., and S. Chartrand. *Flying Safe, Flying Blind.* Avon Books, New York, 1997.
17. National Civil Aviation Review Commission. *Avoiding Aviation Gridlock and Reducing the Accident Rate: A Consensus for Change.* www.faa.gov/NCARC/reports/pepele.htm. Accessed July 31, 2003.
18. Commons, N. *Bonfires to Beacons: Federal Civil Aviation Policy Under the Air Commerce Act, 1926–1938.* FAA, U.S. Department of Transportation, 1978.
19. Oster, C. V., Jr., J. S. Strong, and C. K. Zorn. *Why Airplanes Crash: Aviation Safety in a Changing World.* Oxford University Press, New York, 1992.
20. *AFS-900, System Approach to Safety Oversight—Mission Need Statement.* FAA, U.S. Department of Transportation, 2001.
21. Barnett, A., and A. Wang. *Airline Safety: The Recent Record.* NEXTOR Research Report RR-98-7. Massachusetts Institute of Technology, Cambridge, Mass., 1998.
22. Barnett, A. The Mother Metric? FAA's New Safety Index. Presented at Moving Metrics: A Performance-Oriented View of the Aviation Infrastructure, Asilomar Conference Center, Pacific Grove, Calif., Jan. 27–30, 2004.
23. *Accident/Incident Overview*. FRA, U.S. Department of Transportation. http://safetydata.fra.dot.gov/officeofsafety/. Accessed July 30, 2004.
24. *Department of Transportation PART Assessments*. Office of Management and Budget. www.whitehouse.gov/omb/budget/fy2005/pma/transportation.pdf. Accessed July 31, 2003.
25. *Five-Year Strategic Plan for Railroad Research, Development, and Demonstrations.* FRA, U.S. Department of Transportation, 2002.
26. Molitoris, J. M. Statement of Jolene M. Molitoris. *Testimony before the Senate Committee on Commerce, Science, and Transportation Subcommittee on Surface Transportation and Merchant Marine,* Feb. 25, 1998.
27. *Reactor Oversight Process Self-Assessment for Calendar Year 2003.* SECY-04-0053. Nuclear Regulatory Commission, 2004.

Statements in this paper do not reflect the official positions or policies of the Federal Aviation Administration or the National Center of Excellence for Aviation Operations Research.

The Vehicle User Characteristics Committee sponsored publication of this paper.