Contents: v.1. Properties and selection—irons, steels, and high-performance alloys—v.2. Properties and selection—nonferrous alloys and special-purpose materials—[etc.]—v.21. Composites
1. Metals—Handbooks, manuals, etc. 2. Metal-work—Handbooks, manuals, etc. I. ASM International. Handbook Committee. II. Metals Handbook.
TA459.M43 1990 620.1′6 90-115
SAN: 204-7586
ISBN: 0-87170-704-7
ASM International
Materials Park, OH 44073-0002
www.asminternational.org
Printed in the United States of America
Multiple copy reprints of individual articles are available from Technical Department, ASM International.

Introduction to Failure Analysis and Prevention

James J. Scutti, Massachusetts Materials Research, Inc.; William J. McBrine, ALTRAN Corporation

Introduction

ANALYZING FAILURES is a critical process in determining the physical root causes of problems. The process is complex, draws upon many different technical disciplines, and uses a variety of observation, inspection, and laboratory techniques. One of the key factors in properly performing a failure analysis is keeping an open mind while examining and analyzing the evidence to foster a clear, unbiased perspective of the failure. Collaboration with experts in other disciplines is required in certain circumstances to integrate the analysis of the evidence with a quantitative understanding of the stressors and background information on the design, manufacture, and service history of the failed product or system.

Just as failure analysis is a proven discipline for identifying the physical roots of failures, root-cause analysis (RCA) techniques are effective in exploring some of the other contributors to failures, such as the human and latent root causes. Properly performed, failure analysis and RCA are critical steps in the overall problem-solving process and are key ingredients for correcting and preventing failures, achieving higher levels of quality and reliability, and ultimately enhancing customer satisfaction.

This article briefly introduces the concepts of failure analysis, root-cause analysis, and the role of failure analysis as a general engineering tool for enhancing product quality and failure prevention. The discipline of failure analysis has evolved and matured, as it has been employed and formalized as a means for failure prevention. Consistent with the recent trend toward increased accountability and responsibility, its purpose has been extended to include determining which party may be liable for losses, be they loss of production, property damage, injury, or fatality. The discipline has also been used effectively as a teaching tool for new or less experienced engineers.

The importance and value of failure analysis to safety, reliability, performance, and economy are well documented. For example, the importance of investigating failures is vividly illustrated in the pioneering efforts of the Wright Brothers in developing self-propelled flight. In fact, while Wilbur was traveling in France in 1908, Orville was conducting flight tests for the U.S. Army Signal Corps and was injured when his Wright Flyer crashed (Fig. 1). His passenger sustained fatal injuries (Ref 1). Upon receiving word of the mishap, Wilbur immediately ordered the delivery of the failed flyer to France so that he could conduct a thorough investigation. This was decades before the formal discipline called “failure analysis” was introduced.

Fig. 1 Crash of the Wright Flyer, 1908. Courtesy of the National Air and Space Museum, Smithsonian Institution, Photo A-42555-A

Unfortunately, there are many dramatic examples of catastrophic failures that result in injury, loss of life, and damage to property. For example, a molasses tank failed in Boston in 1919, and *another* molasses tank failed in Bellview, NJ, in 1973 (Ref 2). Were the causes identified in 1919? Were lessons learned as a result of the accident? Were corrective actions developed and implemented to prevent recurrence?

Conversely, failures can also lead to improvements in engineering practices. The spectacular failures of the Liberty ships during World War II were studied extensively in subsequent decades, and the outcome of these efforts was a significantly more thorough understanding of the phenomenon of fracture, culminating in part with the development of the engineering discipline of fracture mechanics (Ref 3). Through these and other efforts, insights into the cause and prevention of failures continue to evolve.

References cited in this section

1. P.L. Jakab, *Visions of a Flying Machine: The Wright Brothers and the Process of Invention,* Smithsonian Institution, 1990, p 226

2. R.W. Hertzberg, *Deformation and Fracture Mechanics of Engineering Materials,* John Wiley & Sons, 1976, p 229–230

3. D.J. Wulpi, *Understanding How Components Fail,* 2nd ed., ASM International, 1999

Introduction to Failure Analysis and Prevention

James J. Scutti, Massachusetts Materials Research, Inc.; William J. McBrine, ALTRAN Corporation

---

Concepts of Failure Analysis and Prevention

Clearly, through the analysis of failures and the implementation of preventive measures, significant improvements have been realized in the quality of products and systems. This requires not only an understanding of the role of failure analysis, but also an appreciation of quality assurance and user expectations.

Quality and User Expectations of Products and Systems. In an era that initially gained global prominence in the 1980s, corporations, plants, government agencies, and other organizations developed new management systems and processes aimed at improving quality and customer satisfaction. Some of these systems include Total Quality Management (TQM), Continuous Improvement (CI), and, more recently prominent, Six Sigma. Historically, these initiatives are founded on the philosophies of the quality visionaries W. Edwards Deming (Ref 4) and Joseph Juran (Ref 5).

In their most basic descriptions, TQM and CI represent full organizational commitment to a system focused on “doing the right thing right the first time” and not merely meeting but exceeding customer requirements (Ref 6, 7). They are focused on process improvements, generally in a production environment. Six Sigma adopts these themes and extends the “reach” of the system to all levels of organizations, with a system to achieve, sustain, and maximize business success (Ref 8). Six

Sigma is founded on the use of measurements, facts, and statistics to move organizations in directions that constantly improve and reinvent business processes (Ref 8). The roots of this business system are in the statistical limits set for the maximum number of defects in a product, as a fraction of the total number of opportunities for such defects to occur. To the practitioners of this system, “six sigma” is a statistical metric referring to six times the statistical standard deviation of a normal distribution, which allows no more than 3.4 defects per million opportunities (equivalent to 99.9997% reliability). This is indeed a lofty goal for any organization (be it a manufacturing company, a petrochemical plant, a service business, or a government agency), but companies committed to Six Sigma have reported significant gains in productivity with simultaneous improvements in organizational culture (Ref 7, 8, 9).

The most positive result of these new management systems is that organizations have responded to the higher expectations of consumers and users and have provided higher-quality products and systems, with attendant increases in customer satisfaction. However, this notion of the *quality* of a product or system is multifaceted. Juran described quality as “fitness for use” (Ref 5). TQM defines quality as the ability to satisfy the needs of a consumer (Ref 10). These characteristics of quality also apply internally to those in organizations, either in the services, or in manufacturing, operating, or administering products, processes, and systems (Ref 10). The intent is to provide not only products and systems that garner high customer satisfaction, but also that increase productivity, reduce costs, and meet delivery requirements.

In general, high quality refers to products and systems manufactured to higher standards, in response to higher expectations of consumers and users. These expectations include such attributes as:

- Greater safety
- Improved reliability
- Higher performance
- Greater efficiency
- Easier maintenance
- Lower life-cycle cost
- Reduced impact on the environment

Some or all of these qualities at one time appeared mutually exclusive. However, customer demands and the aforementioned new business-management systems have provided a means of measuring and quantifying these attributes, creating a new paradigm for business. With the business-culture changes that have occurred through the implementation of one or more of the aforementioned improvement systems, users in recent years have experienced, in general, improvement in all of these areas simultaneously. That translates to reduced product failure and greater likelihood of preventing failures. It is important to recognize that, with all the gains achieved under these management systems, the full potential for maximizing these attributes is yet to be achieved.

Though all of the various improvement systems are unique, they have two aspects in common. They are all *customer focused* and are founded on *problem solving* as a means for improvement.

When addressing customer focus, producers and other organizations have identified that the form, fit, function, and service-life requirements of a product or system are actually defined ultimately by customers. Customer-focused manufacturers strive to meet these requirements in designing, developing, and producing their products or systems. In a broad sense, form, fit, function, and service life represent the technically relevant properties of a product. The form, or physical characteristics of components or products, include the size and shape of a product, as well as the materials of construction and the manufacturing techniques used. The manner in which individual components are assembled into and integrate with the product as a whole describes the fit of components. The function of a product or system is its ability or capability to serve the need for which it was intended. Service life is the duration over which the product or system successfully serves its function. These characteristics define products in the customer's eyes. Arguably the most important characteristics, from a consumer's perspective, are how well a product or system functions and how long it serves a useful life.

Problem Solving, Quality, and Customer Satisfaction. Achieving the levels of quality that meet and exceed customer expectations is paramount to customer satisfaction in a customer-focused management system. Since a customer's perspective of quality is strongly tied to the function and service life of a product or system, it follows that failure to provide adequate measures of function and service life presents problems. One proven technique to improving quality is problem solving. Problems can range broadly, from maintenance training issues, to marginal equipment reliability, to business systems conflicts, to policy inconsistencies, to poor working conditions on the shop floor. When a problem occurs, the responsible organization will analyze the problem to determine the cause and solve it. However, due to various business or cultural pressures, some organizations fall into the following pitfalls when problems arise (Ref 9):

- Do nothing and perhaps hope that the problem will go away.
- Deny that the problem exists, minimize its importance, question the motives of those identifying the problem.
- Troubleshoot in a haphazard fashion (i.e., “shotgun” troubleshooting).

- Chase false leads (i.e., “red herrings”).

In an enlightened organizational culture, products or systems require a systematic approach to problem solving, based on analysis, to achieve the levels of quality and customer satisfaction defined by the new management systems. The cultural aspect is critical, as those who have identified problems must be encouraged to come forward. Furthermore, resources and commitment are required to formulate the solutions and implement necessary changes.

Problem-Solving Models. A wide range of problem-solving methods and models are available in the literature (Ref 4, 5, 6, 8, 9, 10, 11, 12), presenting various details of approaches and processes for solving any of the general types of problems defined previously. All of these methods and models are rooted in the scientific method (summarized as follows) (Ref 6):

1. Define the issue
2. Propose a hypothesis
3. Gather data
4. Test the hypothesis
5. Develop conclusions

A concise problem-solving model, adapted from several of the referenced authors, and that has specific applicability to this Volume, is depicted in Fig. 2. The continuous, circular format in the graphic is significant, indicating that the process reinitiates with the identification of a new problem or problems brought to light as a result of the first problem-solving activity. Note the similarity to the classical scientific method summarized previously.

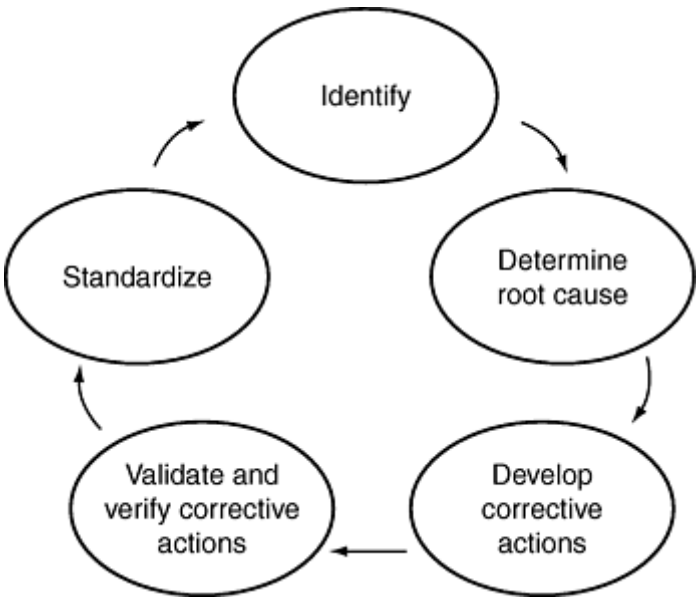


Fig. 2 Problem-solving model

The major steps in the model define the problem-solving process:

1. *Identify:* Describe the current situation. Define the deficiency in terms of the symptoms (or indicators). Determine the impact of the deficiency on the component, product, system, and customer. Set a goal. Collect data to provide a measurement of the deficiency.
2. *Determine root cause:* Analyze the problem to identify the cause(s).
3. *Develop corrective actions:* List possible solutions to mitigate and prevent recurrence of the problem. Generate alternatives. Develop implementation plan.
4. *Validate and verify corrective actions:* Test corrective actions in pilot study. Measure effectiveness of change. Validate improvements. Verify that problem is corrected and improves customer satisfaction.
5. *Standardize:* Incorporate the corrective action into the standards documentation system of the company, organization, or industry to prevent recurrence in similar products or systems. Monitor changes to ensure effectiveness.

The second step in the problem-solving model, *determine root cause,* introduces a very significant process. Solutions to prevent recurrence of problems cannot be developed without identification of the root cause.

Failure Definitions. In the general sense of the word, a failure is defined as an undesirable event or condition. For the purposes of discussion related to failure analysis and prevention, it is a general term used to imply that a component is unable to adequately perform its intended function. The intended function of a component and therefore the definition of failure may range greatly. For instance, discoloration of an architectural feature is a failure of its intended aesthetic function.

Failure can be defined on several different levels. The simplest form of a failure is a system or component that operates, but does not perform its intended function (Ref 13). This is considered a *loss of function.* A jet engine that runs but can only produce partial thrust (insufficient to enable an aircraft to take off) is an example of a loss of function.

The next level of failure involves a system or component that performs its function but is unreliable or unsafe (Ref 13). In this form of failure, the system or component has sustained a *loss of service life.* For example, a wire rope for an elevator has lost its service life when it has sustained fatigue fractures of some of the individual wires, due to irregularities in the wrapping over the sheave. Even though the wire rope continues to function, the presence of fatigue fractures of some of the wires results in an unsafe condition and is therefore considered a failure. Another example of such a failure is the inability of an integrated circuit to function reliably.

In the next level of severity of failure, a system or component is inoperable (Ref 13), such as a pump shaft fracture that causes the impeller to seize or a loss of load-carrying capability of a structural bolt in-service due to fracture.

Failure and Failure Analysis. A logical failure analysis approach first requires a clear understanding of the failure definition and the distinction between an indicator (i.e., symptom), a cause, a failure mechanism, and a consequence. Although it may be considered by some to be an exercise in semantics, a clear understanding of each piece of the situation associated with a failure greatly enhances the ability to understand causes and mitigating options and to specify appropriate corrective action.

Consider the example of a butterfly valve that fails in service in a cooling water system at a manufacturing facility (Table 1). Recognizing the indicators, causes, mechanisms, and consequences helps to focus investigative actions:

- *Indicators(s):* Monitor these as precursors and symptoms of failures.
- *Cause(s):* Focus mitigating actions on these.
- *Failure mechanism(s):* These describe how the material failed according to the engineering textbook definitions. If the analysis is correct, the mechanism will be consistent with the cause(s). If the mechanism is not properly understood, then all true cause(s) will not be identified and corrective action will not be fully effective.
- *Consequence(s):* This is what we are trying to avoid.

Table 1 Example—Failure of a butterfly valve in a manufacturing plant cooling water system

Item	Description	Indicators
Cause	Throttling of valve by the operator outside of the design parameters	Flow gages and records Operator logs
	Low-strength copper nickel alloy construction	Material specifications Laboratory analysis
	Flow-induced cavitation	Rumbling noise in system Vibration of system
Failure mechanism	Erosion-fatigue damage	Laboratory examination of disk, thinning
Consequences	Inability to manufacture at normal production rates	

Life-Cycle Management Concepts. The concept of life-cycle management refers to the idea of managing the service life of a system, structure, or component. There is a cost associated with extending the service life of a component, for example, higher research costs, design costs, material and fabrication costs, and higher maintenance costs. With regard to product failures, it must be understood that failures cannot be totally avoided, but must be better understood, anticipated, and controlled. Nothing lasts and functions forever. For some products, consumers may prefer a shorter life at a more modest cost. In contrast, the useful service life of a product such as an aircraft part may be carefully planned in advance and managed accordingly with routine inspections and maintenance, which may increase in frequency over time. In many cases, avoiding failures beyond a certain predetermined desired life provides no benefit, such as is the case when a surgical implant is designed to far outlive the human recipient. There is also a point of diminishing return on investments related to extending the life of a component. A life-cycle management study of a component would look at these issues as well as other factors such as the issue of obsolescence. How long will it be before the product is obsolete?

Understanding how the typical distribution of failure for a given product must be factored with time is also important when looking at failure patterns (Fig. 3). Early life failures are often associated with fabrication issues, quality-control issues, or initial "shakedown" stresses, while age-related failure rates would increase with time. This is discussed in more detail in the article "Reliability-Centered Maintenance" in this Volume.

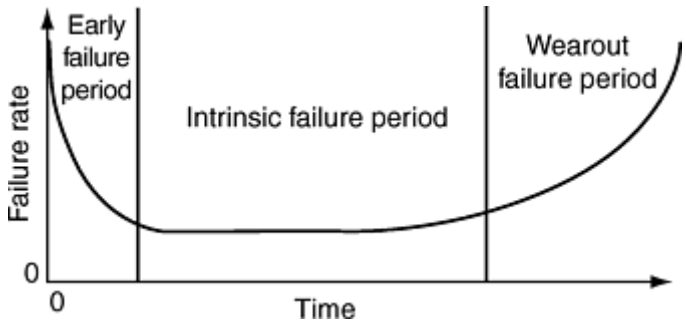


Fig. 3 Typical time distribution of failures ("bathtub curve")

Once the concept of a managed life is prudently adopted over a simple failure prevention concept, design and fabrication costs can be reduced and maintenance and other life-prolonging activities can be optimized.

Diligence in Use of Terminology. Communicating technical information is of paramount importance in all engineering areas, including failure analysis. The choice of technical descriptors, nomenclature, and even what might be considered technical jargon is critical to conveying technical ideas to other engineers, managers, plant personnel, shop personnel, maintenance personnel, attorneys, a jury, and so forth. It is instructive in this introductory article to emphasize that a descriptor can mean something very specific to a technical person and mean something very different to a business manager or an attorney.

For example, the term "flaw" is synonymous with "defect" in general usage. However, to a fracture mechanics specialist, a flaw is a discontinuity such as a crack. Under some circumstances, when the crack is smaller than the critical size (i.e., subcritical), the crack is benign and therefore may not be considered a defect. To the quality-control engineer, flaws are characteristics that are managed continuously on the production line, as every engineered product has flaws, or "deviations from perfection" (Ref 14). On the manufacturing floor, these flaws are measured, compared with the preestablished limits of acceptability, and dispositioned as acceptable or rejectable. A rejectable characteristic is defined as a defect (Ref 14). To the Six Sigma practioner, a defect is considered anything that inhibits a process or, in a broad sense, any condition that fails to meet a customer expectation (Ref 9). To the attorney, a defect refers to many different types of deficiencies, including improper design, inadequate instructions for use, insufficient warnings, and even inappropriate advertising or marketing (Ref 15).

Similar nuances may occur in the basic definitions and interpretations of technical terms used in materials failure analysis. Terms such as ductile and brittle, crack and fracture, and stable and unstable crack growth are pervasive in failure analysis. Even these seemingly basic terms are subject to misuse and misinterpretations, as suggested in Ref 16—for example "brittle cleavage," which is a pleonasm that does not explain anything. Another example noted in Ref 16 is the term "overload fracture," which may be misinterpreted by nonanalysts as a failure caused by a load higher than anticipated by the materials or mechanical engineers. This limited interpretation of overload failure is incomplete, as described in the article "Overload Failures" in this Volume.

Judgmental terminology should be used with prudence when communicating analytical protocols, procedures, findings, and conclusions. Communications during the preliminary stages of an investigation should be factual rather than judgmental. It is important to recognize that some of the terminology used in a failure analysis can be judgmental, and consideration must be given to the implications associated with the use of such terminology. For example, when examining both a failed and an unfailed component returned from service, references to the unfailed sample as "good" and the failed sample as "bad" should be avoided. This is because the investigation may reveal both samples to contain the same defect, and therefore both could be considered "bad." Similarly, ***neither*** may be "bad" if the analysis actually indicates the failed component met all requirements but was subjected to abuse in service. On completion of the failure analysis, judgmental terminology is often appropriate to use if the evidence supports it, such as in the example of a casting defect that has been confirmed in the example bolt failure analysis.

While discussions of the semantics of terminology may seem pedantic, communicating the intended information gleaned from a failure analysis relies heavily on precision in the use of language.

References cited in this section

4. W.E. Deming, *Out of the Crisis*, MIT Center for Advanced Engineering Study, 1986

5. J.M. Juran and F.M. Gryna, Ed., *Juran's Quality Control Handbook*, 4th ed., McGraw-Hill, 1988

6. P.F. Wilson, L.D. Dell, and G.F. Anderson, *Root Cause Analysis: A Tool for Total Quality Management*, ASQ Quality Press, 1993, p 7

7. F.W. Breyfogle III, *Implementing Six Sigma: Smarter Solutions Using Statistical Methods*, John Wiley & Sons, 1999, p xxvii

8. P.S. Pande, R.P. Neuman, and R.R. Cavanaugh, *The Six Sigma Way*, McGraw-Hill, 2000, p xi

9. M. Harry and R. Schroeder, *Six Sigma: The Breakthrough Management Strategy Revolutionizing the World's Top Corporations*, Doubleday & Co., Inc., 1999

10. G.F. Smith, *Quality Problem Solving*, ASQ Quality Press, 1998, p 7

11. B. Anderson and T. Fagerhaug, *Root Cause Analysis: Simplified Tools and Techniques*, ASQ Quality Press, 2000, p 7, 125

12. M. Ammerman, *The Root Cause Analysis Handbook: A Simplified Approach to Identifying, Correcting, and Reporting Workplace Errors*, Max Ammerman/Quality Resources, 1998

13. Engineering Aspects of Failure and Failure Analysis, *Failure Analysis and Prevention*, Vol 10, 8th ed., *Metals Handbook*, American Society for Metals, 1975, p 1–9

14. R.K. McLeod, T. Heaslip, and M. Vermij, Defect or Flaw—Legal Implications, *Failure Analysis: Techniques and Applications*, Conf. Proc. International Conference and Exhibits on Failure Analysis, 8–11 July 1991 (Montreal, Quebec, Canada), ASM International, 1992, p 253–261

15. J.J. Asperger, Legal Definition of a Product Failure: What the Law Requires of the Designer and the Manufacturer, *Proc. Failure Prevention through Education: Getting to the Root Cause*, 23–25 May 2000 (Cleveland, OH), ASM International, 2000, p 25–29

16. D. Broek, Fracture Mechanics as an Important Tool in Failure Analysis, *Failure Analysis: Techniques and Applications*, Conf. Proc. International Conference and Exhibits on Failure Analysis, 8–11 July 1991 (Montreal, Quebec, Canada), ASM International, 1992, p 33–44

James J. Scutti, Massachusetts Materials Research, Inc.; William J. McBrine, ALTRAN Corporation

Root-Cause Analysis

Failure analysis is considered to be the examination of the characteristics and causes of equipment or component failure. In most cases this involves the consideration of physical evidence and the use of engineering and scientific principles and analytical tools. Often, the reason why one performs a failure analysis is to characterize the causes of failure with the overall objective to avoid repeat of similar failures. However, analysis of the physical evidence alone may not be adequate to reach this goal. The scope of a failure analysis can, but does not necessarily, lead to a correctable root cause of failure. Many times, a failure analysis incorrectly ends at the identification of the failure mechanism and perhaps causal influences. The principles of root-cause analysis (RCA) may be applied to ensure that the root cause is understood and appropriate corrective actions may be identified. An RCA exercise may simply be a momentary mental exercise or an extensive logistical charting analysis.

Many volumes have been written on the process and methods of RCA. The concept of RCA does not apply to failures alone, but is applied in response to an undesirable event or condition (Fig. 4). Root-cause analysis is intended to identify the fundamental cause(s) that if corrected will prevent recurrence.

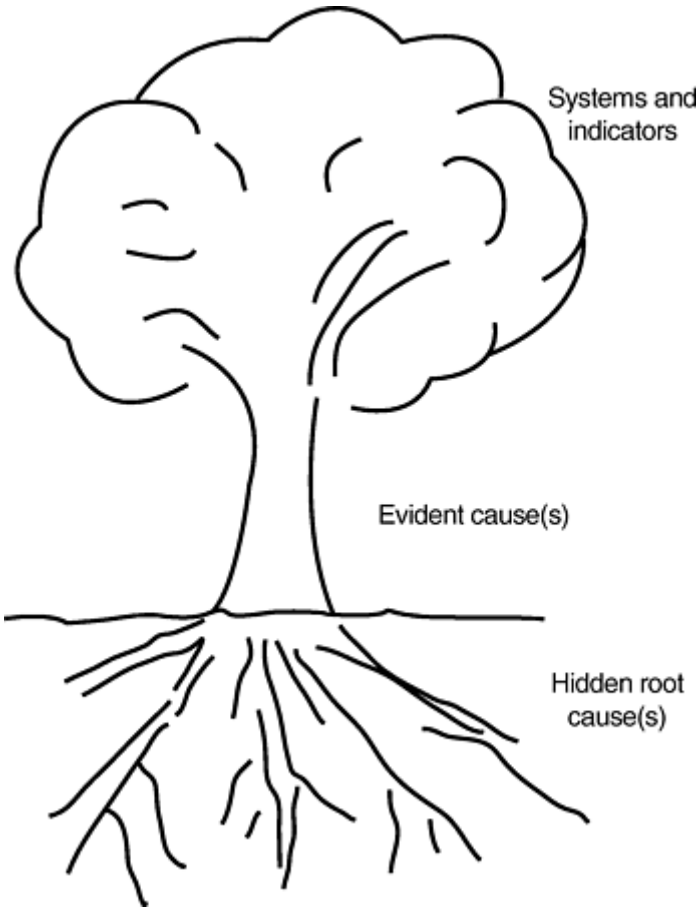


Fig. 4 Root-cause analogy

Levels. The three levels of root-cause analysis are physical roots, human roots, and latent roots (Ref 17, 18, 19, 20, 21). Physical roots, or the roots of equipment problems, are where many failure analyses stop. These roots may be what comes out of a laboratory investigation or engineering analysis and are often component-level or materials-level findings. Human roots (i.e., people issues) involve human factors that caused the failure, an example being an error in human judgment. Latent roots lead us to the causes of the human error and include roots that are organizational or procedural in nature, as well as environmental or other roots that are outside the realm of control. These levels or root cause are best defined by the two examples in Table 2.

Table 2 Examples of root causes of failure of pressure vessel and bolt

Root type	Pressure vessel failure	Bolt failure
Physical roots	Corrosion damage, wall thinning	Fatigue crack; equipment vibration; lack of vibration; isolation
Human roots	Inadequate inspection performed	Improper equipment installed
Latent roots	Inadequate inspector training	Inadequate specification verification process

How deeply one goes into the root causes depends on the objectives of the RCA. These objectives are typically based on the complexity of the situation and the risk associated with additional failures. In most cases, one desires to identify root causes that are reasonably correctable. An example of the variety of possible root causes of an electric motor driven compressor assembly is provided in Table 3 (Ref 22).

Table 3 Possible causes of electric motor driven pump or compressor failures

System design and specification responsibility	Component manufacturer's responsibility	Shipping and storage responsibility	Installation responsibility	Operations and maintenance responsibility	Distress damage or failed components
Application	**Material of construction**	**Preparation for shipment**	**Foundations**	**Shock**	**Distress damages**
Undercapacity	Flaw or defect	Oil system not clean	Settling	Thermal	Vibration
Overcapacity	Improper material	Inadequate	Improper or insufficient grouting	Mechanical	Short/open circuit
Incorrect physical				Improper startup	**Failed components**

System design and specification responsibility	Component manufacturer's responsibility	Shipping and storage responsibility	Installation responsibility	Operations and maintenance responsibility	Distress damage or failed components
condition assumed (temperature, pressure, etc.) Incorrect physical property assumed (molecular weight, etc.) **Specifications** Inadequate lubrication system Insufficient control instrumentation Improper coupling Improper bearing Improper seal Insufficient shutdown devices **Material of construction** Corrosion and/or erosion Rapid wear Fatigue Strength exceeded Galling Wrong hardening method **Design for installation** Unsatisfactory piping support Improper piping flexibility Undersized piping Inadequate foundation	Improper treatment **Design** Improper specification Wrong selection Design error Inadequate or wrong lubrication Inadequate liquid drain Critical speed Inadequate strength Inadequate controls and protective devices **Fabrication** Welding error Improper heat treatment Improper hardness Wrong surface finish Imbalance Lube passages not open **Assembly** Improper fit Improper tolerances Parts omitted Parts in wrong Parts/bolts not	drainage Protective coating not applied Wrong coating used Equipment not cleaned **Protection** Insufficient protection Corrosion by salt Corrosion by rain or humidity Poor packaging Desiccant omitted Contamination with dirt, etc. **Physical damage** Loading damage Transport damage Insufficient support Unloading damage	Cracking or separating **Piping** Misalignment Inadequate cleaning Inadequate support **Assembly** Misalignment Assembly damage Defective material Inadequate bolting Connected wrong Foreign material left in General poor workmanship	**Operating** Slugs of liquid Process surging Control error Controls deactivated/not installed Operating error **Auxiliaries** Utility failure Insufficient instrumentation Electronic control failure Pneumatic control failure **Lubrication** Dirt in oil Insufficient oil Wrong lubricant Water in oil Oil pump failure Low oil pressure Plugged lines Improper filtration Contaminated oil **Craftsmanship after maintenance** Improper tolerances	Sleeve bearing Seal Coupling Shaft Pinion/ball/turning gear Casing Rotor Impeller Shroud Piston Diaphragm Wheel Blades; foil, root, shroud Labyrinth Thrust bearing Pivoted pad bearing Roller/ball bearing Cross-head piston Cylinder Crankshaft

System design and specification responsibility	Component manufacturer's responsibility	Shipping and storage responsibility	Installation responsibility	Operations and maintenance responsibility	Distress damage or failed components
Unsatisfactory soil data Liquid ingestion Inadequate liquid drain Design error	tight Poor alignment Imbalance Inadequate bearing contact Inadequate testing			Welding error Improper surface finish Improper fit General poor workmanship **Assembly after maintenance** Mechanical damage Parts in wrong Parts omitted Misalignment Improper bolting Imbalance Piping stress Foreign material left in Wrong material of construction **Preventive maintenance** Postponed Schedule too long	

Requirements for Effective RCA. Performing an effective RCA requires an interdisciplinary approach in order to ensure that the results are correct and proper corrective actions are identified. In fact, most failures involve factors that spread across many disciplines such as metallurgy, mechanical engineering, hydraulics, electrical engineering, quality control, operations, maintenance, human factors, and others. The analysis team on a complex failure will ideally represent a spectrum of expertise to ensure a very broad perspective.

The best analysis team leader must be a good communicator, have a broad background, be able to integrate factors, and be able to select the best expertise for the project. On less complex failures it is often beneficial to have an individual with a diverse background participate in addition to the specialists, once again to ensure a broader perspective. For example, a metallurgist may be more likely to report a metallurgical deficiency in a product that contributed to the failure, a fabricator is more likely to point to fabrication-related contributors, and a designer is more likely to identify design deficiencies. All of these may be important considerations, but one, all, or none may be a primary root cause. Problems related to people, procedures, environmental concerns, and other issues can also be treated effectively by conducting problem-solving processes and RCAs (although the main focus of this article and this Volume is on materials failure analysis).

References cited in this section

17. P.F. Wilson, L.D. Dell, and G.F. Anderson, *Root Cause Analysis: A Tool for Total Quality Management,* ASQ Quality Press, 1993, p 50

18. G.F. Smith, *Quality Problem Solving,* ASQ Quality Press, 1998, p 127

19. M. Paradise, L. Unger, and D. Busch, *TapRoot® Root Cause Tree™ User's Manual,* Systems Improvement, Inc., 1996, p 9–14

20. R.J. Latino and K.C. Latino, *Root Cause Analysis: Improving Performance for Bottom Line Results,* Reliability Center, Inc., 1999, p 79–89

21. C. Nelms, *What You Can Learn From Things That Go Wrong,* 1st ed., Failsafe Network, Richmond, VA, 1994

22. H.P. Bloch and F.K. Geitner, *Practical Machinery Management for Process Plants,* Vol 2, *Machinery Failure Analysis and Troubleshooting,* Gulf Publishing Co., 1983, p 5–6

James J. Scutti, Massachusetts Materials Research, Inc.; William J. McBrine, ALTRAN Corporation

Primary Physical Root Causes of Failure

Categorizing schemes for the root causes of equipment failures vary among failure analysis practitioners, quality engineers, other engineers, and managers, as well as legal and insurance professionals (Ref 13, 15, 23, 24, 25, 26, 27). Grouping physical root causes into only a few fundamental categories is advantageous and informative because it defines which aspect of a product or system requires corrective action and prevention strategies. Systematic analysis of equipment failures reveals physical root causes that fall into one of four fundamental categories (Ref 28):

- Design deficiencies
- Material defects
- Manufacturing/installation defects
- Service life anomalies

An effective graphical representation of the impact of defects on the service life of a component or system is provided in the application-life diagram (Fig. 5) (Ref 29, 30). The diagram is constructed by plotting the service lives of components having specific characteristics in the design/configuration, as related to the severity of a specific service condition that is anticipated for the application. Typical characteristics include strength, corrosion resistance, heat treatment condition, flaw size, surface finish, bend radius, void content (i.e., in a casting), degree of sensitization, and so forth. Examples of service conditions include magnitude of stress (either cyclic or static), exposure temperature, aggressiveness of environment, radiation exposure, electrical stress, and so forth.

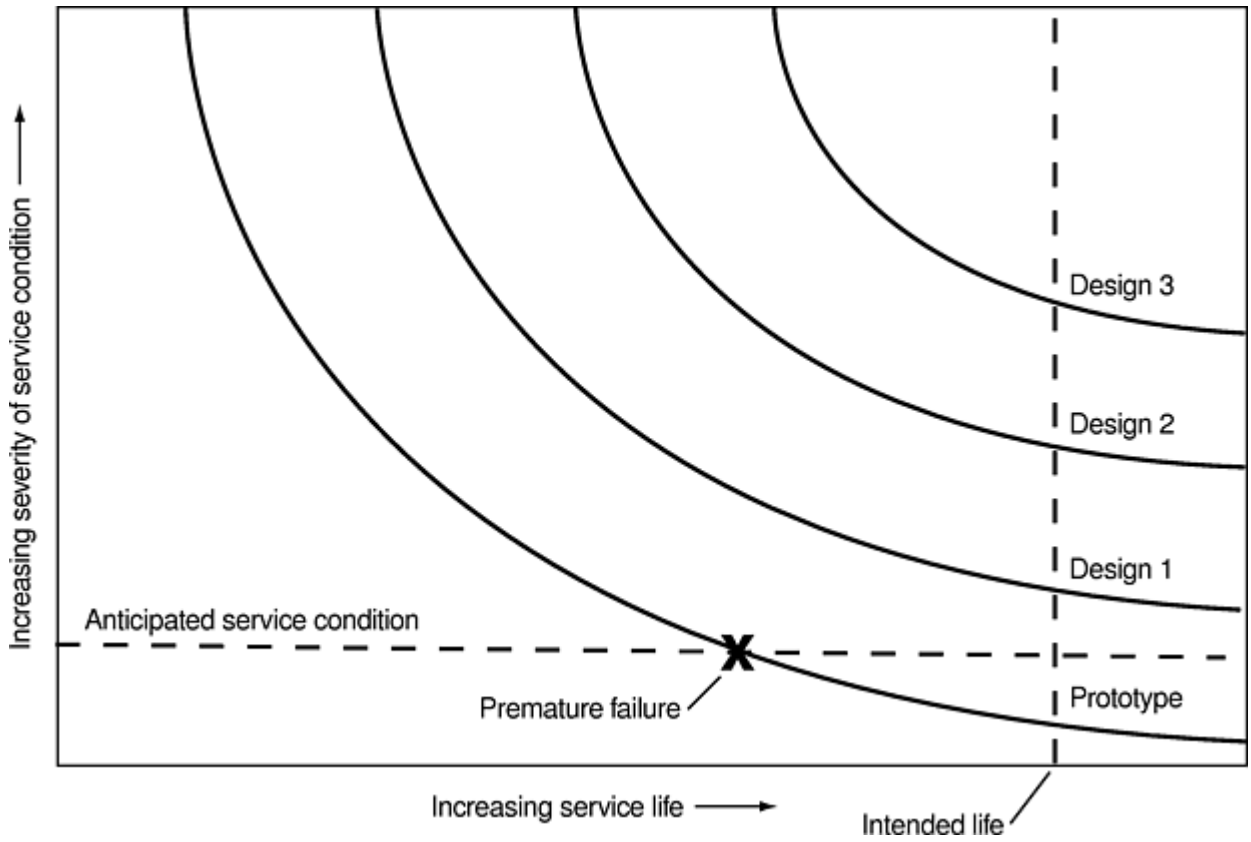


Fig. 5 Application-life diagram comparing the severity of a service condition with the service lives of products having a variable characteristic. This diagram is utilized in specific examples in text.

By varying the characteristics, a family of curves is generated, contrasting the lives of components with the various characteristics and service conditions with the intended service life. Each of the curves represents a different design/configuration characteristic, with increasing degrees of durability as the curves move up the ordinate. Failures can be prevented when the curve for a specific design/configuration lies above the severity of service line, and to the left of the intended service life line. However, if the severity of service conditions increases (either intentionally during operation or as a result of some other change in the system), the propensity for failure may increase, since the characteristics curves intersect the severity of service condition line "to the left," that is, at an earlier point in the service life.

Design Deficiencies

Root causes of failures that stem from design deficiencies refer to unacceptable features of a product or system that are a result of the design process. This process encompasses the original concept development, the general configuration definition, and the detail design, including selection and specification of materials and manufacturing processes. Design involves identifying and defining a need for the product or system, followed by definition of the performance requirements, anticipated service conditions in the application(s), the constraints on the design, and the criticality or risks associated with failure (Ref 31). Discussion of the design process as it relates to failure analysis and prevention is provided in the article "Design Review for Failure Analysis and Prevention" in this Section.

Some examples of design deficiencies include unintended stress raisers due to excessively sharp notches (Ref 32) (e.g., in keyways on shafts) or insufficient radii (e.g., on shafts at bearing journals). Other examples include unanticipated residual stresses associated with heat treating configurations designed with complex geometries, or assembly stresses from configurations that contain unwanted interference. Inappropriate surface treatments could result in failures, such as the use of cadmium plating on an A286 superalloy fastener, subjected to service temperatures above 315 °C (600 °F) (the melting temperature of cadmium is 320 °C, or 610 °F). Two metals specified for use in a wear application could sustain galling if the metals are incompatible, such as sliding wear of components made from 300 series stainless steels.

Selection of a material that is incapable of providing adequate mechanical properties for the application (including strength, fatigue resistance, fracture toughness, corrosion resistance, elevated temperature resistance, etc.) is also a type of design deficiency. Materials can exhibit anisotropy, or variability in properties within a product, such as between the thick

and thin portions of a casting, or between longitudinal and transverse properties in a wrought material. Note that a material can be shown to meet the properties required or specified (i.e., a separately cast tensile bar used to certify a casting, or the longitudinal tensile properties to certify a complex aluminum extrusion), but the specific properties required for the application may rely on the strength, toughness, or stress-corrosion cracking resistance in a direction other than longitudinal.

Design-caused failures include inappropriate geometries (as defined on the engineering drawing), which may lead to a compromise of component or system capabilities. Examples of inappropriate geometries include improper joint preparation for welding or brazing, such as an insufficient or missing groove for a groove weld, insufficient fit-up relief in a socket weld, or inadequate joint overlap in a brazed joint. Other geometry-caused failures can result from insufficient section thickness for a failure based on gross yielding, excessive section thickness in the presence of a flaw for a material of limited fracture toughness, or a fabrication configuration with an excessively sharp forming bend, with the resulting high residual stresses causing a reduction in the fatigue life.

For the example of the excessively tight cold-formed bend radius described previously, an application-life diagram can be constructed as shown in Fig. 6. The service condition considered is stress, and the characteristic that is varied is the radius of the cold-formed bend. Upon examination of the relationship between the characteristic curves and the intended service life, the components having the large and moderate bend radii are found to meet the intended service life at the severity of stress that is anticipated in the specific application. However, in this illustration, the component with the small bend radius sustained a premature failure at the anticipated stress level in the application, since the curve intersects the severity of stress line prior to reaching the intended service life.

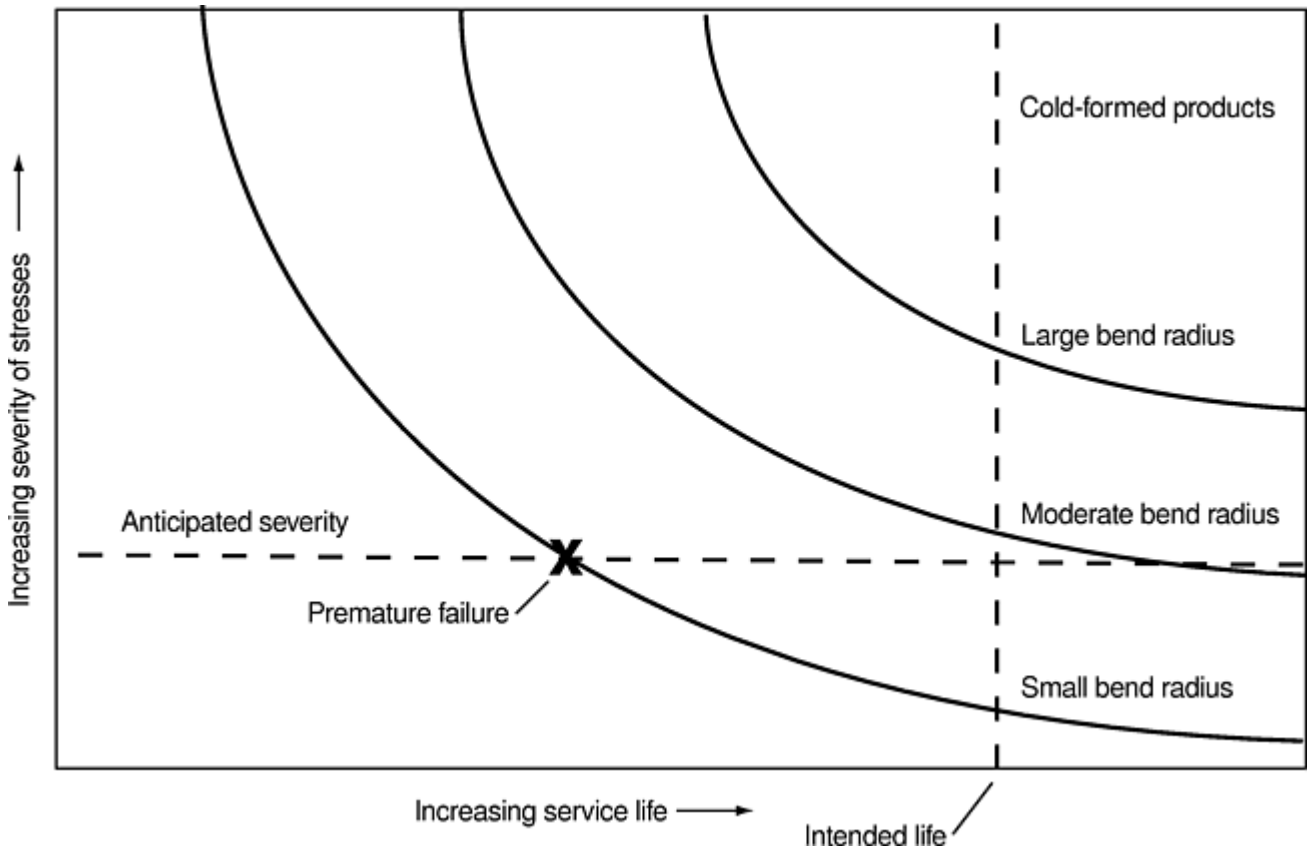


Fig. 6 Application-life diagram for design deficiency

Some of the aforementioned deficiencies in design as well as application-life diagram concepts are illustrated in the following two case histories.

Example 1: Ice Cream Drink Mixer Blade Failures. Excessive assembly stresses and inappropriate detail design caused the premature failures of ice cream drink mixer blades shortly after the mixing machines were introduced into service. A mixer blade as-manufactured is shown on the left side of Fig. 7. As assembled (right side of Fig. 7), the mixer blade is slightly deformed by the contact between the wavy washer at the bottom of the assembly and the bends at the bottom shoulders of the two mixer arms. When properly torqued, the screw that fastens the wavy washer and the mixer blade to the spindle in the center of the assembly places an upward force on the bottoms of the arms (as indicated by the pair of upward facing arrows in Fig. 7). This results in the observed inward deflection of the arms (as indicated by the right and left facing arrows). More significantly, this bending force places the inside radii of the two shoulders of the mixing blade

arms (at the bottom of the blade) in tension. When the mixer is running, the rotational forces further add to the tensile loads on the inside radii of the shoulders.

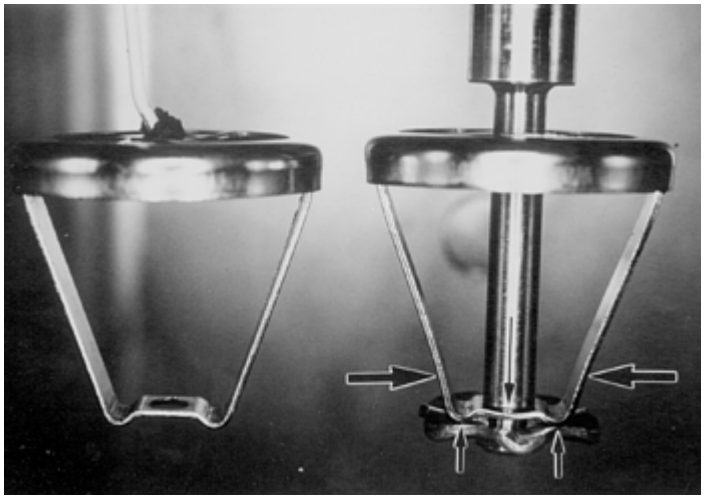


Fig. 7 Ice cream mixer blade as manufactured (left) and assembled to spindle (right)

Analysis of the failed mixer blades revealed multiple fatigue crack origins on the inside radii of the bends at the bottom shoulders (Fig. 8). Metallographic examination of the arm materials revealed additional problems with the configuration: the shoulders on the arms were cold bent, introducing tensile residual stresses on the inside radii of the shoulders and creating a localized area of fatigue susceptibility due to the inherent notch sensitivity of cold-formed 300 series stainless steel.

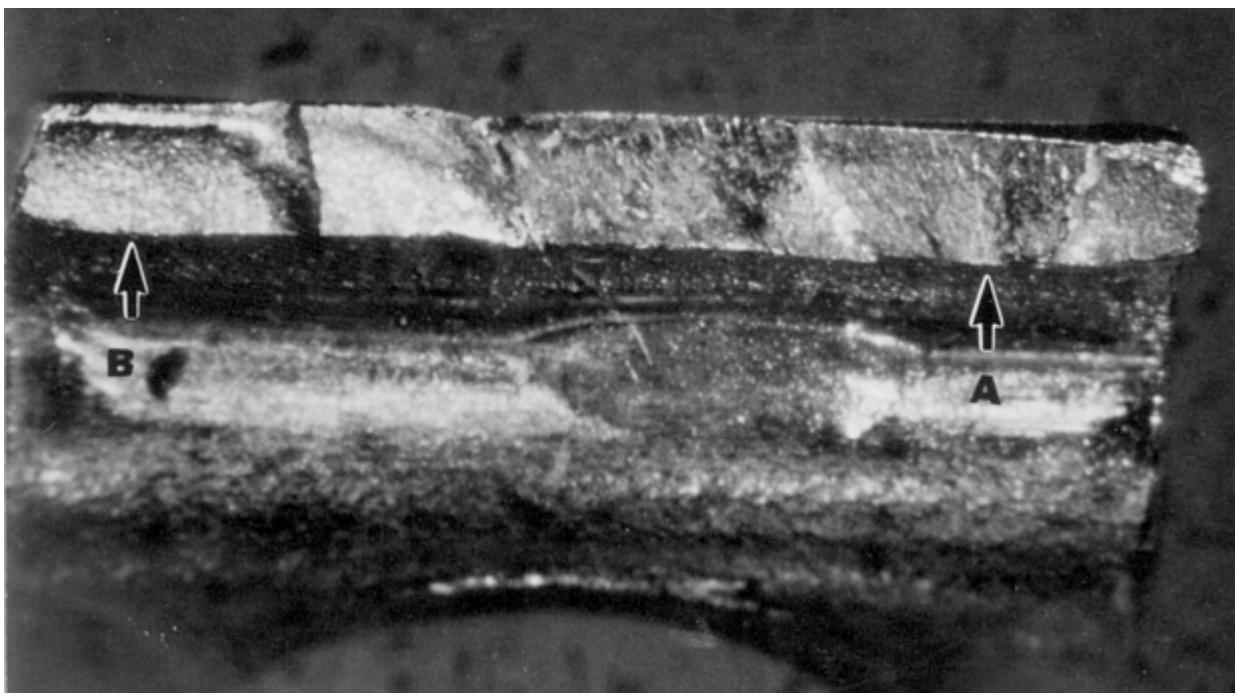


Fig. 8 Fracture surface of failed ice cream mixer blade. Arrows indicate fatigue crack origins. 13×

Clearly, the physical root cause is the design of the mixer blade, which defined two bend areas that contained tensile residual stresses, tensile assembly stresses, and a notch-sensitive microstructure that added to the normal operating rotational and vibratory stresses. The net effect was a reduction in the life of the blade causing loss of function. Corrective-action recommendations included the addition of a stand-off washer between the wavy washer and the bottom shoulders of the blade, or modification of the shape of the wavy washer to prevent contact with the blade shoulders as assembled.

Example 2: Sprocket Locking Device Failure. (Ref 33). A design deficiency involving improper materials selection was revealed through the analysis of a failed tapered-ring sprocket locking device. The device is used to attach a chain sprocket to a shaft without the use of a locking key, enabling the shaft to either drive or be driven anywhere on the shaft (see Fig. 9). The configuration consists of an assembly of four tapered rings (Fig. 10) that are retained by a series of cap screws. As shown in Fig. 11, when the screws are tightened, the middle wedge-shaped rings are pulled closer, forcing the split inner ring to clamp tightly onto the shaft, and the split outer ring to force tightly against the inside diameter of the sprocket. When properly assembled and torqued, the sprocket is fixed to the shaft.

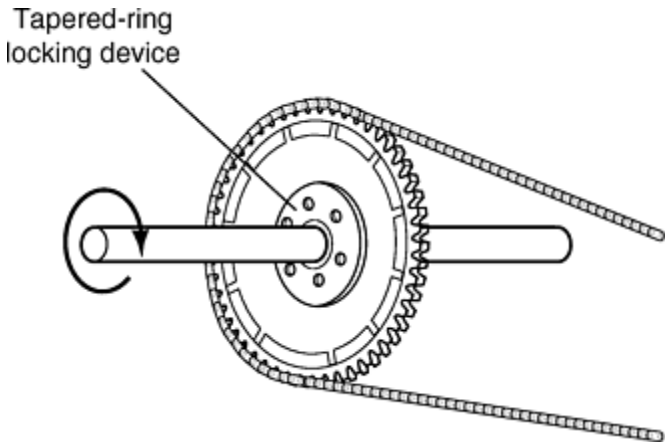


Fig. 9 Sketch of tapered-ring locking device application

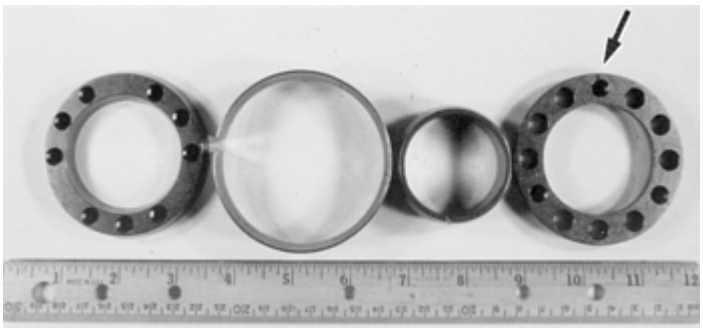


Fig. 10 Four tapered rings of locking device. Arrow indicates crack in one of the middle rings.

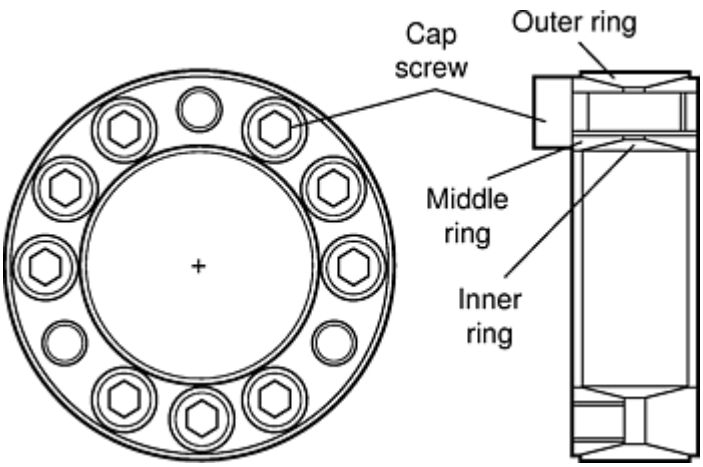


Fig. 11 Plan view (left) and cross section (right) through tapered-ring locking device assembly.

During initial assembly of a new locking device by the manufacturer during a bench test, one of the wedge-shaped middle rings fractured prior to having been fully torqued, preventing the sprocket from being locked to the shaft. The failed assembly was investigated for root cause. One of the middle rings had cracked (Fig. 10, 12a). Examination of the fracture

revealed “woody” fracture features (Fig. 12b), as a result of decohesion between a high volume fraction of manganese sulfide stringers and the matrix (Fig. 13). The matrix fracture features showed ductile dimple rupture.

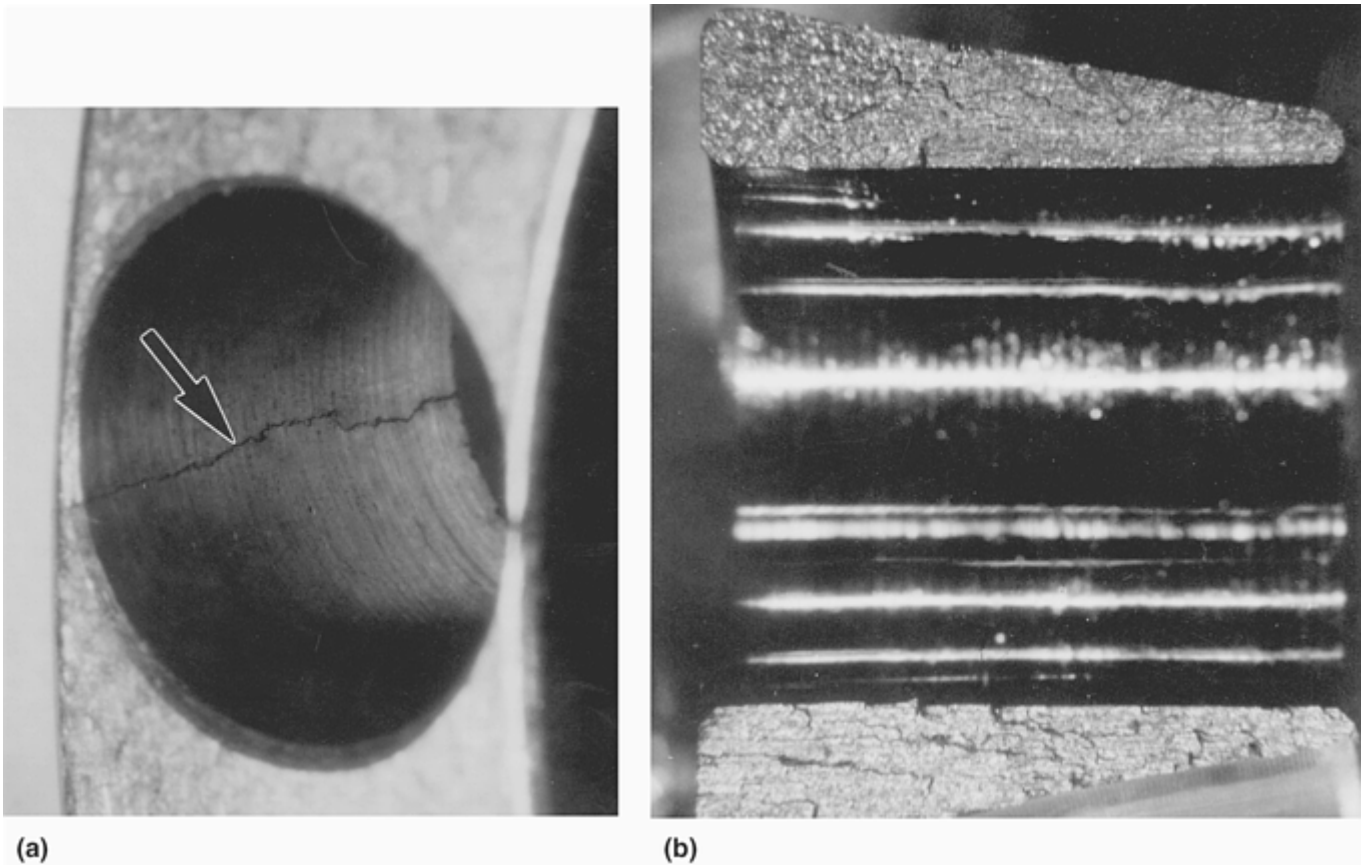


Fig. 12 Crack (a) and broken-open fracture surface (b) of failed wedge-shaped middle tapered ring. 6×

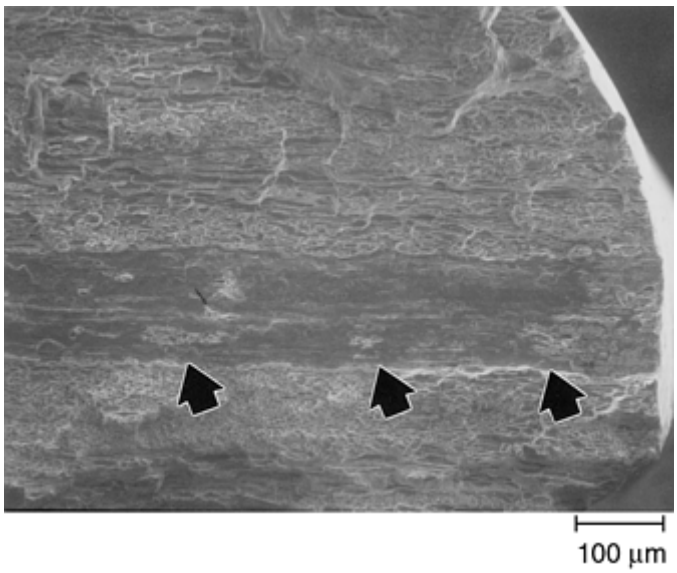


Fig. 13 Higher-magnification view of fracture surface shown in Fig. 12 at origin of cracking. Arrows indicate large manganese sulfide inclusion at origin.

Chemical analysis of the material revealed a resulfurized grade of carbon steel (SAE type 1144, UNS G11440), as required by the manufacturer. This type of steel is marketed as having a rather unusual combination of high strength *and* high machinability. The source of the high strength is in the carbon content and the cold-drawing process used to produce the bar material, giving rise to enhanced *longitudinal* tensile properties. The high volume fraction of manganese sulfide inclusions (Fig. 14) impart the high machinability properties, due to the well-documented enhancement to chipmaking

during machining. The trade-off to this combination of properties, however, is the loss of transverse properties, including strength, ductility, and toughness.

Fig. 14 Significant volume fraction of manganese sulfide inclusions in wedge-shaped tapered ring microstructure. 73×

Analysis of the forces present in the tapered-ring locking device revealed that when the fastening screws were torqued, a significant hoop stress was placed on the middle rings due to the wedging action between the inner and outer rings as well as the relatively small cross section of the middle rings at the fastener holes (see Fig. 11). Since the large inclusion was present at the minimum section thickness zone of the middle ring, the stresses applied to the middle rings during normal torquing caused failure at the inclusion. Since the material contained a high volume fraction of these inclusions, this material choice was not appropriate for this application. The material was weak in an orientation of relatively high stress. Failure prevention recommendations involved specification of a nonresulfurized grade of a low-alloy steel.

Example 2 illustrates some of the complexity and subtlety of RCA. The material was no doubt chosen for its ease of machining. The designer may not have been heavily involved in the material specification or may not have realized the sensitivity of this particular design to material anisotropy. The material itself was not defective or bad, and the part design was reasonable too, except for the material selection, which turned out to be the critical factor in this case.

Material Defects

Unacceptable imperfections or discontinuities in materials are defects, and some types of imperfections may be generally detrimental to the performance or appearance of a product or system. Some of the classical types of material discontinuities that have been identified as causal factor(s) in failures include:

Metal product form	Types of discontinuities
Forgings	Laps Bursts Flakes Segregation

Metal product form	Types of discontinuities
	Cavity shrinkage Centerline pipe Parting line grain flow Inclusions
Castings	Porosity, gas, and microshrinkage Cavity shrinkage Segregation Cold shuts Inclusions
Plate and sheet	Edge cracking Laminations Flakes
Extrusions and drawn products	Edge cracking Seams Steps Central bursts

More detailed descriptions, with physical characteristics and mechanisms for the creation of these defects, are contained in subsequent sections of this Volume. Problems that may develop during subsequent processing, such as heat treating and welding, are discussed in the section “Manufacturing/Installation Defects” in this article.

These material defects can be generally described as discontinuities that degrade the performance of a product in some way. Despite measures taken to control, document, measure, analyze, and improve the processes involved in manufacturing the metal product (such as in TQM and Six Sigma systems), material defects occur. Many defective products are prevented from leaving the mill, foundry, or forge through diligence in adhering to internal procedures and quality-assurance systems. Yet defective materials are sometimes delivered. Depending on the criticality, periodic field inspection may be required and may reveal defects not previously identified. A case study of one such occurrence illustrates the effectiveness of a maintenance plan that includes periodic inspection.

Example 3: Forging Laps in Ski Chair Lift Grip Components. Alloy steel forgings used as structural members of a ski chair lift grip mechanism were identified to have contained forging laps during an annual magnetic particle inspection of all chair lift grip structural members at a mountain resort. A lap in one of the lift grip components (Fig. 15) measured 4.8 mm ($\frac{3}{16}$ in.) long on the surface. An example of the metallurgical cross section through a similar lap is provided in Fig. 16. In accordance with the ASTM standard for magnetic particle inspection, the paint on the forgings was stripped prior to performing the magnetic particle inspection, since the thickness of the paint slightly exceeded the maximum allowable 0.05 mm (0.002 in.) thick paint layer. It should be noted that prior annual inspections, performed at a contracted magnetic particle inspection facility, revealed no significant indications on these forgings. However, the paint was not stripped prior to the magnetic particle inspection at that time.

Fig. 15 Forging lap on ski lift fixed jaw

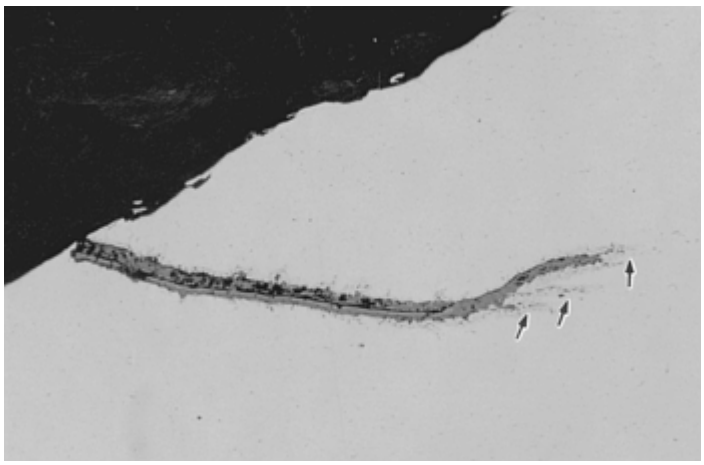


Fig. 16 Microstructure of forging lap in another ski lift grip component. As-polished. 111×

The presence of the laps, which are rejectable according to the manufacturer's drawings, indicates the forgings were delivered from the manufacturer in this condition. Aside from the obvious procedural roots related to the quality system of the manufacturer, the present issue was whether or not the laps (i.e., sharp-notched discontinuities) had “grown” in a progressive manner, such as by fatigue or stress-corrosion cracking, during the five years that the components had been in service.

The material was confirmed to be 34CrNiMo6 (a European Cr-Ni-Mo alloy steel containing 0.34% C), as required. The broken-open lap (Fig. 17) revealed a darkened area on the fracture surface that was consistent with the dimensions of the lap. The darkened area extended 0.89 mm (0.035 in.) deep. Adjacent to the darkened area, a small area of bright, fibrous fracture features was observed, as well as a transition to a bright, faceted fracture appearance. Scanning electron microscope examination in conjunction with energy-dispersive x-ray spectroscopy (EDS) revealed a heavy oxide on the dark area of the fracture surface (Fig. 18). The bright area adjacent to the dark area contained ductile dimple rupture, which changed to cleavage fracture beyond this area. It was determined through stereomicroscopy, fractography, and metallography that the oxidized portion of the fracture was the preexisting forging lap and that both bright fracture areas were created in the laboratory during the breaking-open process. A cross-sectional view of the broken-open lap is shown in Fig. 19, depicting the field of oxides in the material beneath the lap surface.

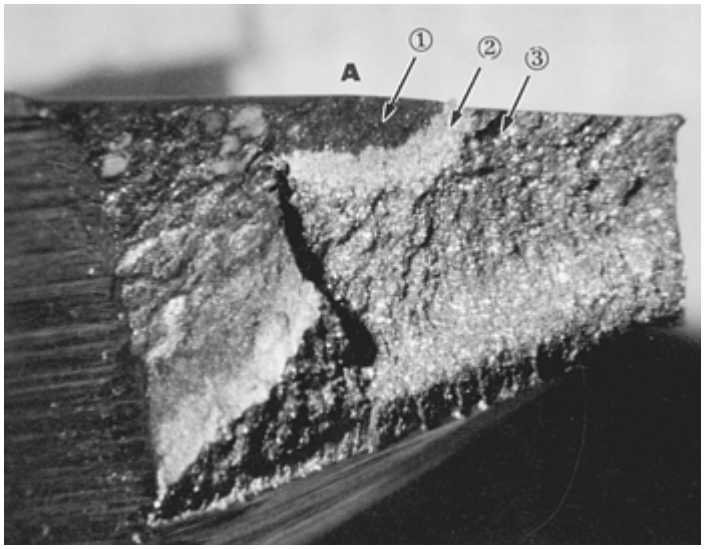


Fig. 17 Broken-open lap. 6×

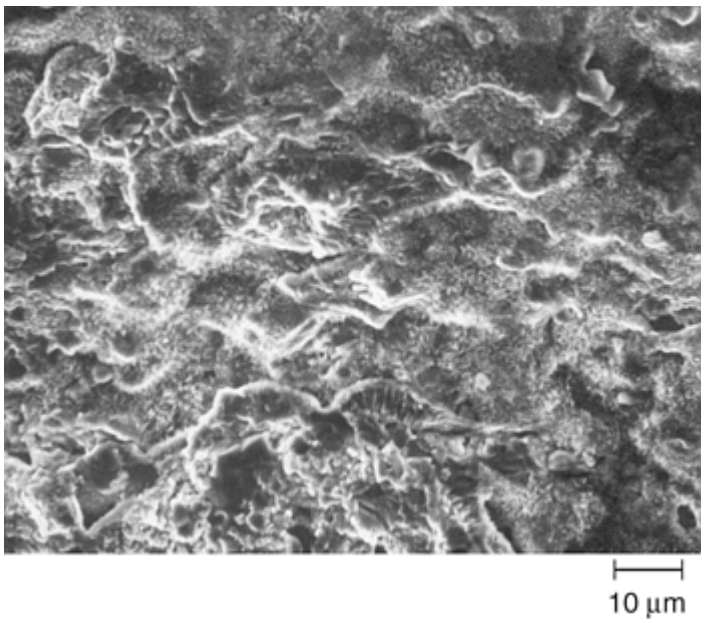


Fig. 18 Scanning electron micrograph of surface features in dark area.

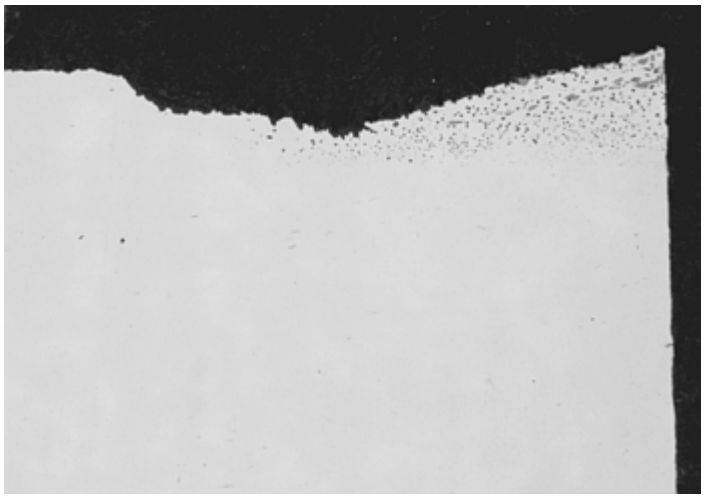


Fig. 19 Micrograph of lap. As polished. 58×

This case is particularly significant in that it is a successful example of *failure prevention* through periodic field inspections. The previously unknown defects were discovered only after magnetic particle inspection procedures adhering

to ASTM standard practices were rigorously followed. Subsequent investigation and analysis of the indications revealed no growth of the laps in service. Nevertheless, the corrective action defined that all forgings showing laps be removed from service. Preventive measures involved critical review and revision of the forging process (so that future lots would be properly forged) and revisions to the nondestructive evaluation (NDE) procedures at the forging supplier.

Building an application-life diagram around this case (Fig. 20) (Ref 29), one can explore the impact of material defects of various sizes on service life. In one possible scenario, the lower curve in Fig. 20 could describe the observed lap, being detectable by NDE and of a size sufficient to sustain growth under the anticipated service conditions *at some time in the future.* However, *at the time of the inspection,* the defect was smaller than that required for crack growth, since the date of the inspection is relatively early in the intended service life of the component. The risk of crack growth and premature failure at some time in the future (as shown by the "X" in Fig. 20) prompted the removal from service of all forgings showing NDE indications.

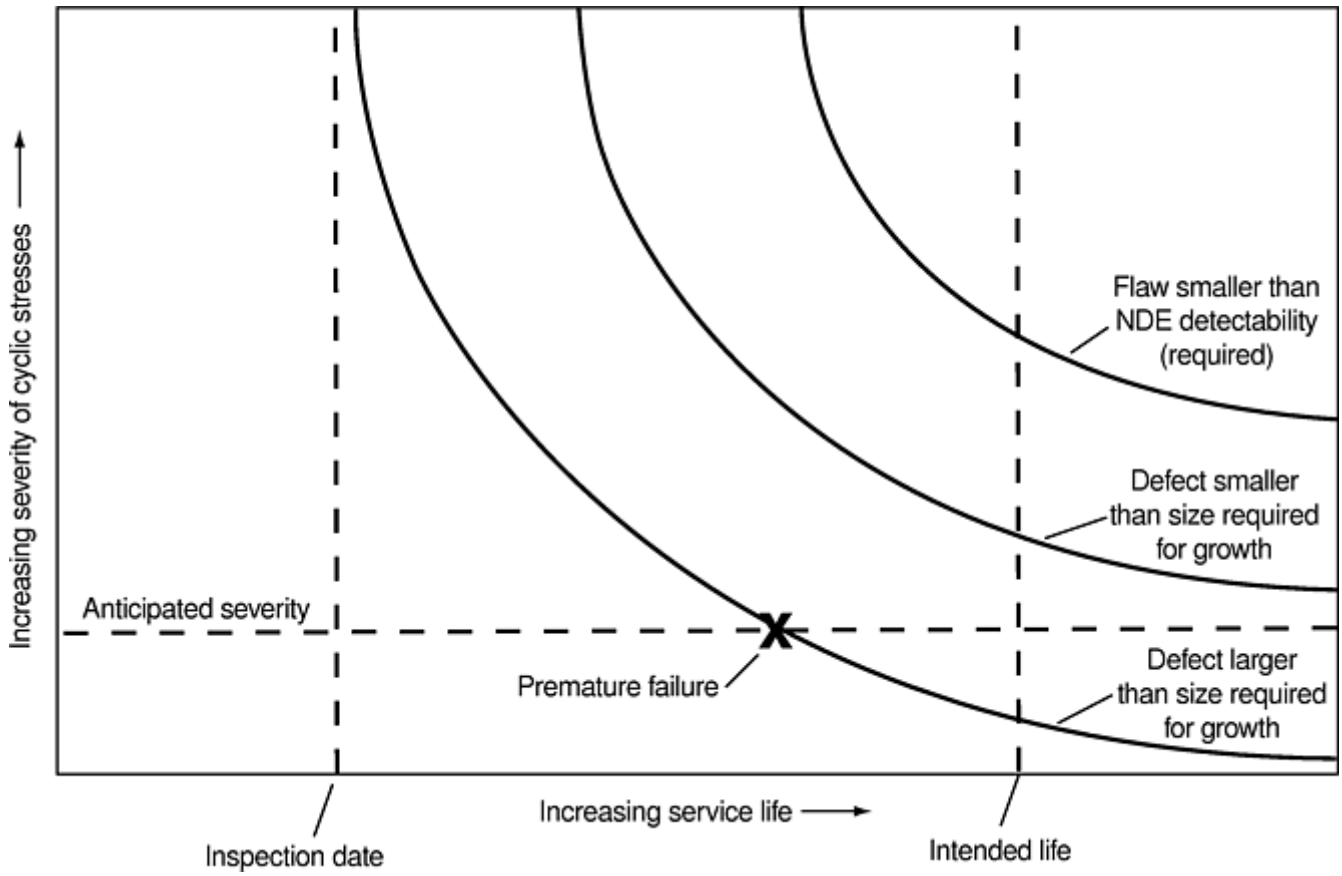


Fig. 20 Application-life diagram showing effects of different sized material discontinuities on service life

Manufacturing/Installation Defects

Manufacture refers to the process of creating a product from technical documentation and raw materials, generally performed at a factory. Installation can be considered manufacturing in-place, such as at a construction site or a new plant. Products can be designed properly using sound materials of construction, yet be defective as delivered from the manufacturer, due to rejectable imperfections (i.e., defects) introduced during the manufacturing process or due to errors in the installation of a system at a site. A wide variety of manufacturing-caused defects exist; each and every manufacturing/installation process has many variables that, when allowed to drift toward or to exceed control limits, can result in a defective product (Ref 34).

Some examples of such manufacturing/installation anomalies are listed below (Ref 35, 36). Failures associated with metalworking, welding, and heat treating operations are also discussed in more detail in other articles in this Volume, and example 4 also illustrates the effects of manufacturing anomalies on the life of a component.

Metal Removal Processes

- Cracks due to abusive machining
- Chatter or checking due to speeds and feeds

- Microstructural damage due to dull tool
- Grinding burn
- Electrical discharge machining recast layer cracking
- Electrochemical machining intergranular attack
- Residual stress cracking due to overheating

Metalworking Processes

- Cracking, tears, or necking due to forming/deep drawing
- Laps due to thread rolling/spinning
- Tool marks and scratches from forming
- Surface tears due to poor surface preparation prior to working
- Residual stress cracking due to flowforming
- Lüders lines due to forming strain rate
- Microstructural damage due to shearing, blanking, piercing
- Overheating damage during spring winding
- Laps and cracks due to shot peening
- Stress-corrosion cracking due to use of improper die lubricants

Heat Treatment

- Grain growth
- Incomplete phase transformation
- Quench cracks
- Decarburization
- Untempered martensite
- Temper embrittlement and similar embrittlement conditions
- Inadequate precipitation
- Sensitized microstructure
- Inhomogeneities in microstructure
- Loss of properties due to overheating during post-plating bake

Welding

- Lack of fusion
- Brittle cracking in heat-affected zone (HAZ)
- Sensitized HAZ
- Residual stress cracking
- Slag inclusions
- Cratering of fusion zone at endpoint
- Filler metal contour out of specification
- Hot cracking
- Cracking at low exposure temperatures
- Hydrogen embrittlement due to moisture contamination
- Liquid metal embrittlement from plating contamination

Cleaning/Finishing

- Corrosion due to inadequate cleaning prior to painting
- Intergranular attack or hydrogen embrittlement due to acid cleaning
- Hydrogen embrittlement due to plating
- Stress corrosion from caustic autoclave core leaching of castings

Assembly at Factory/Installation at Site

- Misalignment
- Missing/wrong parts

- Improper fit-up
- Inappropriate fastening system, improper torque
- Improper tools
- Inappropriate modification
- Inadequate surface preparation

Inspection Techniques

- Arc burn due to magnetic particle inspection
- Intergranular attack or embrittlement due to macroetch
- Fatigue or quench crack from steel stamp mark

Example 4: Forming Process Anomalies in Diesel Fuel Injection Control Sleeve (Ref 28). A user complained of a diesel engine that failed to start in cold weather. Troubleshooting isolated the problem to the diesel fuel control assembly, which was changed out, fixing the problem. Teardown of the fuel control assembly by the manufacturer revealed that a small subcomponent known as the cold start advance solenoid sleeve (Fig. 21) was leaking through the wall. The sleeve operates under relatively high pressure cycles in service. This component is a tubular product with a “bulb” section at one end and threads on the other. The manufacturing method used to create the bulb shape was hydroforming, using a 300 series stainless steel tube in the full-hard condition.

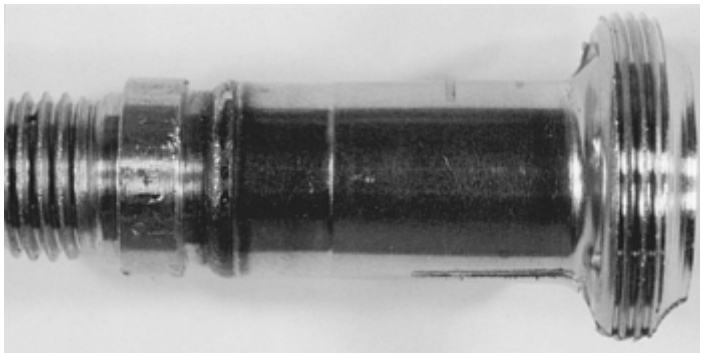


Fig. 21 Cold start advance solenoid sleeve. 0.85×

The leak was attributed to a crack in the sleeve (Fig. 22), in the radius between the bulb area and the cylindrical portion of the sleeve. Scanning electron microscope examination of the broken-open crack revealed fatigue cracks initiated at multiple sites near the outside diameter (OD) of the sleeve (Fig. 23). The crack origins were determined to be extending from shallow (0.013 mm, or 0.0005 in.) zones exhibiting ductile shear (see area between arrows in Fig. 23). Viewing the OD surface of the sleeve adjacent to the fracture plane revealed an extensive network of microcracks on the OD in the radius between the bulb and cylindrical portions (Fig. 24). A cross section through one of the fatigue crack origins revealed slip bands emanating from the microcracks (Fig. 25).

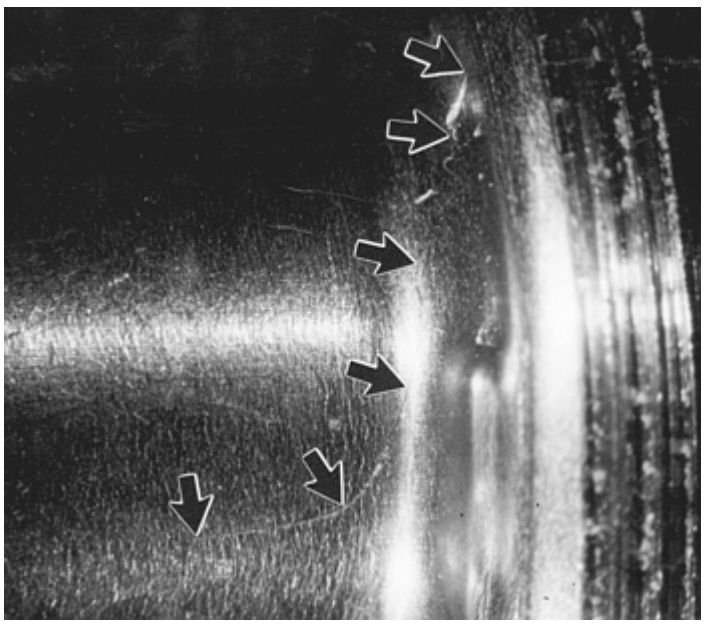


Fig. 22 Crack in sleeve (arrows). 2.5×

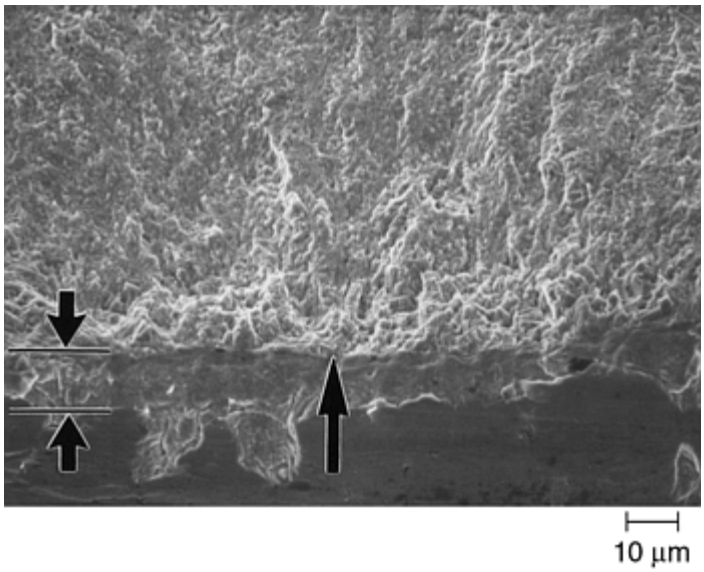


Fig. 23 Fatigue cracking from the outside diameter (OD) of the sleeve (large arrow). Area between small arrows shows evidence of ductile shear at OD surface.

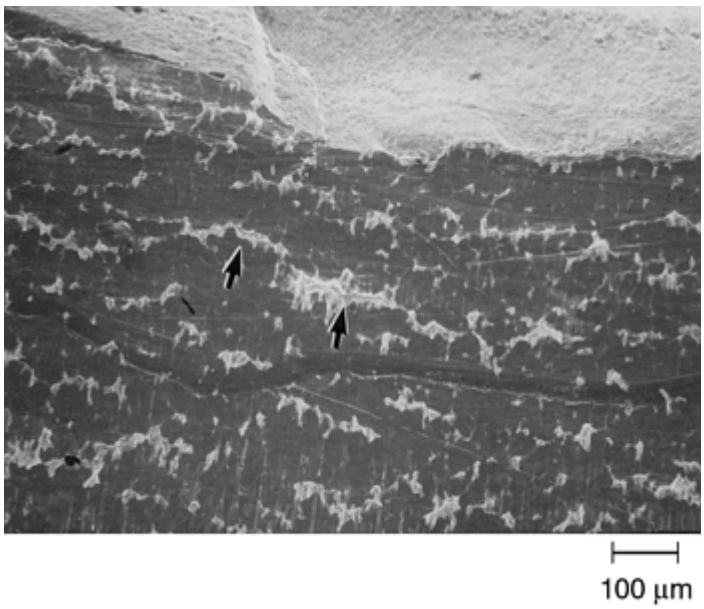


Fig. 24 Network of microcracks (arrows) on the outside diameter surface of the sleeve (lower portion of the micrograph).

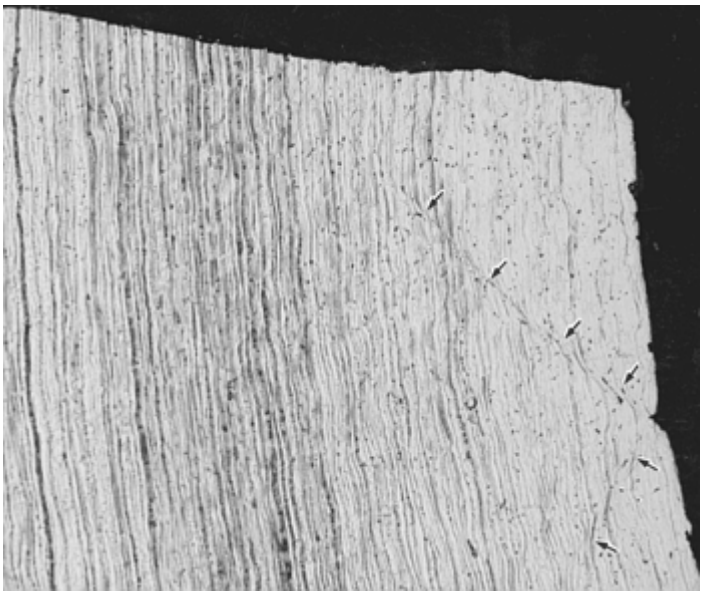


Fig. 25 Microstructure of cross section through outside diameter surface of sleeve adjacent to fracture. Fracture surface is along top of micrograph. Outside diameter surface is along right side of the micrograph. Note slip banding (arrows) emanating from microcrack. 116×

The analysis revealed that during the hydroforming process, heavy biaxial strains were imparted to the sleeve wall, in the radius between the bulb and cylindrical portions of the sleeve. When combined with the heavy strains inherently present in the full-hard 300 series stainless steel, the hydroforming strains in the radius caused the microcracking. The ductile shear areas observed at the origins (see Fig. 23) are microcracks that served to intensify the cyclic service stresses, resulting in fatigue cracks initiating and propagating from these flaws through the wall, causing the leak.

The physical root cause for this failure is a manufacturing process that omitted an intermediate stress relief or annealing treatment prior to hydroforming to the final shape.

Some time later, a similar complaint was received at the factory for a nonstart condition in cold weather. The sleeve was again identified to be leaking due to a through-wall crack. Analysis of the broken-open crack (Fig. 26) revealed fatigue cracks initiated on the *inside diameter* (ID) of the sleeve. This time, the flaw that led to the failure was shallow (approximately 0.005 mm, or 0.0002 in.) intergranular attack on the ID surfaces due to overly aggressive acid cleaning or insufficient rinsing after the acid-cleaning operation. Examination of the OD surfaces revealed no microcracking or evidence of localized strain. Thus a *second* manufacturing defect affecting the same component was identified through failure analysis to have caused the *identical* complaint from the field.

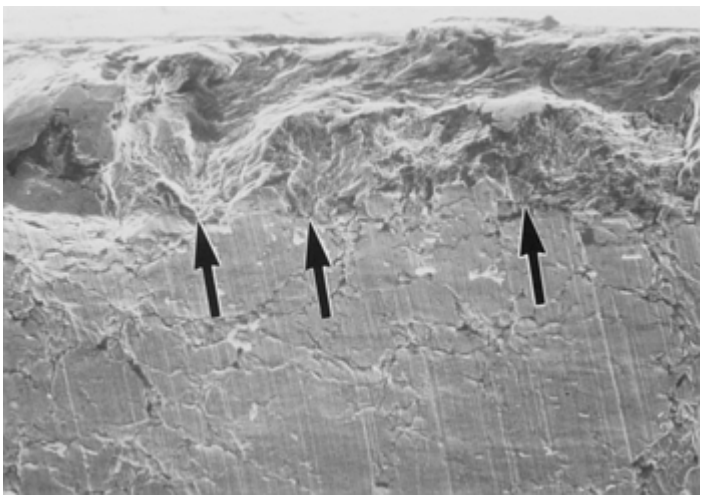


Fig. 26 Multiple fatigue crack origins (arrows) initiating in a network of intergranular attack on the inside diameter of the sleeve. 155×

Using the application-life diagram, the strong effects of minute surface anomalies in this fracture critical component is clearly apparent (Fig. 27). As a result of the severity of the pressure cycles in service, the sleeve cannot tolerate surface flaws.

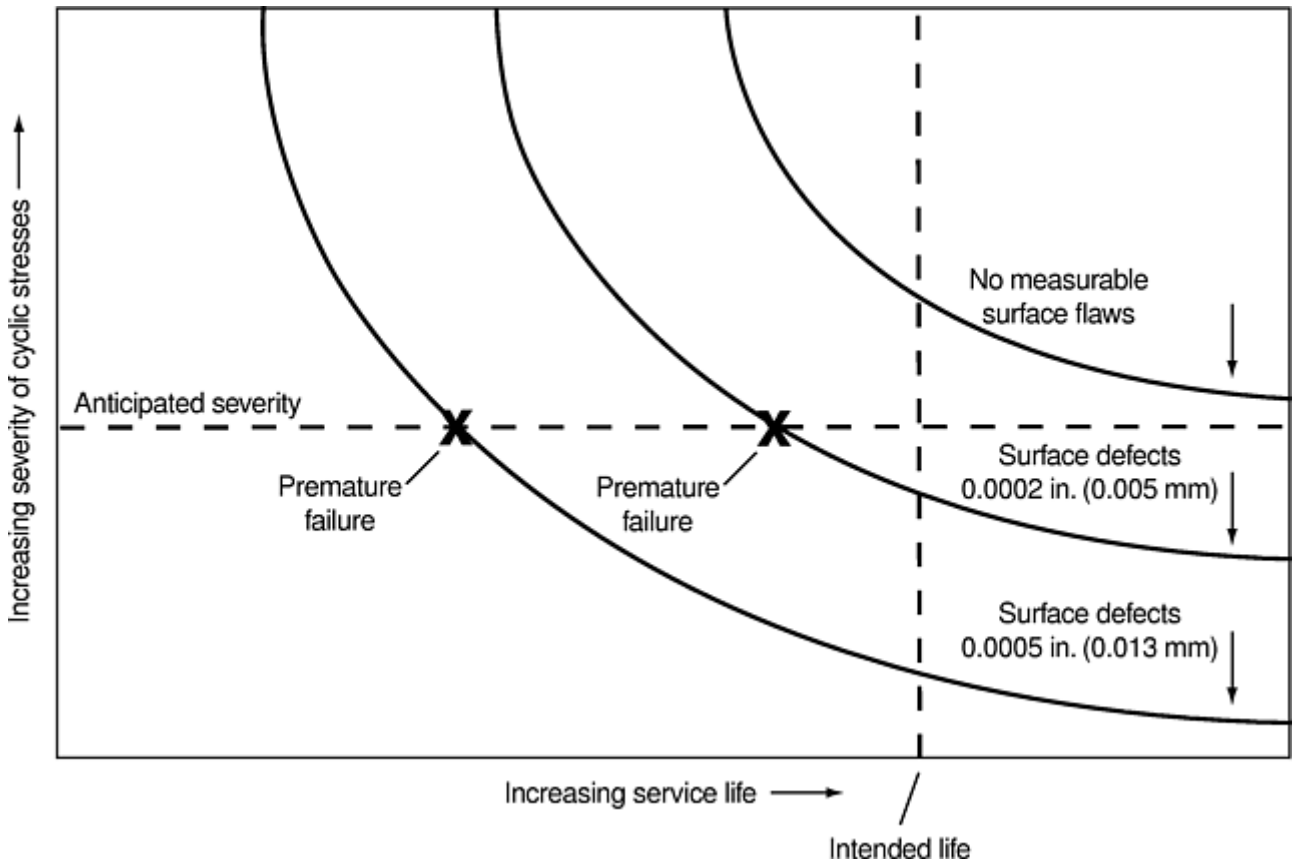


Fig. 27 Application-life diagram showing effects of manufacturing-caused surface discontinuities on service life

Service Life Anomalies

The life of a component or system is heavily dependent on the conditions under which the product operates in service. The service life of a product includes its operation, maintenance, inspection, repair, and modification. Failures due to anomalies in any one of these aspects of service life are unique from those created during the design, procurement of materials, and manufacture of products, as described above. Examples of the types of root causes of failures that result from unanticipated service conditions (Ref 30) are summarized in the following paragraphs.

Operation of the equipment outside of the manufacturer's design parameters would include an example such as a military fighter aircraft in a turn that causes “g” forces that are outside of the operating envelope of the aircraft. Another example is inlet-flow blockage on a high-performance air compressor resulting in excessive cyclic loads applied to the blades causing blade (Fig. 28, 29) and drive shaft (Fig. 30) failures. Failure analysis revealed both the compressor rotor and the shaft sustained fatigue failures.

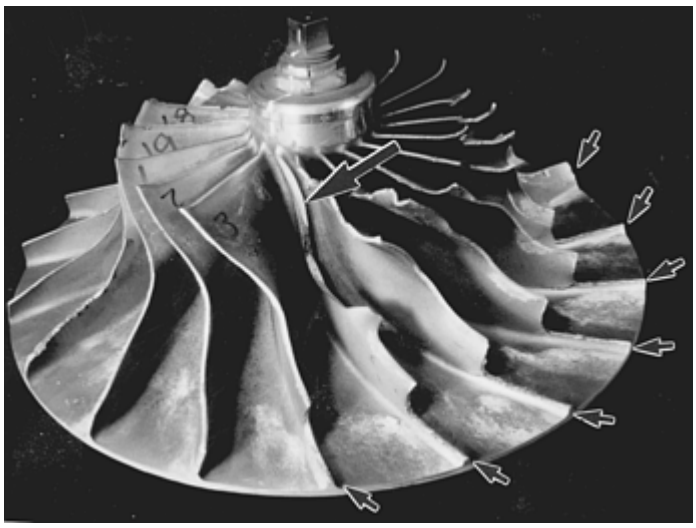


Fig. 28 Failed compressor rotor. Arrows indicate fractured portions of blades. 36×

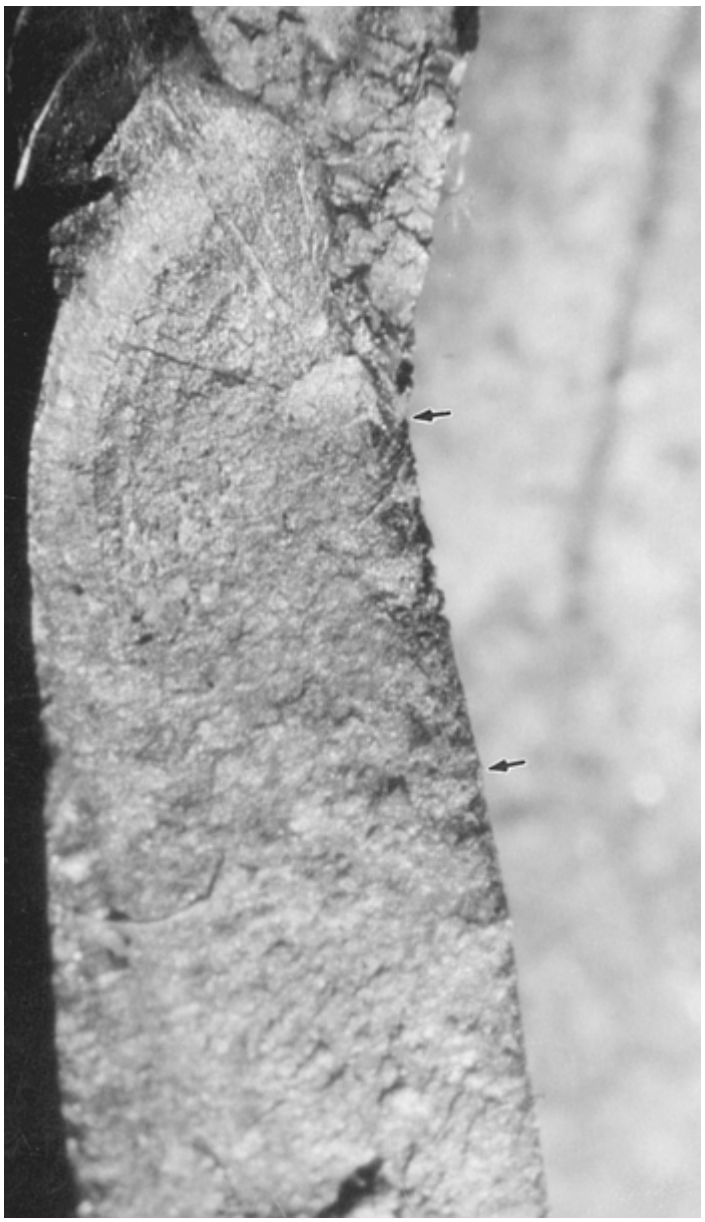


Fig. 29 Compressor blade fracture surface showing fatigue origins on low pressure (i.e., right) side of blade, as indicated by the arrows. 13×

Fig. 30 Failed compressor rotor shaft. Fracture occurred at radius between large and small diameters. Arrows indicate some of fatigue origins. 1×

Careful fracture analysis revealed fatigue cracks initiated on the low-pressure side of the blades, which are in compression during normal compressor operation. However, when the inlet flow is blocked, particularly when the blockage is only partial, the blades sustain alternating tensile forces, one load cycle per revolution, on the low-pressure side of the blades, resulting in the observed blade fractures. The shaft failed subsequently, due to the severe imbalance and rubbing caused by the blade failures.
Exposure of the product or system to environments more aggressive than forethought would include examples such as:

- Microbiologically influenced corrosion in a cooling-water system using river water in which the ecosystem has changed
- Organic chloride-containing environment exposing a titanium centrifuge bowl, resulting in stress-corrosion cracking
- Faulty sensor cable resulting in an overtemperature condition in a jet engine, which consumes the high-pressure turbine blade life

Improper maintenance would include examples such as:

- Installing a metallic fuel line onto the mating fitting by forcing the tube to align with the mating fitting. Adding the installation stress to the normal cyclic stresses results in a leak due to fatigue cracking.
- Weld repair of a material that is sensitive to high heat cycles, causing brittle cracking and subsequent fatigue failure
- Misalignment of a bearing during rebuild, causing bending loads on the shaft and resulting failure by rotating bending fatigue

Inappropriate Modifications. An example of this would be part-through drill holes in bicycle handlebar stem resulting in fatigue initiation at holes and subsequent fracture (Fig. 31, 32).

Fig. 31 User-modified bicycle handlebar stem failed in service

Fig. 32 Multiple fatigue initiations at part-through drill holes in user-modified bicycle handlebar stem. 3×

The application-life diagram is useful in exploring the effects of service-life anomalies on the lives of products. For the compressor inlet blockage case described previously, the Fig. 33 depicts the significant loss of service life when the rotor blades sustain the unintended cyclic stresses that occur during an inlet blockage event.

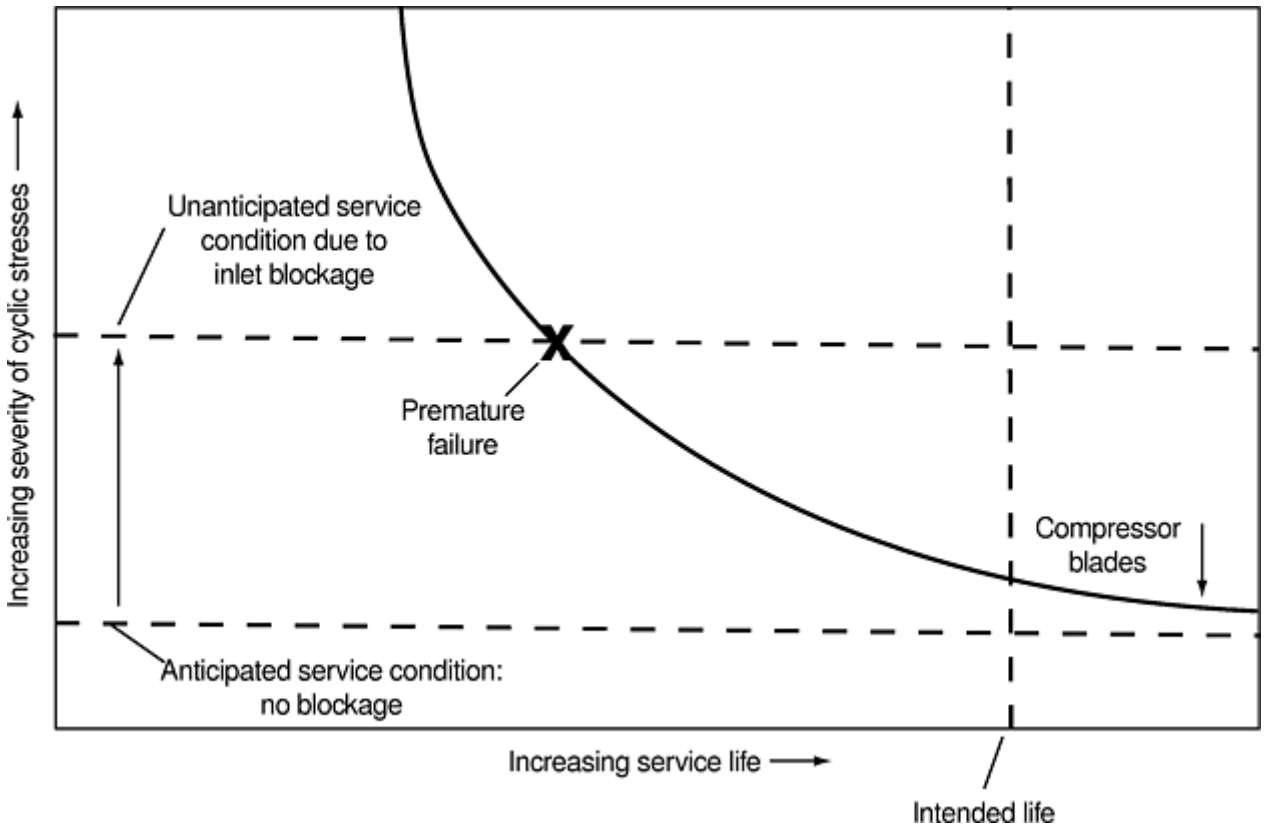


Fig. 33 Application-life diagram showing effects of increasing the severity of the service condition

References cited in this section

13. Engineering Aspects of Failure and Failure Analysis, *Failure Analysis and Prevention,* Vol 10, 8th ed., *Metals Handbook,* American Society for Metals, 1975, p 1–9

15. J.J. Asperger, Legal Definition of a Product Failure: What the Law Requires of the Designer and the Manufacturer, *Proc. Failure Prevention through Education: Getting to the Root Cause,* 23–25 May 2000 (Cleveland, OH), ASM International, 2000, p 25–29

23. P.F. Wilson, L.D. Dell, and G.F. Anderson, *Root Cause Analysis: A Tool for Total Quality Management,* ASQ Quality Press, 1993

24. G.F. Smith, *Quality Problem Solving,* ASQ Quality Press, 1998

25. R.K. Mobley, *Root Cause Failure Analysis,* Butterworth-Heinemann, 1999, p 37–39

26. M. Ammerman, *The Root Cause Analysis Handbook: A Simplified Approach to Identifying, Correcting, and Reporting Workplace Errors,* Max Ammerman/Quality Resources, 1998, p 67

27. *Failure Analysis, The British Engine Technical Reports,* American Society for Metals, 1981

28. F.A. Hossain and J.J. Scutti, Four Fundamental Root Causes of Failure: Case Histories, *Failure Analysis: A Foundation for Diagnostics and Prognostics Development,* Proc. 53rd Meeting of Society for Machinery Failure Prevention Technology, 19–22 April 1999 (Virginia Beach, VA), 1999, p 185–195

29. J.J. Scutti, "Where Things Go Wrong: Where to Look for Failure Prevention Opportunities," Presented to ASM International Materials Solutions Conference, 9–11 Oct 2000

30. J.J. Scutti and F.A. Hossain, Unanticipated Service Conditions, *Proc. Failure Prevention Through Education: Getting to the Root Cause,* 23–25 May 2000 (Cleveland, OH), ASM International, 2000, p 141–148

31. C.O. Smith and B.E. Boardman, Concepts and Criteria in Materials Engineering, *Stainless Steels and Special-Purpose Metals,* Vol 3, 9th ed., *Metals Handbook,* American Society for Metals, 1980, p 825–834

32. D. Krashes and J.J. Scutti, Poor Surfaces and Intersections of Surfaces Still Cause Trouble Just Like They Used to Do, *Technology Showcase: Integrated Monitoring, Diagnostics and Failure Prevention,* Proc. Joint Conference, 22–26 April 1996, Society for Machinery Failure Prevention Technology (MFPT), 1996, p 681–690

33. F. Hossain and J.J. Scutti, Failure of Components Although the Causes are Simple & Well Documented, *1998 Technology Showcase,* Proc. Joint International Conference, JOAP International Condition Monitoring Conference, 20–24 April 1998 (Mobile AL), JOAP-TSC, 1998, p 455–464

34. G.F. Smith, *Quality Problem Solving,* ASQ Quality Press, 1998, p 53

35. P.F. Wilson, L.D. Dell and G.F. Anderson, *Root Cause Analysis: A Tool for Total Quality Management,* ASQ Quality Press, 1993

36. R.P. Baggerly, Preventing Failures Resulting from Machining Issues, *Proc. Failure Prevention through Education: Getting to the Root Cause,* 23–25 May 2000 (Cleveland, OH), ASM International, 2000, p 118–123

Introduction to Failure Analysis and Prevention

James J. Scutti, Massachusetts Materials Research, Inc.; William J. McBrine, ALTRAN Corporation

Charting Methods for RCA

Many tools exist to assist in performing RCA. The most important element, however, is the preservation of an open mind by the investigator or investigating team. Preconceived ideas or the existence of an investigative bias often obstructs effective root-cause investigations.

A visual representation of an RCA is more easily understood than a long narrative description. Many charting methods have been developed that facilitate the logical organization of information as an aid in performing an RCA. Although such techniques can be invaluable for completeness and logistical analysis, one must not inhibit creativity and an open mind.

The following paragraphs outline a brief and somewhat simplified description of several common charting methods that may be useful in performing an RCA.

A fault-tree analysis is a deductive analysis that identifies a top event, in this case a failure, and then evaluates all credible ways in which this event could have occurred by identifying the interrelationships of basic events or conditions that lead to the failure. The tree is organized by identifying all event strings that lead to the top event and connecting them with a "gate" that depicts the logical relationship. Figure 34 depicts a simplified fault tree.

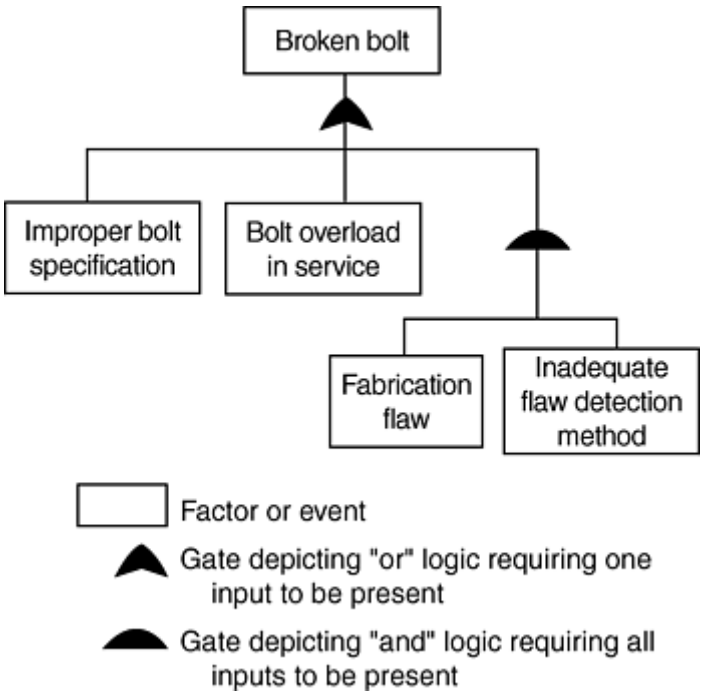


Fig. 34 Simplified fault-tree example

Event and causal factor analysis charting is a very flexible tool that is very useful for performing a logical analysis of the chronological sequence of events and causal factors. The construction starts with a basic timeline with the addition of related conditions, secondary events, and presumptions.

To construct the chart, enclose events in rectangles and connect them in sequence from left to right using solid arrows. The terminal event should be listed at the right-hand end within a circle. In ovals, list conditions, causal factors, and contributing factors and show the relationship between events with dashed arrows.

Barriers may also be added to the chart to identify barriers that failed, allowing events to occur. A barrier can take many forms including a physical barrier such as a locker door or a procedural barrier that was not properly implemented.

The basic elements of the event and causal factor chart (Fig. 35) are primary events, secondary events, and conditions. Events make up the backbone of the chart, while conditions are circumstances pertinent to the situation. The goal of the analysis is to identify the key equipment failures, process failures, or human errors that allowed the loss event to occur. Once the chart is laid out, the causal factors are identified. These are identified as the factors that if eliminated would have prevented the occurrence or lessened the severity of the loss event (Ref 35).

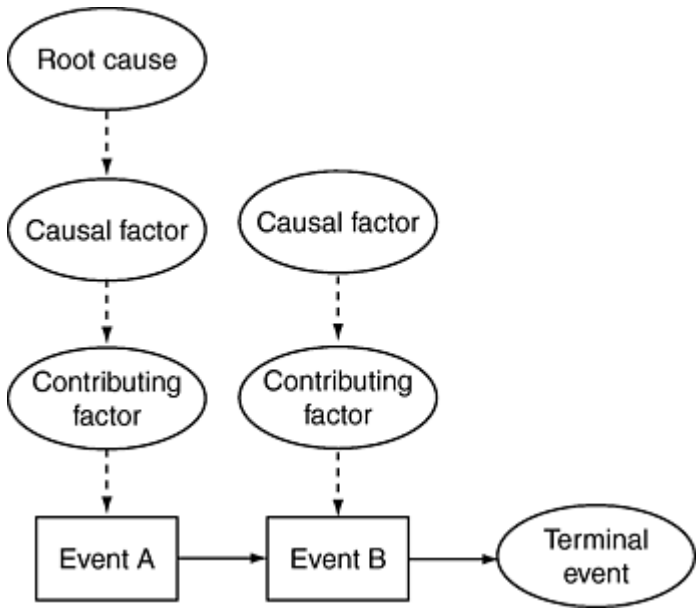


Fig. 35 Simplified event and causal factor chart

Cause-and-Effect Analysis. Failures are always caused to happen. A cause-and-effect analysis is a way to relate causes to a failure in an attempt to find the root cause. Causes can be design problems, human performance, poor fabrication, and so forth. A simple cause-and-effect analysis can take the form of a fishbone diagram (Fig. 36) that can be constructed as follows:

1. Clearly describe the failure at the right side of the diagram.
2. Identify the main cause categories as branches converging on the failure.
3. Brainstorm and list all causes on each branch.
4. Analyze the data until the root cause(s) are identified (Ref 11).

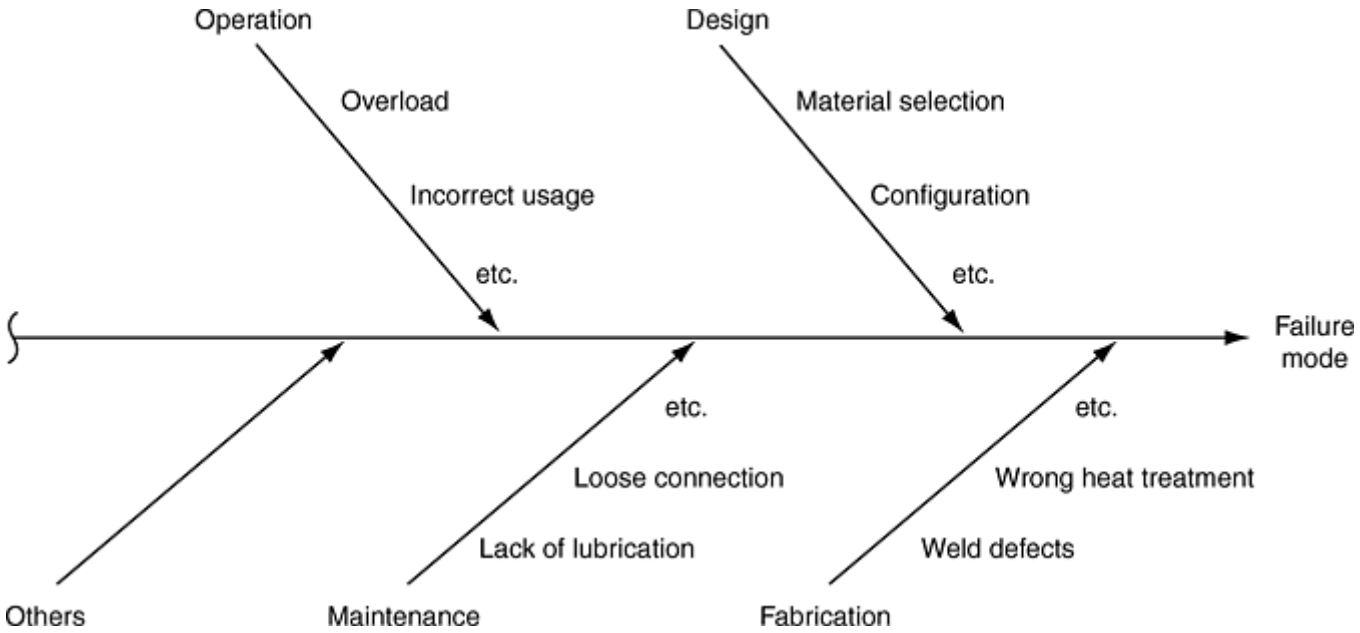


Fig. 36 Simplified fishbone diagram

Five Whys is a simple technique that is intended to lead the user into deeper levels of cause identification, thus leading one further into root cause. The overall objective is to ask "why" after each cause has been identified until true root causes are identified. There actually may be more or less than five "whys" to reach the root-cause level desired (Ref 11). The following example demonstrates this simple concept:

- Event—Highway bridge failure
- Why?—Corrosion damage on structural steel
- Why?—Water collection
- Why?—Debris clogging drainage pipes
- Why?—No maintenance performed to clean pipes
- Why?—Maintenance funding reductions (root cause)

References cited in this section

11. B. Anderson and T. Fagerhaug, *Root Cause Analysis: Simplified Tools and Techniques,* ASQ Quality Press, 2000, p 7, 125

35. P.F. Wilson, L.D. Dell and G.F. Anderson, *Root Cause Analysis: A Tool for Total Quality Management,* ASQ Quality Press, 1993

Other Failure Analysis Tools

There are many other “tools” that must be considered in performing a failure analysis. In addition to root-cause techniques, tools available to the analyst include:

- Review of all sources of input and information
- People interviews
- Laboratory investigations
- Stress analysis
- Fracture mechanics analysis

Sources of Input

Physical data such as failed parts, samples of environmental influences, photographs, data collection records (pressure, temperature, speed, etc.), and background data are an important part of the investigative process. Forensic analysis of such parts and data is the backbone of any failure investigation. Some of the key elements to an investigation include:

- *Physical evidence:* Broken parts, samples, malfunctioned components, positions, configurations, and so forth. The timely preservation, collection, and recording of physical evidence are essential to any effective failure investigation. The preservation of evidence is done by restricting access to a failure site, preserving of configurations and positions, taking a photographic record of the as-found situation, making sketches, recording of process variables (pressure, temperature, position, etc.), marking and tagging pieces and positions.
- *Background data:* Design data, specifications, technical data, analysis or simulation results, and so forth
- *People:* Witnesses, operators, designers, maintenance personnel, participants, experts

People Interviews

Interviews can provide an essential source of information in any failure investigation. This information, if solicited and documented properly, augments that collected by physical data or research. A very effective way to collect information from people is through the interview process. There are three reasons one would collect data through interviews:

- Firsthand data (witnesses, participants, etc.)
- Background and circumstantial data (historical experiences, related events, situational insights, etc.)
- Expert information (to elicit technical knowledge)

It is essential that those having firsthand data be interviewed as soon as possible after the failure. Important information can be corrupted by losing some of the subtle points over time or by the tendency to have one's firsthand knowledge evolve when discussing the event with other individuals.
Some important points to consider when performing interviews include:

- Explain why the interviews are being performed and maintain confidentiality when possible.
- Interview individually or small groups when possible. Never interview somebody with one's supervisor or manager present or in any other influencing or restricting environment.
- Make the interview environment as comfortable and unintimidating as possible.
- Ask open-ended questions and do not guide the responses.
- Distinguish between firsthand and secondhand knowledge.
- Solicit specific quantitative data, qualitative data, and opinions.
- Get referrals to others who may have pertinent information and other sources of data.
- Recognize biases and paradigms when interpreting answers.

Laboratory Investigations

After pertinent data and samples have been collected, a laboratory investigation is often needed to fully analyze the physical evidence and to identify the failure mechanism. Good procedures in a laboratory begin with good sample collections and handling.

In-Situ Sample Collection and Laboratory Receipt. When collecting samples for laboratory examination, it is a good rule of thumb to collect failed parts, nearby fragments, and lubricant and fluid samples. Collect evidence beyond what is apparent at the time of the initial assessment. Collect undamaged samples of similar components for comparison to the damaged one. Draw diagrams to indicate the position of parts and sample collection locations. Do not be afraid to take many photographs while photodocumenting the scene. Take shots from every angle and always have a scalable object in the photo, preferably a ruled scale. Make in situ markings of fluid levels or other positions that should be recorded prior to disturbing. Having the appropriate documentation and collection tools at a failure site is important to be prepared for activities that may not be anticipated prior to arrival.

Generally, samples should be collected in polyethylene jars or bags using protective gloves and appropriate collection tools. Liquid samples should be collected in glass jars with Teflon-lined covers. Samples for microbiological analysis should be collected in sterile containers and kept cool for prompt analysis. Surfaces should be free from fingerprints or other sources of contamination. Protect samples, particularly delicate items and fracture surfaces, from each other and from other sources of damage.

Tag or label samples in order to indicate when and why it was collected, how it was oriented, who removed it, and what were relevant in situ observations. Generally, it is desirable to collect the largest reasonable sample for laboratory examination prior to sectioning and removal of smaller samples.

Samples received in a laboratory can range from a large component that requires a high-capacity crane to move to something that can only be seen under a microscope. After appropriate collection, receipt, handling, labeling, and appropriate storage of the sample, it is essential to ensure that important evidence is not lost or altered. Samples should always be kept in a dry, secure location and a storage record maintained. A materials safety data sheet (MSDS) should be acquired, and appropriate storage requirement of hazardous material observed. An experienced investigator will also anticipate the disposal of hazardous material after the investigation is completed. For many such materials, disposal in the trash or down the drain is no longer an option. Specialists must be called in to remove and dispose of the material.

Laboratory Analysis. Steps taken in a laboratory after proper receipt may include:

- Initial examination
- Photodocumentation
- Nondestructive examination
- Material verification
- Fractographic examination
- Metallurgical analysis
- Mechanical properties determination
- Analysis of evidence
- Writing of a report

Handling of samples and laboratory techniques employed in a failure analysis are discussed in greater detail in other sections of this Volume (see the article “The Failure Analysis Process: An Overview” for an introduction).

Stress Analysis. Performance of a stress analysis is often a critical part of a structural failure analysis. Stress-analysis techniques are typically used to determine the state of stress as a result of external loadings or other sources of stress such as thermal transients or applied accelerations. Available stress-analysis techniques include hand calculations using theories of strength of materials, approximations derived from reference sources, empirically derived sources and methods, and computerized techniques such as the finite-element analysis (FEA) method.

The FEA method is widely used as both a design tool and a failure analysis investigative tool. Finite-element analysis can be applied to many areas useful in failure analysis, the most common being stress analysis, heat transfer and fluid flow, and electromagnetic properties. Finite-element analysis is able to model complex conditions and handle transient and nonlinear conditions that are typically too complex to perform using hand calculations or other analytical approximations.

The use of FEA in a failure analysis is different from its use in a product design capacity. In a failure analysis, special attention is directed to the failure location. This area of the FEA model may have a finer meshing to capture localized stress concentrations or other localized effects. Applied loadings should include actual load histories that are associated with the failure, including events that are not associated with normal design considerations. This is in contrast to an FEA model used for design that would be used to capture stresses in the entire component as a result of loadings anticipated by design. The results of a failure analysis model would be compared to failure criteria such as shear strength, yield strength, and so forth, or actual observed component deformation. Design models would then be used to qualify the component against the applicable design criteria such as would be published in a code or standard. (See the article "Finite Element Modeling in Failure Analysis" in this Volume for a more in-depth discussion regarding the use of FEA.)

Fracture Mechanics and Failure Analysis

Historically, the discipline of fracture mechanics was developed to understand the relationships among cracklike imperfections, stresses, and crack tolerance for the purpose of fabricating durable structures. As development of this body of knowledge continues, the usefulness of fracture mechanics in failure analysis has been recognized and is appropriately applied as one of the tools for failure analysis (Ref 2, 37, 38).

An in-depth discussion of fracture mechanics as it relates to failure analysis is beyond the scope of this article; more thorough treatment of this subject can be found in *Fatigue and Fracture,* Volume 19 of the *ASM Handbook,* and in the references cited previously. It is instructive to note that the technique is useful in some failure analyses. By performing careful measurements of relevant fracture features, incorporating known material properties (such as tensile strength and fracture toughness), and analyzing the loads and mechanics of the application, relationships can be developed to obtain an estimate of the loads and/or stresses that were operating at the time of fracture or to determine that the material in fact did not have the assumed properties. These can be compared with the loads or stresses either measured or calculated (Ref 37). Note that this is only a very brief summary and an oversimplification of the process. Extreme care must be exercised in performing such a fracture mechanics analysis, since there are uncertainties in failure analysis and in the stress-intensity-factor solutions of the failed component. The results of the stress analysis and fracture mechanics analysis must be consistent with the macroscale and microscale fractographic information and the microstructural information.

When a failure occurs by a progressive form of fracture, such as fatigue or stress-corrosion cracking, fractography can be performed to establish the fatigue striation spacing, or the crack arrest profile, across the fracture surface (as is practicable). These data can be put into appropriate equations to estimate the stress-intensity factors for either fatigue or stress-corrosion cracking, or, under some circumstances, both (Ref 2, 37, 38). Measured fatigue striation densities can be used in fracture mechanics calculations to determine either the stress range or the stress-intensity factor range, when the actual cycle counts for a given length of crack extension are known (Ref 2, 39). The usefulness of fracture mechanics as a tool for failure analysis continues to develop. One goal is to be able to reconstruct the size and growth rate of the crack over time and consider questions such as:

- Was a detectable crack somehow missed during inspection?
- Was the inspection interval appropriate?
- Did a rebuilding, overhaul, or other maintenance operation somehow contribute to the cracking?
- Did a change in service conditions or operating parameters contribute to the cracking?

References cited in this section

2. R.W. Hertzberg, *Deformation and Fracture Mechanics of Engineering Materials*, John Wiley & Sons, 1976, p 229–230

37. A.R. Rosenfield, Fracture Mechanics in Failure Analysis, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 450–456

38. R.M. Pelloux and A.S. Warren, Fatigue Striations and Failure Analysis, *Failure Analysis: Techniques and Applications*, J.I. Dickson, et al., Ed., ASM International, 1992, p 45–49

39. J.C. McMillan and R.M.N. Pelloux, Fatigue Crack Propagation under Program and Random Loads, *Fatigue Crack Propagation*, STP 415, ASTM, 1967, p 505–535

James J. Scutti, Massachusetts Materials Research, Inc.; William J. McBrine, ALTRAN Corporation

Categories of Failure

There are many ways to categorize failures and material damage in terms of forms, mechanisms, or cause. No one system is necessarily complete and consistent with the multitude of possibilities. However, categories can help prioritize or identify avenues of investigation, as long as the categories do not limit critical inquiry.

Categories of Material Stressors

To determine the cause of material failure, one must consider the active stressors. A stressor is an external influence that can be a direct or indirect cause of failure. Understanding these influences is important for effective failure analysis and determining root cause. Likewise, mitigation of the stressors is often the most logical solution to reducing susceptibility to failure. The influence of stressors is heavily dependent on the susceptibility of the component, performance criteria, the magnitude of the stressor, exposure, and the material susceptibility.
The six stressors are:

- *Mechanical:* Applied static, dynamic or cyclic loads, pressure, impact, fabrication-induced residual stresses, applied end movements
- *Chemical:* Inadvertent acute or chronic exposure to an aggressive chemical environment, material compatibility issues
- *Electrochemical:* A susceptible metal in a corrosive aqueous environment
- *Thermal:* Exposure to elevated temperatures resulting in materials degradation
- *Radiation:* Ultraviolet lighting, sunlight, ionizing radiation from nuclear power plants, and so forth
- *Electrical:* Applied electrical stress due to the presence of an electric field

Four Categories of Failures

The physical failure of materials can be placed in one of many categories depending on the classification system. The following four categories are a convenient way to descriptively categorize and discuss failures, with the ultimate goal of understanding causes and preventing failures (Ref 3):

- Distortion or undesired deformation
- Fracture
- Corrosion
- Wear

These four categories represent the general forms of failure, and each form of failure may have a variety of different underlying mechanisms (e.g., fatigue crack propagation in the case of fracture or galvanic effects in metal corrosion). It is important to point out that two or more mechanisms can occur simultaneously in some failures. These failure categories integrate with the four fundamental root causes of failures discussed in the section “Primary Physical Root Causes of Failure” in this article. As presented in Table 4, each observed failure category can be associated with any one of the four root causes.

Table 4 Examples of root causes that result in the four failure types

Failure type	Design deficiency	Material defect	Manufacturing defect	Service life anomaly
Distortion	Insufficient section thickness of a tee section results in buckling under normal load.	Cavity shrinkage in a highly stressed area of a complex structural casting used in a gas turbine engine results in permanent deformation in service, and consequential loss of clearances between the rotor and the stator housed by the casting.	Abusive thread rolling causes heavy slip banding in titanium fastener, resulting in localized stretching of the fastener upon torquing as required and an associated inability to adequately clamp joint.	Exposure of an aluminum aircraft structure to excessively high temperatures results in permanent deformation by creep and subsequent buckling.
Fracture	Cold-formed and galvanized carbon steel sheet sustains brittle fracture under normal service loads, due to strain-age embrittlement.	Lap in forging, loaded cyclically in service, grows into a fatigue crack and subsequently fails catastrophically.	Welding of alloy steel with moisture-contaminated filler metal wire results in hydrogen embrittlement and consequential brittle cracking in service.	Inappropriate hole drilling of aluminum structural bicycle component by owner results in fatigue cracks initiating and propagating in service, ending in final failure of the component.
Corrosion	Gray cast iron underground pipes used to transport hazardous materials sustain rupture due to dealloying, or "graphitization."	Iron impurities in wrought aluminum alloy suspension component for railroad car create pitting susceptibility, resulting in loss of structural integrity.	650 °C (1200 °F) stress-relief treatment of a 304L stainless steel formed and welded screen for pulp processing sustained intergranular corrosion, cracking, and failure.	Increased usage of road salt in wintertime in northeastern U.S. results in vehicle electrical problems, traced to corroded electrical contacts
Wear	Incompatible wear couple is specified in the design of an injection mold/ejector pin assembly, resulting in galling and seizure.	Improper melting and hot-working processes lead to poor distribution of primary carbides in tool steel, resulting in rapid tool wear.	Poorly machined surface of a sliding machine element leads to accelerated wear and subsequent mechanical malfunction.	Insufficient lubrication during maintenance results in premature wearout of bearing on pump shaft.

For any of these failure types, materials performance plays a critical role. Just as the performance of a component or system is dependent on the behavior of the materials of construction under the service conditions, the manner in which a component or system sustains a physical failure is strongly affected by materials performance. For example, corrosion failures of dissimilar metals in physical contact in an aggressive environment are associated with the differences in the electrochemical behavior as a result of the chemical compositions of the two metals. This illustrates that one of the most basic tenets in materials science and engineering applies to failures: the interaction of the composition, processing, structure, and properties defines materials performance (Fig. 37), whether satisfactory or unsatisfactory (Ref 40).

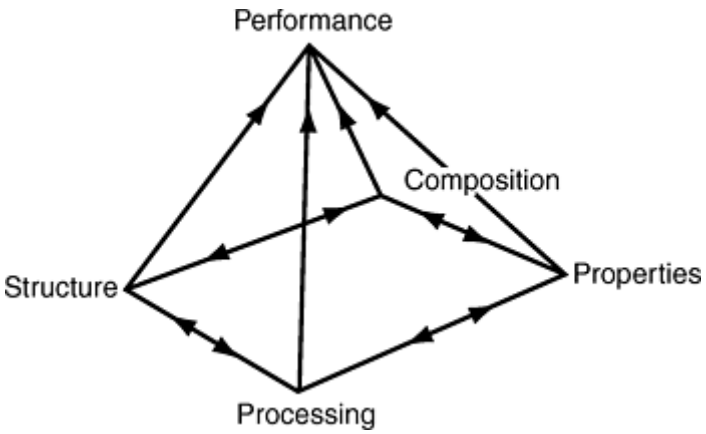


Fig. 37 Materials performance as a result of the interactions among composition, processing, structure, and properties. Source: Ref 40

Distortion. A distortion failure occurs when geometrical changes prevent a component from functioning properly such as a swollen polymer bearing in a pump or bent linkage in a transmission. Geometry change will generally be in the form of volume changes (e.g., swelling or shrinkage) or shape changes (e.g., warping, bending, or buckling).

Common causes of volume-distortion failures include temperature-induced phase changes or thermal expansion in metals, fluid absorption of nonmetallics, and curing shrinkage such as may occur in grouts and adhesives. Common causes of geometry-induced failure include inadequate design, flexural stiffness under load, stress-induced material yielding (Fig. 38), and uneven heating while in service.

Fig. 38 Example of distortion in an overloaded valve stem

Fracture. A fracture is generally defined as material separation. There are many causes and forms of fracture including brittle fracture (Fig. 39), ductile fracture, and many progressive cracking mechanisms that can lead to final fracture. An understanding of the component design, service loading, environment, and the application of sound laboratory investigative techniques such as interpretation of the fracture surfaces (fractographic examination) are essential to an effective failure analysis in the case of component fracture.

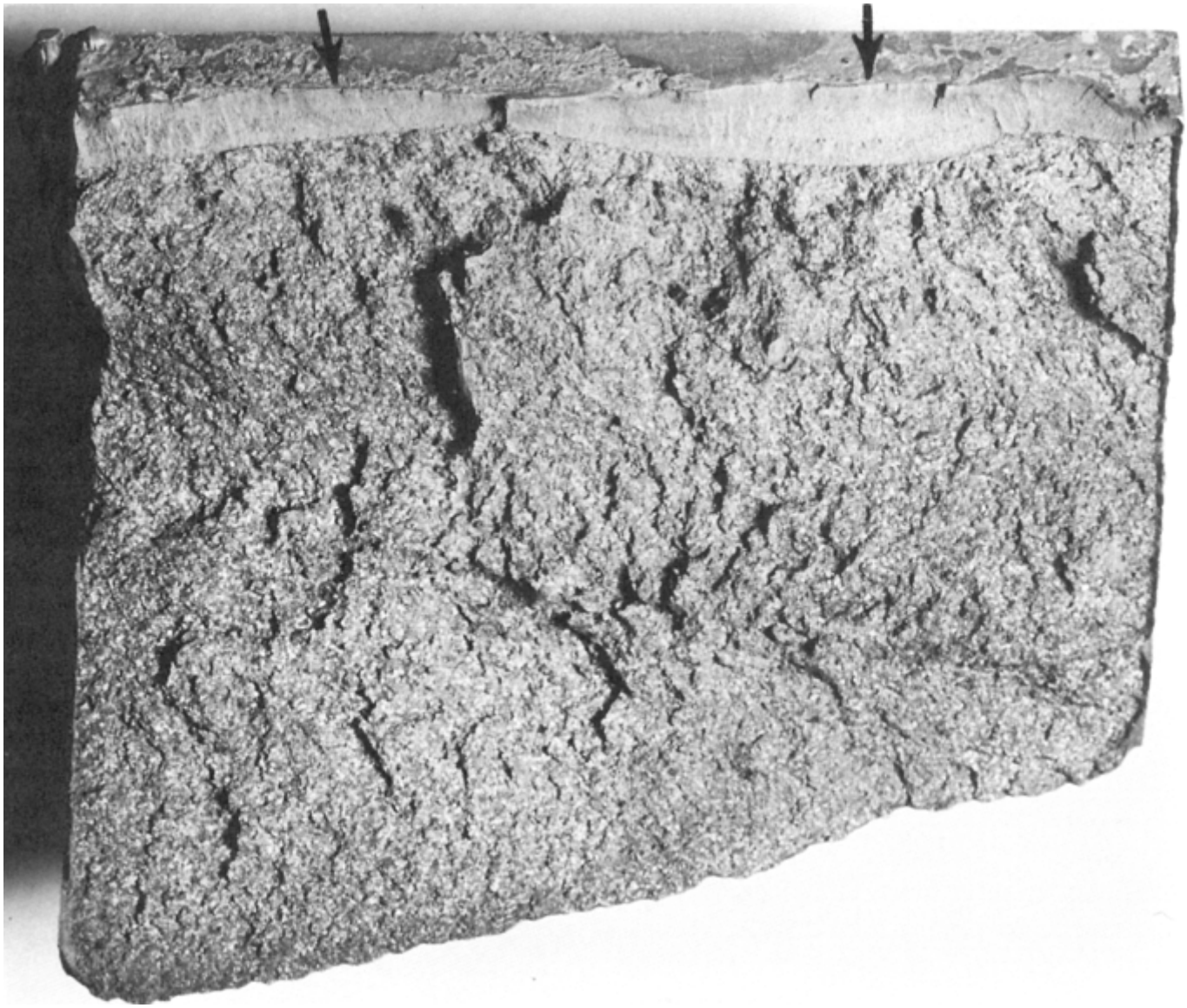


Fig. 39 Example of a brittle fracture of A36 structural steel, after sustaining fatigue cracking initially (at arrows). Source: Ref 41

Material Behavior under Load. Understanding the behavior of materials under load is important to the understanding of fracture modes. The macroscopic behavior of materials under loading is often characterized through tensile testing. It is customary to measure load and elongation during these tests and to plot the results in the form of a stress-strain diagram. Experimentally derived stress-strain diagrams can vary widely between different materials and are influenced greatly by parameters such as the speed of the test and temperature of the specimen during the test. Figure 40 depicts typical stress-strain diagrams. One curve is characteristic of mild steel, and the others are characteristic of other types of materials. However, generally speaking, each material has its own curve. Ductile materials are those that are capable of withstanding relatively large strains prior to fracture as opposed to brittle material to which the converse applies. Nonuniform and unstable transverse contraction referred to as necking in ductile materials indicates a severe overload. It reduces the effective stressed area and results in a distinction between the true stress-strain curve and the engineering stress-strain curve, which considers the original cross section when calculating the stress. A study in dislocation theory further explains the plastic behavior of metals beyond the elastic stress range (Ref 42).

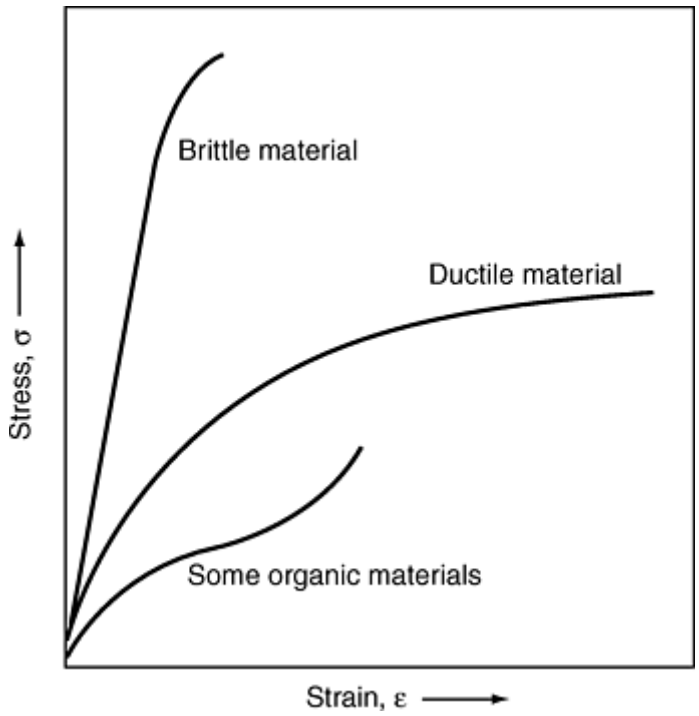


Fig. 40 Typical stress-strain diagrams. Source: Ref 42

Fundamental Fracture Mechanisms. Figure 41 illustrates the three most common fracture mechanisms in metals. Ductile fractures initiate with the nucleation, growth, and coalescence of microscopic voids that often begin at second phase particles or inclusions. Although cleavage fracture is most commonly thought of as a brittle fracture that propagates along crystallographic planes, it can also be preceded by a high degree of plasticity and ductile crack growth (Ref 43).

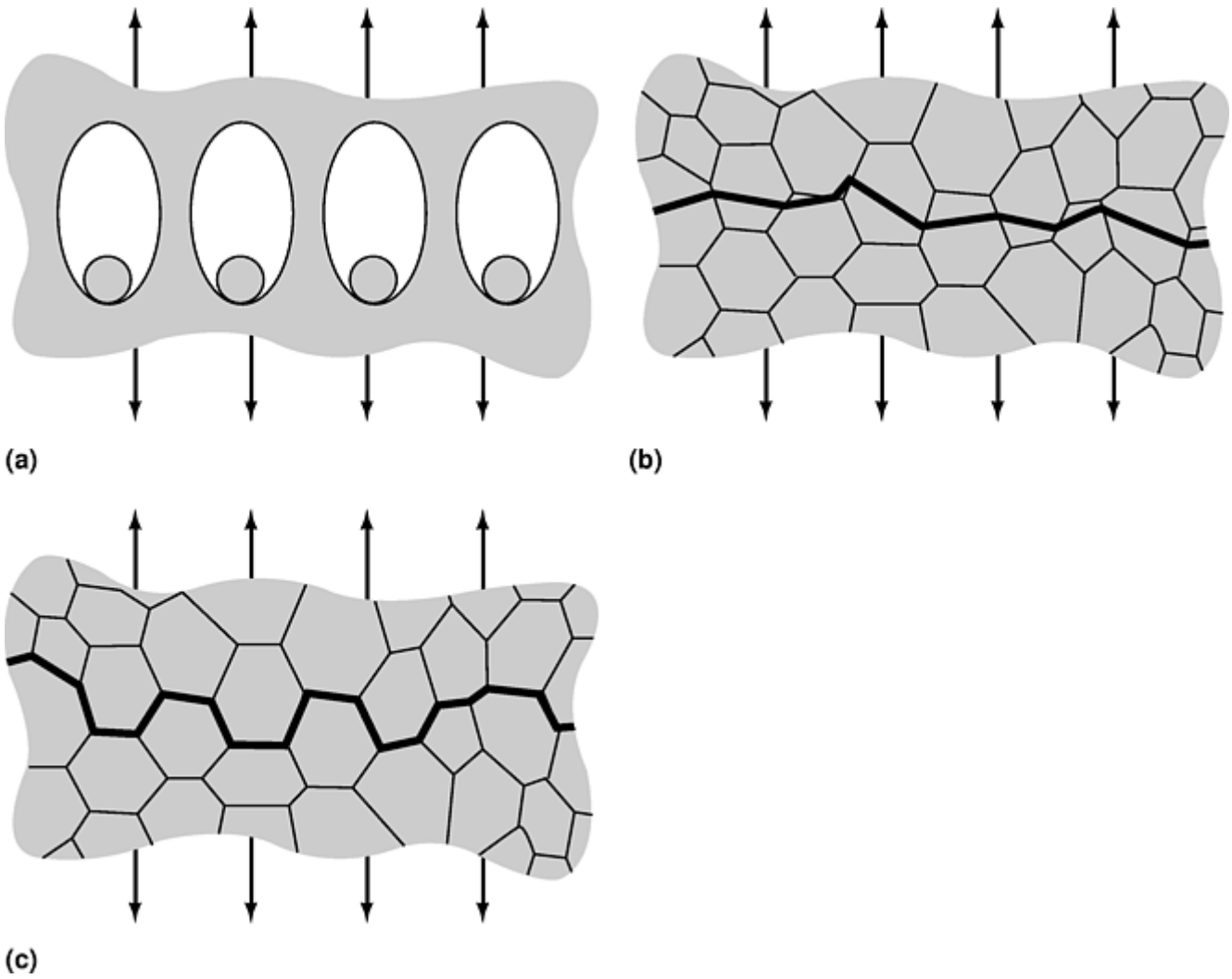


Fig. 41 Three micromechanisms of fracture in metals. (a) Ductile fracture. (b) Cleavage fracture. (c) Intergranular fracture. Source: Ref 43

Corrosion is the environmental degradation of materials. In metals, the most common type of corrosion is an electrochemical phenomenon that occurs on the surface of susceptible metal or metal alloys when exposed to a corrosive aqueous environment. Other forms of corrosion that do not involve electrochemical action include liquid metal embrittlement, corrosion in molten salts, high-temperature oxidation, and so forth. The result of corrosive attack can take the physical form of uniform surface wastage, local wastage, pitting, cracking, embrittlement, and so forth. The loss of material can eventually lead to an overload failure or through-wall penetration. The buildup of oxide scale that has a significantly increased volume when compared to the unoxidized metal can also be a problem by applying wedging load in crevices. Mitigation of corrosive attack involves a change of materials, removal of the corrosive environment, providing a surface barrier such as a coating, or providing cathodic protection. An example of piping system corrosion due to the effects of microbiological activity is shown in Fig. 42.

Fig. 42 Microbiologically influenced corrosion in a cooling water piping system

Wear failures result from the removal or displacement of surface material through contact and relative motion with a solid, liquid, or gas. There is a significant influence of friction and lubrication on the rate and severity of wear damage. Wear generally results in loss of material and load-carrying capability, adhesion, increased friction, and debris generation. Whether or not wear damage constitutes failure of a component depends on the performance criteria of the component, such as in a failed diesel engine main bearing that sustained excessive wear and a subsequent loss of control of the crankshaft radial movement (Fig. 43). Slight wear on metal valve seats may result in unacceptable leakage, while severe wear in a less critical application may be anticipated and without consequence and thus be perfectly acceptable. Controlled wear such as is the case with automotive brake pads may be part of the design criteria for a consumable component. The generation of debris could also be a critical consideration if, for example, the contamination of an ultrapure water system is at risk.

Fig. 43 Example of a wear failure in a diesel engine bearing

References cited in this section

3. D.J. Wulpi, *Understanding How Components Fail,* 2nd ed., ASM International, 1999

40. M. Cohen, MIT Lecture, 1977

41. G.F. Vander Voort, Ductile and Brittle Fractures, *Met. Eng. Quart.,* Vol 16 (No. 3), 1976, p 32–57

42. E. P. Popov, *Introduction to Mechanics of Solids,* Prentice-Hall, 1968, p 101–110

43. T.L. Anderson, *Fracture Mechanics: Fundamentals and Applications,* 2nd ed., CRC Press, 1995

James J. Scutti, Massachusetts Materials Research, Inc.; William J. McBrine, ALTRAN Corporation

Failure Prevention

Failure prevention begins with a state of mind in the specification, design, manufacture/fabrication, installation, operation, and maintenance of any component. However, before failure prevention measures are taken, the degree of reliability required in a specific situation must be determined.

There is a cost associated with failure prevention, and of course there is a cost associated with accepting failures. As shown in Fig. 44, many times it may be reasonable to accept failures should the cost of reliability enhancement outweigh the benefits. For example, the consequence of an aircraft structural failure is very high, thus demanding a high assurance of reliability. In contrast, the failure of a screwdriver may be low cost, although certainly a nuisance.

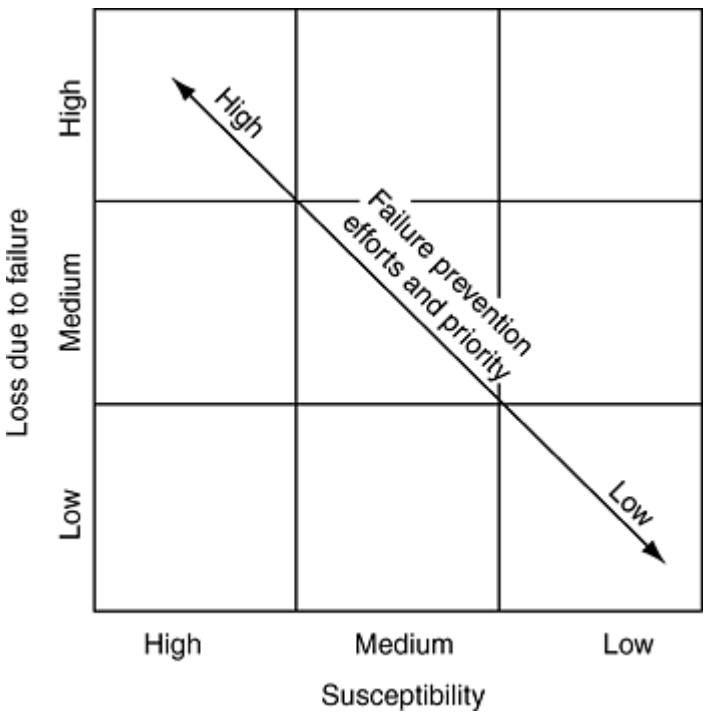


Fig. 44 Failure prevention effort prioritization

Building in reliability from the start is the most efficient way to achieve levels of reliability that will reduce or prevent failures. First, develop a performance specification that establishes the criteria for acceptable performance, answering critical questions. What are the important aspects of form, fit, and function? How long should it last? What are both the

expected and unexpected stressors? The element of life-cycle management becomes an important consideration as previously described.

Failure Modes and Effects Analysis. Just as an effective failure analysis requires a multidisciplinary approach, so does an effective failure-resistant design. Designers, material scientists, engineers, fabricators, and quality-control specialists contribute to failure modes and effects analysis (FMEA).

The FMEA procedure involves examining each item, considering how that item can fail, and then determining how that failure will affect the operation of the entire component or system. The process of identifying possible component failure modes and determining their effects on the system operation helps the analyst to develop a deeper understanding of the relationships among the different system components and to make any necessary changes to either eliminate or mitigate the possible undesirable effects of a failure. The steps involved in performing a FMEA, as identified by J. Bowles in the article "Failure Modes and Effects Analysis" in this Volume include:

1. Identify all item failure modes.
2. Determine the effect of the failure for each failure mode both locally and on the overall system being analyzed.
3. Classify the failure by its effects on the system operation and mission.
4. Determine the failure probability of occurrence.
5. Identify how the failure mode can be detected. (This is especially important for fault-tolerant configurations.)
6. Identify any compensating provisions or design changes to mitigate the failure effects.

Activities that constitute the FMEA complement and add value at every stage of the development cycle.

Applying Codes, Standards, and Regulations. The necessity to ensure interchangeability and compatibility of parts and safety factors in design led to the initial development of codes, standards, and regulations. In general, "codes" are considered to be a collection of laws or regulations that are a result of legislation to control activities. The term ***standards*** is often considered to be interchangeable with ***specifications.*** However, specifications are generally considered to refer to a more specialized and specific situation. Standards may be categorized as:

- Government regulations (i.e., requirement mandated by the government such as Occupational Safety and Health Administration, or OSHA, regulations)
- Government standards (federal specifications such as Military Specifications)
- Consensus standards (e.g., ASTM and ANSI standards)
- Technical society, trade association, and industry standards
- Company standards (both the supplier and the purchaser company may have their own standards)
- Standards of good practice
- Standards of consumer expectation

These standards include mandatory standards such as those published by government agencies. An example of these are those specified by the U.S. Code of Federal Regulation (CFR). Also, voluntary standards are often specified for mandatory compliance by a manufacturer or buyer of a product. There are many codes and standards such as those published by the American Society of Mechanical Engineers (ASME), the American Society for Testing and Materials (ASTM), and the American National Standards Institute (ANSI) (Ref 44).

The ASME Boiler and Pressure Vessel (BPV) Code is an example of a code that evolved from the necessity to prevent failures. Typical boiler operating pressures increased gradually from 206 kPa (30 psi) in the mid-19th century to more than 1378 kPa (200 psi) by 1900 and were accompanied by a much more widespread use of steam power. This led to a drastic increase in boiler explosions to a rate of approximately one per day in the United States. The evolution of the code resulted from the need to avoid increasing boiler failures as well as providing a basis for uniformity in the commercial bidding process. Today the ASME BPV Code is widely adopted and specifies acceptable materials and designs as well as fabrication, inspection, and repair methods (Ref 45).

Safety Factors and Reliability. An important element of design is the concept of a safety factor, which is typically a driving influence in the development of failure prevention concepts in codes and standards. A safety factor is generally defined as the ratio of failure load to anticipated load, if the safety factor is applied to stress. However, it can also be applied to fracture toughness, ductility in forming, casting quality, or other failure criteria that are established. Designing and manufacturing a product to adequately perform its intended function is not sufficient. A safety factor must consider an imperfect world, including manufacturing or construction tolerances, material variability, unanticipated stressors, and the effects of aging. An example of such a practice is a code specifying allowable material stresses that are much less than the strength of the material. The advent of modern materials and engineering design has reduced levels of uncertainty and has allowed a reduction of safety factors over the years. Selecting an appropriate safety factor for a given product includes consideration of:

- Degree of uncertainty about loading

- Degree of uncertainty about material strength
- Degree of uncertainty in relating applied loads to material strength
- Consequences of failure in terms of human safety and economics
- Cost of providing a large safety factor

Recommended safety factors for a performance factor (e.g., yield strength or some other failure criterion) may start from a low value of 1.25 to 1.5, where the materials are exceptionally reliable. That is, they are used under controlled conditions with frequent maintenance and inspection and are subjected to loads and stresses that are determined with certainty through testing of statistically significant material populations and/or analysis. Factors such as these are in most cases used where low weight is a particularly important consideration.

More common safety factors are in the range of 3 to 4 or higher when the loads or materials are less certain. Higher safety factors also apply in situations where repeated loads are applied, impact forces exist, materials are brittle, or there is other uncertainty. In the end, the appropriate safety factor is dictated by the applicable code or standard as well as situation specific considerations (Ref 46).

The concept of reliability is closely related to the concept of safety factors, which often incorporate a statistical approach. One must ask: if 1000 “identical” parts are put into service, what is the acceptable failure rate? The usefulness of a reliability approach depends on having adequate information on the statistical distribution of loading applied to parts in service as well as the statistical distribution of strength coming from production runs of manufactured parts. These variables are used in various statistical models with predicted failure rates compared with those considered to be acceptable (Ref 46).

Materials Selection. Design and material selection are fundamentally important in minimizing failures and hence ensuring component reliability. Selecting the most appropriate material for an application is highly product dependent and situation dependent. All functional requirements and environments must be considered in order to satisfy design requirements as well as economic considerations. Significant engineering expertise is required to ensure the material selections are appropriate for the intended function and service (including an understanding of the stressors) because trade-offs are usually required.

One of the common considerations in selecting materials is determining the desired mechanical properties. For instance, having a fracture-tolerant component is often an objective that can be achieved by selecting a material that is ductile and flaw tolerant, reducing the likelihood of brittle fracture. The trade-off is that ductility is often achieved by sacrificing overall strength, wear resistance, and resistance to deformation. In order to achieve ductility and maintain wear resistance, one may select a surface treating process such as a case-hardening process. In metals, the properties that must be considered to both ensure the desired function and reduce the likelihood of failures include:

- Tensile strength
- Yield strength
- Modulus of elasticity
- Ductility (percent elongation)
- Fatigue strength
- Fracture toughness
- Hardness
- Shear strength
- Machinability
- Coefficient of friction
- Impact strength
- Corrosivity
- Density
- Coefficient of thermal expansion
- Thermal conductivity
- Electrical resistivity
- Other physical properties

In the typical application of polymers, there are often other material properties considerations with regard to both performance and failure prevention such as:

- Stiffness
- Chemical, thermal, and ultraviolet resistance
- Electrical resistance
- Dimensional stability
- Resistance to moisture absorption

There are also unique properties associated with other material such as composites and ceramics that must be considered (Ref 47). A more complete treatment of the importance of materials selection in preventing failures is found in the article entitled “Materials Selection for Failure Prevention” in this Volume.

Operation, Maintenance, and Inspection. Achieving the expected service life of a component or system (and consequently preventing failures) requires diligence on the part of the user in operating and maintaining the component or system within established bounds. This process begins, however, with the definition of appropriate operating conditions by the manufacturer during product development and testing. The deliverable product that results from that development effort is an instruction manual that is provided with the product. Information in the manual should include installation instructions, recommended methods for activating, using, and shutting down the component or system, and maintenance recommendations or requirements. Responsibility also lies with the manufacturer to anticipate misuse of a product and adequately warn of the dangers and risks associated with such misuse (Ref 15). Warnings typically are communicated to the user through both the instruction manual, labels prominently affixed to the product, or through public announcements by the manufacturer or a government agency. Such warnings often distinguish risks of personal injury or death versus risks of damaging the product. Design of the warnings also distinguishes hazards that are obvious versus hazards that are not.

Proper maintenance of products is of paramount importance in realizing the expected service life. The user (or the actual owner) is ultimately responsible for proper upkeep of components or systems. However, initially, the responsibility also lies with the manufacturer in developing an appropriate maintenance plan for the anticipated service conditions. Such methodologies as reliability-centered maintenance can be employed to build maintenance plans that optimize requirements as appropriate for specific types of products in specific applications. Emphasis is placed on maintenance of the components or systems with the greatest impact in the event of a failure (see the article “Reliability-Centered Maintenance” in this Volume). Petrochemical and chemical-processing industries also use extensive methods for predictive maintenance for prevention of corrosion failures (see the article “Analysis and Prevention of Corrosion-Related Failures” in this Volume).

Maintenance plans could be as simple as periodic cleaning of, for example, a toaster. Conversely, the plan could be as complex as a comprehensive product-management system involving rigidly defined inspections, servicing, replacements-for-cause, and life-limited component changeouts required at various maintenance levels (that is, sites with specific capabilities). For example, in some aircraft fleet maintenance plans three levels of maintenance are defined, with easily accessible component replacements (including entire engines) allowed at the flight operations sites, partial teardown and rebuilding of more complex components (including parts of engines) at intermediate maintenance sites, and full teardown and rebuilding of all serviceable components (no matter how complex) at maintenance depots. The plan could also include the updating of maintenance manuals, training of maintenance personnel, spare parts procurement, implementation of maintenance directives, and so forth.

Maintenance and repair activities can be provided by:

- The manufacturer, for specialized, complex, and critical systems or for low-volume products
- An approved repair/overhaul company, an approach commonly used for aircraft, and other specialized, complex, and critical systems of moderate to high volume
- Independent repair providers, for less complex systems of high volume, as with automobile service stations
- The end user, as with simple systems such as yard machines

These options are listed in the order of decreasing input from the manufacturer and hence control of the repair processes used. Defining the appropriate service provider requirements can prevent failures and improve service life and reliability.

An important aspect of any maintenance plan for complex or critical products and systems is inspection. As shown previously in Example 3, sound inspection is effective in failure prevention. In general, periodic inspection programs are required for critical systems in which reducing the risk of safety or health issues is desired or required. Other conditions under which inspection programs are implemented include situations where equipment downtime has excessively high cost, such as in a paper mill and where output of a process creates a significant loss in the ability to meet demand, as in the availability of electric power during peak usage.

Inspection programs identify degradation or loss of function of equipment and unanticipated service conditions. In safety and health critical systems, federal regulations often require inspection programs. For example, periodic inspection (or condition assessment) of pressure vessels, tanks, and piping that store or transport hazardous substances is required by federal law under OSHA 1910.119. The inspections are typically performed by visual and nondestructive techniques, documenting internal and external corrosion, corrosion under insulation, poor welds or joint failures that might leak, inappropriate support, lining failures, and so forth, in accordance with the appropriate American Petroleum Institute (API) and ASME codes. Other U.S. government-required inspection programs include aircraft (Federal Aviation Administration, or FAA), transportation of hazardous substances (Department of Transportation), and the manufacture of pharmaceutical products (Food and Drug Administration, or FDA).

Incorporating Lessons Learned. A proper failure analysis and root-causal analysis can provide valuable information into the entire design and production process of a product or system. The implementation of these analyses alone, however, does not ensure that benefit is gained. These results must be communicated into corrective-action recommendations that are routed back to the proper stage of the product life cycle. Implementation of these actions and final verification of these actions should be performed in order to ensure that the desired outcome is obtained in terms of failure prevention. Organizing a multidisciplinary lessons-learned meeting to discuss the outcome of the failure analysis is important from a learning perspective and to ensure proper communications (Fig. 45), particularly in a large company with many departments involved in the evolution of a product (Ref 48).

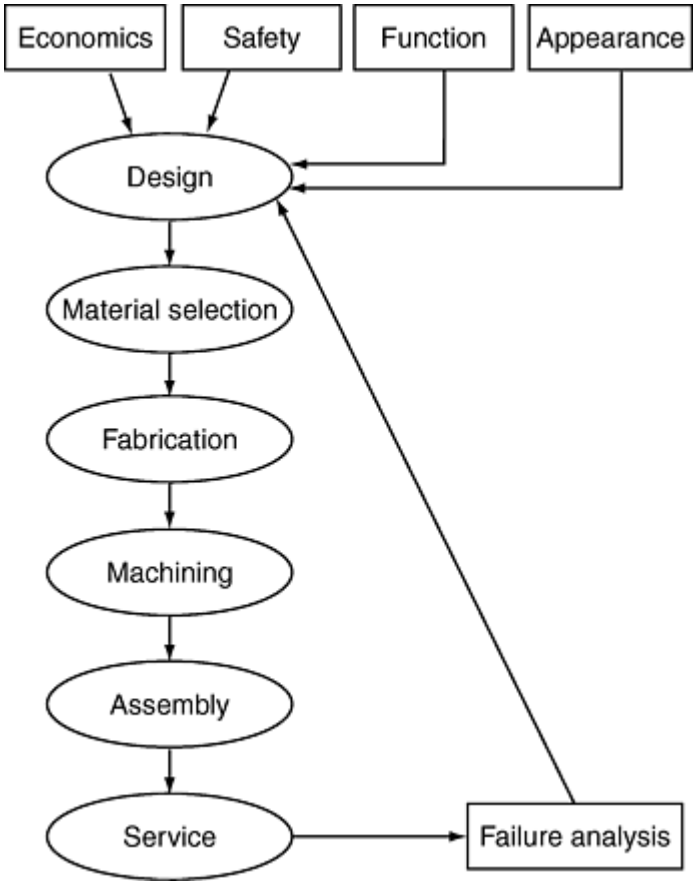


Fig. 45 Feedback of a failure analysis to product evolution. Source: Ref 48

Implementing Corrective Actions. Having completed the failure analysis process and identified the root causes, the next step in preventing future failures involves developing, verifying, and implementing a corrective-action plan. In general, corrective-action plans involve one or more of the three general types of plans: short-term, mid-term, and long-term.

Short-term corrective-action plans involve simple tasks selected to minimize the impact on the operation of the machine or system, such as:

- Repair the inoperative machine or system to get it back up and running
- Identify and manage the suspect population
- Modify the service conditions (make less severe, if possible)
- Issue warnings to other users/maintainers

While these actions may serve to enable continued operation of the machine or system, the amount of service time gained is usually limited. Therefore, mid-term corrective actions can be developed:

- Implement field repair
- Modify design and retrofit in field, at repair facilities, or at the factory

Preventing the failure from recurring in the long run involves implementing long-term “fixes”:

- Redesign
- Implementation of redesign by attrition or retrofit

Standardizing Corrective Actions. Preventing recurrence of failures may require the corrective actions developed through the root-cause failure analysis to be implemented on a much broader basis than for the failed product alone. If similar products are in service that could present risk of failure, those products should be included in the population affected by the corrective actions.

Some corrective actions are standardized internally by the manufacturer, incorporating the actions into division- or corporate-wide standards, such as design guides. Industry associations often step in to standardize corrective actions and lessons learned, in such organizations as ASME in the Boiler and Pressure Vessel Code and the Bridge Welding Code, the API for piping, tanks, and storage vessel inspection programs, the American Welding Society (AWS) for welding standards, the Society of Automotive Engineers (SAE) for fastener standards and material specifications, or the ANSI for safety standards of a wide array of products. The standardization can also be government driven, such as by the FAA, FDA, Consumer Product Safety Commission (CPSC), and so forth in the United States. Clearly, examining failures from a broader perspective enables a much wider impact on preventing failures, through standardization across products, markets, and industries.

References cited in this section

15. J.J. Asperger, Legal Definition of a Product Failure: What the Law Requires of the Designer and the Manufacturer, *Proc. Failure Prevention through Education: Getting to the Root Cause,* 23–25 May 2000 (Cleveland, OH), ASM International, 2000, p 25–29

44. J.E. Shigley and C.R. Mischke, *Standard Handbook of Machine Design,* McGraw-Hill Inc., 1986

45. M.D. Berstein, Design Criteria for Boilers and Pressure Vessels, *J. Pressure Vessel Technol.,* Vol 110, Nov 1988, p 430–432

46. R.C. Juvinall and K.M. Marshek, *Fundamentals of Machine Component Design,* 2nd ed., John Wiley & Sons, 1991

47. R.L. Mott, *Machine Elements in Mechanical Design,* 3rd ed., Prentice Hall Inc., 1999

48. P.A. Thornton and V.J. Colangelo, *Fundamentals of Engineering Materials,* Prentice Hall Inc., 1995

Acknowledgments

The authors gratefully acknowledge the efforts of Dr. Fahmida Hossain, Veda-Anne Ulcickas, Leonard Norman, and Thomas Painter of Massachusetts Materials Research, Inc. for their enlightening analyses and contributions to the case histories presented. We also express our appreciation to Mr. Randy Fach of Dove Consulting for his insights and resources on the topics of Six Sigma, TQM, CI, problem solving, and quality improvement and to Dr. David French for his review and commentary. We especially acknowledge the inspiration and contributions of Dr. Regis Pelloux, Professor Emeritus, M.I.T. Finally, we extend our appreciation to Dr. Walt Griffith of the Air Force Materials Laboratory for his historical perspective on the Wright Brothers, during his informative presentation on “Materials Usage on the Wright Flyer” to a local chapter of ASM International.

References

1. P.L. Jakab, *Visions of a Flying Machine: The Wright Brothers and the Process of Invention,* Smithsonian Institution, 1990, p 226

2. R.W. Hertzberg, *Deformation and Fracture Mechanics of Engineering Materials,* John Wiley & Sons, 1976, p 229–230

3. D.J. Wulpi, *Understanding How Components Fail,* 2nd ed., ASM International, 1999

4. W.E. Deming, *Out of the Crisis,* MIT Center for Advanced Engineering Study, 1986

5. J.M. Juran and F.M. Gryna, Ed., *Juran's Quality Control Handbook,* 4th ed., McGraw-Hill, 1988

6. P.F. Wilson, L.D. Dell, and G.F. Anderson, *Root Cause Analysis: A Tool for Total Quality Management,* ASQ Quality Press, 1993, p 7

7. F.W. Breyfogle III, *Implementing Six Sigma: Smarter Solutions Using Statistical Methods,* John Wiley & Sons, 1999, p xxvii

8. P.S. Pande, R.P. Neuman, and R.R. Cavanaugh, *The Six Sigma Way,* McGraw-Hill, 2000, p xi

9. M. Harry and R. Schroeder, *Six Sigma: The Breakthrough Management Strategy Revolutionizing the World's Top Corporations,* Doubleday & Co., Inc., 1999

10. G.F. Smith, *Quality Problem Solving,* ASQ Quality Press, 1998, p 7

11. B. Anderson and T. Fagerhaug, *Root Cause Analysis: Simplified Tools and Techniques,* ASQ Quality Press, 2000, p 7, 125

12. M. Ammerman, *The Root Cause Analysis Handbook: A Simplified Approach to Identifying, Correcting, and Reporting Workplace Errors,* Max Ammerman/Quality Resources, 1998

13. Engineering Aspects of Failure and Failure Analysis, *Failure Analysis and Prevention,* Vol 10, 8th ed., *Metals Handbook,* American Society for Metals, 1975, p 1–9

14. R.K. McLeod, T. Heaslip, and M. Vermij, Defect or Flaw—Legal Implications, *Failure Analysis: Techniques and Applications,* Conf. Proc. International Conference and Exhibits on Failure Analysis, 8–11 July 1991 (Montreal, Quebec, Canada), ASM International, 1992, p 253–261

15. J.J. Asperger, Legal Definition of a Product Failure: What the Law Requires of the Designer and the Manufacturer, *Proc. Failure Prevention through Education: Getting to the Root Cause,* 23–25 May 2000 (Cleveland, OH), ASM International, 2000, p 25–29

16. D. Broek, Fracture Mechanics as an Important Tool in Failure Analysis, *Failure Analysis: Techniques and Applications,* Conf. Proc. International Conference and Exhibits on Failure Analysis, 8–11 July 1991 (Montreal, Quebec, Canada), ASM International, 1992, p 33–44

17. P.F. Wilson, L.D. Dell, and G.F. Anderson, *Root Cause Analysis: A Tool for Total Quality Management,* ASQ Quality Press, 1993, p 50

18. G.F. Smith, *Quality Problem Solving,* ASQ Quality Press, 1998, p 127

19. M. Paradise, L. Unger, and D. Busch, *TapRoot® Root Cause Tree™ User's Manual,* Systems Improvement, Inc., 1996, p 9–14

20. R.J. Latino and K.C. Latino, *Root Cause Analysis: Improving Performance for Bottom Line Results,* Reliability Center, Inc., 1999, p 79–89

21. C. Nelms, *What You Can Learn From Things That Go Wrong,* 1st ed., Failsafe Network, Richmond, VA, 1994

22. H.P. Bloch and F.K. Geitner, *Practical Machinery Management for Process Plants,* Vol 2, *Machinery Failure Analysis and Troubleshooting,* Gulf Publishing Co., 1983, p 5–6

23. P.F. Wilson, L.D. Dell, and G.F. Anderson, *Root Cause Analysis: A Tool for Total Quality Management,* ASQ Quality Press, 1993

24. G.F. Smith, *Quality Problem Solving,* ASQ Quality Press, 1998

25. R.K. Mobley, *Root Cause Failure Analysis,* Butterworth-Heinemann, 1999, p 37–39

26. M. Ammerman, *The Root Cause Analysis Handbook: A Simplified Approach to Identifying, Correcting, and Reporting Workplace Errors,* Max Ammerman/Quality Resources, 1998, p 67

27. *Failure Analysis, The British Engine Technical Reports,* American Society for Metals, 1981

28. F.A. Hossain and J.J. Scutti, Four Fundamental Root Causes of Failure: Case Histories, *Failure Analysis: A Foundation for Diagnostics and Prognostics Development,* Proc. 53rd Meeting of Society for Machinery Failure Prevention Technology, 19–22 April 1999 (Virginia Beach, VA), 1999, p 185–195

29. J.J. Scutti, "Where Things Go Wrong: Where to Look for Failure Prevention Opportunities," Presented to ASM International Materials Solutions Conference, 9–11 Oct 2000

30. J.J. Scutti and F.A. Hossain, Unanticipated Service Conditions, *Proc. Failure Prevention Through Education: Getting to the Root Cause,* 23–25 May 2000 (Cleveland, OH), ASM International, 2000, p 141–148

31. C.O. Smith and B.E. Boardman, Concepts and Criteria in Materials Engineering, *Stainless Steels and Special-Purpose Metals,* Vol 3, 9th ed., *Metals Handbook,* American Society for Metals, 1980, p 825–834

32. D. Krashes and J.J. Scutti, Poor Surfaces and Intersections of Surfaces Still Cause Trouble Just Like They Used to Do, *Technology Showcase: Integrated Monitoring, Diagnostics and Failure Prevention,* Proc. Joint Conference, 22–26 April 1996, Society for Machinery Failure Prevention Technology (MFPT), 1996, p 681–690

33. F. Hossain and J.J. Scutti, Failure of Components Although the Causes are Simple & Well Documented, *1998 Technology Showcase,* Proc. Joint International Conference, JOAP International Condition Monitoring Conference, 20–24 April 1998 (Mobile AL), JOAP-TSC, 1998, p 455–464

34. G.F. Smith, *Quality Problem Solving,* ASQ Quality Press, 1998, p 53

35. P.F. Wilson, L.D. Dell and G.F. Anderson, *Root Cause Analysis: A Tool for Total Quality Management,* ASQ Quality Press, 1993

36. R.P. Baggerly, Preventing Failures Resulting from Machining Issues, *Proc. Failure Prevention through Education: Getting to the Root Cause,* 23–25 May 2000 (Cleveland, OH), ASM International, 2000, p 118–123

37. A.R. Rosenfield, Fracture Mechanics in Failure Analysis, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 450–456

38. R.M. Pelloux and A.S. Warren, Fatigue Striations and Failure Analysis, *Failure Analysis: Techniques and Applications,* J.I. Dickson, et al., Ed., ASM International, 1992, p 45–49

39. J.C. McMillan and R.M.N. Pelloux, Fatigue Crack Propagation under Program and Random Loads, *Fatigue Crack Propagation,* STP 415, ASTM, 1967, p 505–535

40. M. Cohen, MIT Lecture, 1977

41. G.F. Vander Voort, Ductile and Brittle Fractures, *Met. Eng. Quart.,* Vol 16 (No. 3), 1976, p 32–57

42. E. P. Popov, *Introduction to Mechanics of Solids,* Prentice-Hall, 1968, p 101–110

43. T.L. Anderson, *Fracture Mechanics: Fundamentals and Applications,* 2nd ed., CRC Press, 1995

44. J.E. Shigley and C.R. Mischke, *Standard Handbook of Machine Design,* McGraw-Hill Inc., 1986

45. M.D. Berstein, Design Criteria for Boilers and Pressure Vessels, *J. Pressure Vessel Technol.,* Vol 110, Nov 1988, p 430–432

46. R.C. Juvinall and K.M. Marshek, *Fundamentals of Machine Component Design,* 2nd ed., John Wiley & Sons, 1991

47. R.L. Mott, *Machine Elements in Mechanical Design,* 3rd ed., Prentice Hall Inc., 1999

48. P.A. Thornton and V.J. Colangelo, *Fundamentals of Engineering Materials,* Prentice Hall Inc., 1995

Materials Selection for Failure Prevention

Brett A. Miller, Stork Technimet, Inc.

Introduction

MATERIALS SELECTION is an important engineering function in both the design and failure analysis of components. In design, materials selection can be a complex, iterative process that solves a particular set of engineering objectives for a given component. Materials selection is just one part of this overall design process, which may involve a complex set of relationships regarding product function, shape, materials, and manufacturing process (Fig. 1, Ref 1). In the past, engineering design was performed as a sequential procedure, with the material decisions made last, sometimes literally as an afterthought. After the dimensions and property requirements were identified, the cheapest material meeting those requirements was sought. This philosophy may have been more justifiable when fewer material choices were available or when less sophistication in design and processing was required. However, with the heightened awareness of efficient design, optimized performance, economic competition, environmental impacts, safety concerns, and legal liability, current methods of materials selection are viewed more and more as a simultaneous and integral procedure of the entire design process, even during the early stages of design.

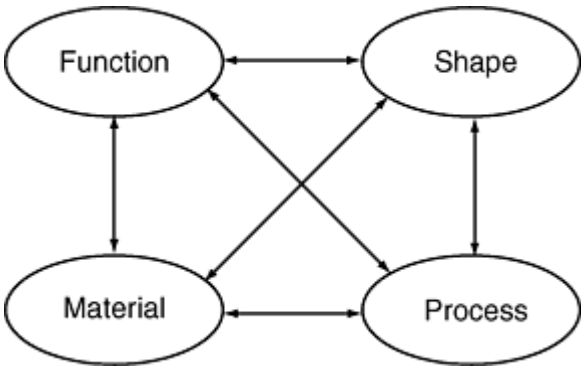


Fig. 1 Interrelated factors involved in the design process. Source: Ref 1

Materials selection and design are also closely related to the objectives of failure analysis and prevention. These processes are inextricably intertwined, probably to an extent that is not readily apparent to most engineers. Failure analysis augments the development process by real-time identification of design inadequacies, providing opportunities for optimization. Finding the root cause of a failure also often takes as much imagination as the original design concept. In fact, failure analysis can be viewed as a figurative reassembly of the component in the original condition, or design in reverse. For example, Fig. 2 is a diagram comparing the general procedures for both engineering design and failure analysis. The basic philosophies of the two processes are reversed. Design is the process of synthesizing and analyzing conditions into the reality of an actual or hypothetical component. In contrast, failure analysis is the dissection of an actual component in order to synthesize and understand the significance of a hypothetical design in a given failure. It is important to note that the analysis and synthesis of engineering factors are prominent in different areas of each process, although the individual steps within the processes contain both.

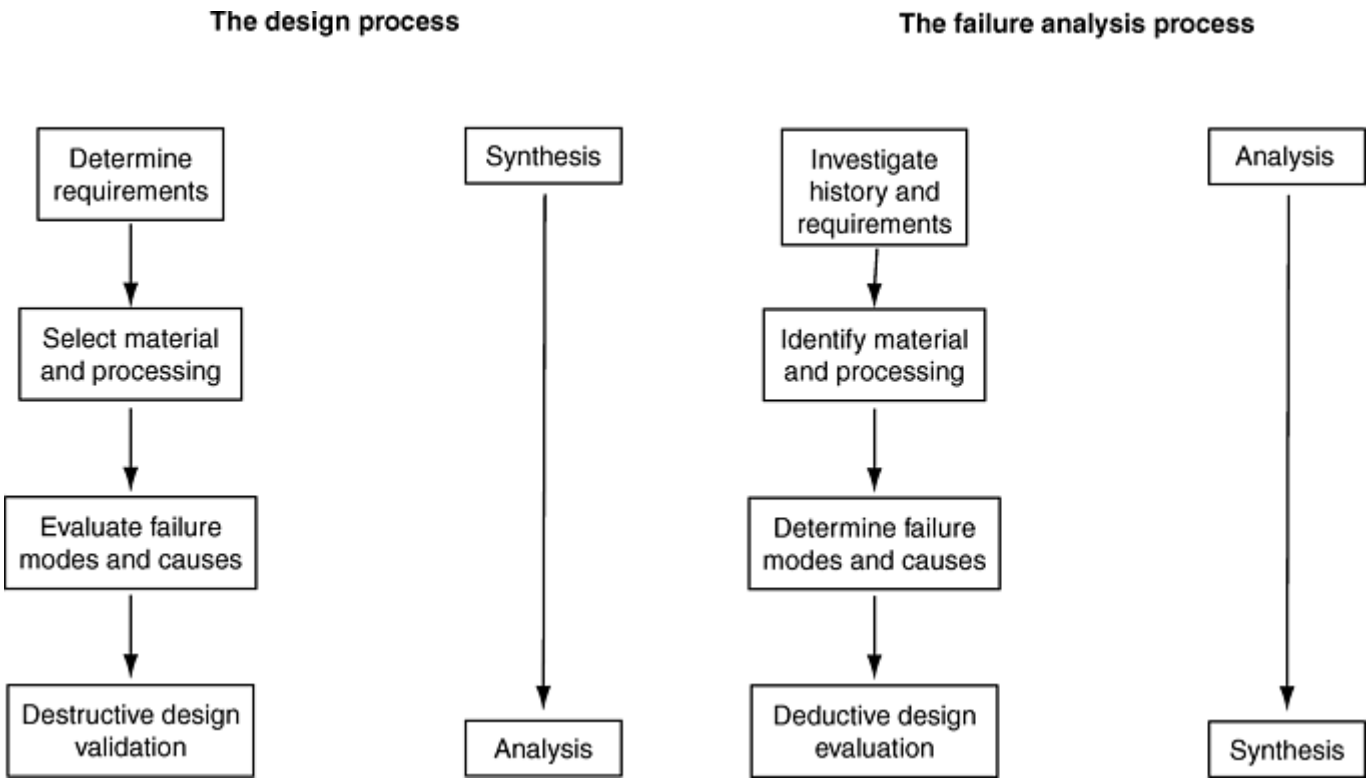


Fig. 2 General steps and the roles of synthesis and analysis in the processes of design and failure analysis

Materials selection augments and is supported by failure analysis in several ways. Of course, a basic objective of any successful design is to prevent failure (which can be defined here in the general context of a part not being able to perform its intended function). Therefore, failure analysts need to understand the underlying principles and practices of design and materials selection as basic tools in failure prevention. However, failure analysts also recognize that the synergistic effects of service conditions, manufacturing effects, and material characteristics are not always captured within the axioms and discrete attributes of a design process. Therefore, the analysis of failed parts can provide important insights for metallurgists and other engineers involved with design in general and materials selection in particular.

Design also sometimes can be an emulative process, whereby a successful design is adapted for a similar but separate service condition. Emulative design assumes that the new intended service is analogous, that the materials and processing are the same, and that the knowledge of the prior design is complete. Using prior design as a pattern involves implicit and possibly unappreciated assumptions, which may not account for the synergistic effects of service conditions, manufacturing effects, and material characteristics. It is easy to imagine, for example, that one structural member in an assembly may appear to be sufficiently strong when, in actuality, its portion of the load may have been displaced and accommodated by the overdesigned, surrounding structure.

This article briefly reviews the general aspects of materials selection as a concern in both proactive failure prevention during design and as a possible root cause of failed parts. This article cannot detail the many particulars of materials selection, because every industry or component application has many specific requirements, guidelines, or procedures, some of which may be mandated by federal or state statute. Therefore, coverage is more conceptual with general discussions on the following topics:

- Design and failure prevention
- Materials selection in design
- Materials selection for failure prevention
- Materials selection and failure analysis

Because materials selection is just one part of the design process, the overall concept of design is discussed first in the section “Design and Failure Prevention.” The next section, “Materials Selection in Design,” then describes the role of the materials engineer in the design and materials selection process. The other sections of this article focus on the significance of materials selection in both the prevention and analysis of failures. Portions of this article contain adapted content from *Materials Selection and Design,* Volume 20, *ASM Handbook,* with citation to more detailed references on materials selection.

Reference cited in this section

1. H.W. Stoll, *Product Design Methods and Practices,* Marcel Dekker, 1999, p 40, 148

Materials Selection for Failure Prevention

Brett A. Miller, Stork Technimet, Inc.

Design and Failure Prevention

The basis of all engineering is design, and the terms are often used synonymously. The primary objective of design is to develop a useful component or structure that performs an intended function in as safe a manner as possible. Therefore, the prevention of failure (generally defined here as any loss of intended function) is a principal concern of any design process. Simply restated, the primary measure of a successful design process is foreseeing and avoiding failure.

Design generally requires specific engineering expertise and is performed by a wide variety of engineering disciplines, such as:

- Civil engineers design large structural forms, such as bridges, highways, buildings, and power-generation and water supply facilities. Codes and standards regulate many of the materials and design features of these structures, due to safety concerns. Civil engineering designs typically use reliable and economical materials that are not particularly exotic.
- Mechanical engineers design a wide variety of components, such as pressure vessels, vehicles, and machinery of all types. These designs often contain moving parts and use materials with highly specific performance requirements. Materials selection can include all materials and processes. Many codes and standards are also applicable to mechanical engineering designs. Mechanical engineering encompasses such a broad range of equipment that designers generally have expertise in specific functions or types of component.
- Chemical engineers typically use materials in various chemical- and petrochemical-processing industries with design requirements involving corrosion resistance and elevated-temperature service.

Many other engineering disciplines also have unique and special requirements for design and materials selection. In electrical engineering design, for example, the physical property requirements (e.g., magnetic, electrical, electronic, or thermal properties) typically supersede the mechanical property requirements. Other design disciplines include industrial, automotive, welding, mining, aerospace, nuclear, and computer engineering.

Each type of engineering discipline requires specialized design expertise that is beyond the scope of this article. However, the general process of engineering design can be described as an iterative procedure that can be roughly divided into two basic stages (Fig. 3, Ref 1):

- *A conceptual design stage* involving the definition of product specification (or functions) and the underlying physical concept and preliminary layout to achieve the intended functions
- *A detailed design stage* involving both the qualitative definition of part configuration and the quantitative analysis of design parameters (e.g., dimensions, tolerances, materials properties, etc.) to achieve a final layout for a product, assembly, or system

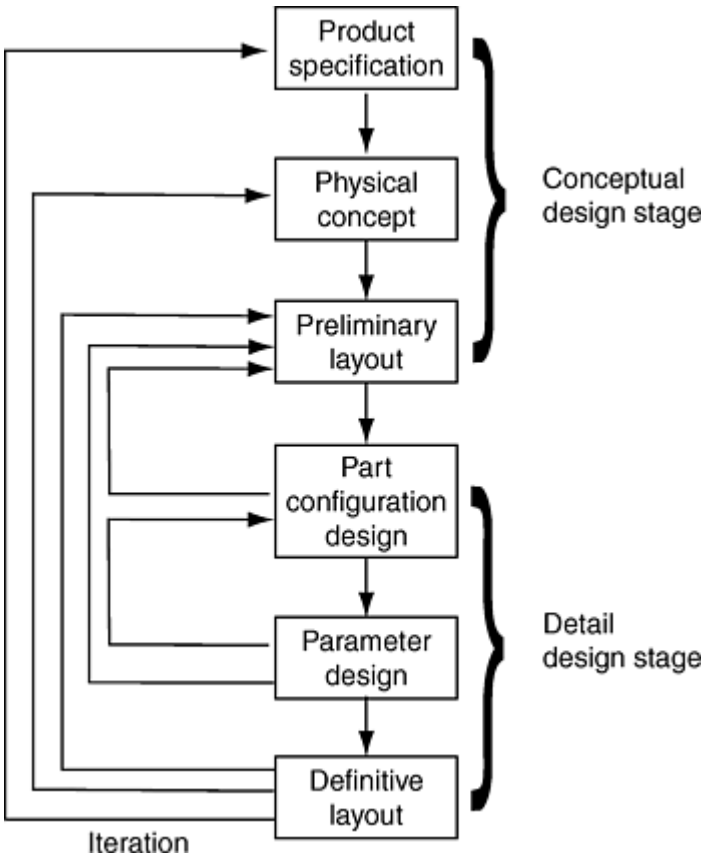


Fig. 3 Stages and steps in the iterative process of design. Source: Ref 1

As shown in Fig. 3, iteration is a key feature for any step in the design process. In fact, engineering design can be thought of as a process of guided iteration (Ref 2), which is a problem-solving methodology that formulates a problem, generates alternative solutions, evaluates the alternatives, and redesigns from the evaluation results if they are unacceptable. This methodology is fundamental to design processes. It is repeated hundreds or thousands of times during a complicated product design. It is used again and again in recursive fashion for the conceptual stage to select materials and processes, to configure parts, and to assign numerical values to dimensions and tolerances (i.e., parametric or parameter design).

Designers and materials engineers are key participants in this iterative process, and failure analysts also need to appreciate and understand the roles and activities of the design process. This provides the basis to identify possible technical root causes of failures and to advise or recommend further design iterations either on a proactive basis or as a lesson learned. Therefore, the basic steps or stages of engineering design are briefly described in the following sections, beginning with the stage of conceptual design and ending with a brief overview on methods of risk assessment in design. The main focus is on the design of part, as opposed to the design of a system or assembly. Systems or assemblies may have many components that involve parameters beyond the typical dimensions, tolerances, and properties of a distinct part. For example, when a system involves human interaction, the design process must address the influence of human factors. This includes not only human factors in equipment design but also a wide range of activities that can include operational methods and procedures, testing and evaluating these methods and procedures, job design, development of job aids and training materials, and selection and training of people (Ref 3).

Conceptual Design

The first stage of design is to determine the physical concept by which the product will function. This includes the physical principles by which the product will work and an abstract physical model that will employ the principles to accomplish the desired functionality. For example, suppose the required function is simply to support a load over an open space. In this case, the physical model could be a beam of uniform cross section or truss. In addition, there is not usually a unique solution for implementing a physical concept, although a concept and its function are inextricably linked.

When a product is more complex, it consists of an assembly of subassemblies and parts. In this case, the physical concept of the system or assembly must be “decomposed” into a set of principal functional subassemblies. For example, an automobile is a set of subassemblies identified as the engine, drivetrain, frame, body, suspension system, and steering system. The physical principles thus require sufficient information about how each of these functional subassemblies will interact with all of the others to accomplish the required product functions. The term “decomposition” is generally used to

describe the part of the design process that identifies the subassemblies comprising a product or larger assembly. That is, in the conceptual design of an automobile, it could be decomposed into the engine, drivetrain, frame, and so forth.

Decomposition in Conceptual Design. Two basic methods of decomposition are used in conceptual design: physical decomposition and functional decomposition. Many design concepts are based on the method of physical decomposition, when existing or emulative designs are used with an implicit functionality. For example, the decomposition of an automobile into engine, drivetrain, frame, body, and so on is an example of physical decomposition. This method is common, but the method of functional decomposition also has benefits. In functional decomposition, the design concept is defined purely in terms of functions, with physical embodiments (or configurations) selected to fulfill the functions. In an automobile, for example, the function of the engine is to convert a source of on-board energy to rotational mechanical power. This function could be achieved with the usual internal combustion engine, but it could also be provided by an electric motor, a turbine powered by compressed gas, human-powered pedals, and many other alternatives. In the case of an automobile, the available alternative sources of power are very familiar. In a new, less-familiar product, however, the advantage of function-first decomposition is that it stimulates designers to consider many ways of fulfilling a given function instead of choosing the most common embodiment that comes to mind.

The whole purpose of decomposition is to provide a systematic description of parts that make up a system or assembly. For an initial concept, it is usually sufficient to perform only one level of functional or physical decomposition, but all subassemblies thus created will ultimately, as a part of their own conceptual design, be decomposed again and again. For example, a lawn mower engine may be decomposed into, among other things, an engine block and a carburetor. Then, in turn, the carburetor may be decomposed into, among other things, a float and a cover. Thus, the process of conceptual decomposition repeats (or recurs) until no new subassemblies are created, that is, until only parts or standard components are obtained. Physical decomposition and functional decomposition are not always mutually exclusive; they also can be used simultaneously in a design.

Conceptual Design of a Part. When the process of conceptual decomposition is completed and a list of parts is obtained, then the next step is conceptual design for each individual part. Conceptual design of a part involves the following steps (Ref 4):

- Determining whether the part is really necessary
- Identifying the required functions of the part
- Selecting the material and a manufacturing process for production

Definition of part function is the essential objective of conceptual design, and examples of common part type or features are listed in Table 1 for various functions. Some parts may have more than one function, and often parts have special features to enhance manufacturing or reduce material costs. Examples of these are described in Table 2. For economic reasons, the process of conceptual design may also address the possibility of eliminating a part or combining functions, because one complex part may be less expensive overall than two or more simpler parts. This step may involve consideration of materials and manufacturing costs and economics, but the reference book by Boothroyd and Dewhurst (Ref 5) provides one relatively easy method for determining whether a proposed part should be assembled from separate components.

Table 1 Functions served by parts

Function	Examples of part types or features
Transmit or support force(s) or torque(s)	Brackets, beams, struts, columns, bolts, springs, bosses, knobs
	Levers, wheels, rollers, handles
	Parts that fasten, hold, or clamp, such as bolts, screws, nails
Transmit or convert energy	
Heat	Heat fins, electric resistance heating elements
Mechanical power	Shafts, connecting rods, gears
Electricity	Wires, lightbulb elements, resistors
Provide a barrier (for example: reflect, cover, enclose, or protect)	
Light	Walls, plugs, caps
Heat	Thermal insulators, thermal reflecting surfaces
Electricity	Electrical insulators, magnetic shields
Sound	Walls, sound-absorbing wall surfaces
Control motion	Cams, grooves, slots, gears
Allow passage (of light, rods, shafts, wires, pipes, etc.)	Holes, windows, grooves
Control or regulate the passage of	
Fluids	Nozzles, orifices, pipes, ducts

Function	Examples of part types or features
Light	Shutters, wheels
Indicate	Clock hands, instrument needles, colors, embossing
Locate or guide	Grooves, holes, bosses, tabs, slots

Source: Ref 4

Table 2 Special features designed into parts to aid manufacturing or to reduce material cost

Function	Examples of part features
Aid manufacturing	Fillets, gussets, ribs, slots, holes
Add strength or rigidity (e.g., stiffen)	Ribs, fillets, gussets, rods
Reduce material use	Windows or holes through walls, ribs that allow thinner walls, slots
Provide a connection or contiguity (so the part can be a single part)	Walls, rods, ribs, gussets, tubes

Source: Ref 4

The process of conceptual design may also involve at least a preliminary decision on a material and manufacturing process to be employed. A physical concept of the materials and manufacturing process is generally required here, because most designs can never proceed very far without this information. In a more detailed approach to engineering design, Dixon and Poli (Ref 2) suggest a four-level approach to materials selection:

- *Level I:* Based on critical properties, determine whether the part will be made from metal, plastic, ceramic, or composite.
- *Level II:* Determine whether metal parts will be produced by a deformation process (wrought) or a casting process; for plastics, determine whether they will be thermoplastic or thermosetting polymers.
- *Level III:* Narrow options to a broad category of material. Metals can be subdivided into categories such as carbon steel, stainless steel, and copper alloys. Plastics can be subdivided into specific classes of thermoplastics and thermosets, such as polycarbonates and polyesters.
- *Level IV:* Select a specific material according to a specific grade or specification.

In this approach, materials and process selection is a progressive process of narrowing from a large universe of possibilities to a specific materials and process selection. Level I and level II often may suffice for conceptual design, while level III is needed for embodiment (configuration) design and sometimes for conceptual design. Level IV usually can be postponed until detail (parametric) design.

The four levels of materials selection in the previous list are just a starting point in narrowing options, because the process of materials selection requires the evaluation of many factors, as briefly summarized in more detail in the section “Materials Selection in Design” in this article. However, the key point is that materials selection is an up-front concern with important consequences for processing, product design, cost, availability, recyclability, and performance of the final product. This is why materials and processes selection can be a critical issue in the early stages of design. Moreover, the proliferation of new and specialized engineering materials has changed the complexion of design to the point that no engineer in a design capacity is conversant in all families of potential materials that can be used. The more critical an application is, the more important the materials selection becomes. Specialized materials expertise is mandated by the complexity of critical service, which very often includes extremes of temperature, stress, environment, or all three, as in the case of jet engine components. The services of a materials engineer should be obtained to foresee the complex material and property interactions and synergistic effects that may be attendant to a design. It is a logical conclusion that, particularly for complex and critical engineering designs, a cross-functional approach is best. A materials engineer might specify the materials and associated processes for an engine part but may not be able to design one. Similarly, a mechanical engineer may design an engine part, but may not be able to determine the materials and processes necessary for fabrication.

Integrated Product Development Teams. The integration of diverse engineering disciplines in design and materials selection is important even in the conceptual design. One form of integrating engineering functions to optimize design is the integrated product development (IPD) team concept. One of the strengths of the design team approach is that all disciplines have input early in the process, while decisions are easily changed and inexpensive improvements can be made. Cross-functional IPD teams are formed temporarily in many organizations for a particular product but are also formed somewhat permanently in others for continuous design support.

The IPD approach has been shown to lead to better results faster. For example, the use of an IPD team approach can be useful during configuration design (described subsequently), when designers may inadvertently create parts with geometric features that place severe restrictions on the selection of manufacturing processes, with even less freedom

remaining for materials selection. Similarly, overly restrictive and independent selection of the material will limit the manufacturing processes available. This is all the more reason to use IPD methods. However, until the IPD approach is in more common use, an alternative approach, referred to as a materials-first approach, may be useful. The materials-first approach depends on a thorough understanding of the service environment and advocates choices based on properties that satisfy those performance needs (see the section “Materials Selection in Design” in this article).

Other Specialized Approaches and Tools. Many other specialized design team approaches have been developed to evaluate design, primarily for critical systems where failure can be catastrophic. These methods include failure mode and effects analysis, failure mode, effect, and criticality analysis, fault tree analysis, and fault hazard analysis. These formalized methodologies use systematic evaluation and sophisticated computer programs to predict failure in complicated designs and can be an invaluable aid in materials selection. The section “Risk Assessment in Design” briefly reviews the use of risk and hazard analysis in design.

In addition to multifunctional design approaches, sophisticated design tools have been developed to assist in preventing failures. Fracture mechanics is often used to create flaw-tolerant designs for critical applications. Finite-element analysis (FEA) techniques have been implemented to dynamically evaluate the effects of material characteristics and geometry changes. These tools also can be of great benefit during configuration design, parametric design, design validation, or systematic investigation of a failed component. More details on these methods are discussed in other articles in this Volume.

Configuration Design (Embodiment)

The first step in the detailed design stage of Fig. 3 is configuration design. After the preliminary steps of concept design have been completed, the designers must define the features of the configuration. Ultimately, designers must determine exact numerical values for the dimensions and tolerances of parts during parametric design. However, before this can be done, designers need to define the general configuration of a part in terms of its physical arrangement and connectivity.

Configuration design is a qualitative (i.e., nonnumerical) process that defines the general features of a part in terms of functional interactions with other parts or its surrounding environment. These interactions include forces (loads and available support areas), energy or material flows, and physical matings or other spatial requirements (e.g., certain spaces may be unavailable to the part). The types of dimensional features that are defined during configuration design may include (Ref 4):

- Walls of various kinds, such as flat, curved, and so forth
- Add-ons to walls, such as holes, bosses, notches, grooves, ribs, and so forth
- Solid elements, such as rods, cubes, tubes, spheres, and so forth
- Intersections among the walls, add-ons, and solid elements

Usually, there are several—and sometimes many—ways to configure a part, and the best approach is generating, evaluating, and modifying a number of alternatives. A three-dimensional sketch that shows these interactions to approximate scale is generally a very helpful starting point of configuration design. The sketch shows the essential surroundings of the part and locates loads, possible support points or areas, heat or other energy flows, adjacent parts, forbidden spaces, and so on.

Configuration designs and the various alternatives need to be evaluated before numerical dimensions and tolerances are established. As described in Ref 4, this evaluation process can be guided by qualitative physical reasoning about the functionality of the part configuration and manufacturing. Even when actual dimensions have not been determined in the configuration stage of part design (i.e., when sizes and spatial relationships of the features are still only approximate), knowledge of physical principles and manufacturing processes can still be applied to help create the most effective alternative designs for further evaluation. General physical reasoning involved in the generation of part configuration and manufacturing alternatives is discussed in more detail in Ref 1, 2 and 4.

In addition to qualitative physical reasoning about functionality, effective part configurations also are strongly influenced by manufacturing issues and materials selection. At this point in the part design process, it is necessary to decide on a manufacturing process and at least a class of materials (e.g., aluminum, thermoplastic, steel). However, unless the information is needed for evaluation of the configurations, selection of the exact material (e.g., the particular aluminum alloy or thermoplastic resin) may be postponed until the parametric stage. Consultation with materials and manufacturing experts is, of course, strongly advised, and other factors, such as recycling concerns and existing business relationships, also may be relevant.

Finally, once the set of the most practical part configurations has been generated, a more formal evaluation should be performed. The evaluation can be done by Pugh's method (Ref 6) or by other methods presented in Ref 4. In any method, the comparison criteria for alternative configurations should include the following:

- *Functionality:* Can exact dimensions and tolerances be imposed that will enable the part to perform its function properly and reliably?
- *Use of materials:* When dimensions are imposed, will the configuration provide for efficient use of all the required material?
- *Mechanical failure:* When dimensions are imposed, can the risks of failure from mechanical causes, such as fatigue, excessive stress, buckling, and so forth, be made suitably low?
- *Analyzability:* Does the configuration enable analyses to be performed for stresses, vibrations, heat flow, and so forth?
- *Manufacturability:* Can the selected manufacturing process hold the tolerances that will be needed for the configuration to meet the required functionality? Does the configuration allow for ease of handling and insertion for assembly? Are there special issues that will influence the time required for tooling and production?

Parametric Design and Analysis

Conceptual and configuration designs are based primarily on qualitative reasoning about physical principles and manufacturing processes. In parametric design, however, numerical computations become much more important. The attributes of parts identified at the configuration stage become the design variables, which must be identified and analyzed during the step of parametric design.
Evaluation in parametric design requires computation of performance parameters as well as selection and implementation of a method for evaluating the overall quality of the trial design. During parametric design, the design or process engineer seeks to optimize performance by:

- Identifying design variables and their allowable range
- Identifying performance parameters whose values will be computed or measured to evaluate the performance of trial designs
- Identifying the analysis methods that will be used to compute values for the evaluation

Often, parametric design values and procedures are governed by design codes, standards, or test methods. These may be specific to a particular discipline, industry, or class of material. In addition, design tools such as fracture mechanics and FEA can be used to analyze designs and evaluate the effects of material characteristics and geometry changes. These tools also can be of great benefit during configuration design, parametric design, and design validation.

Risk Assessment in Design

All designs balance expected benefits against potential risks. Therefore, the notion of explicitly looking for technical and manufacturing risks in a design is useful. The idea is to look for any previously unquestioned assumptions that have been made while generating the configuration and while doing the evaluations—that is, to look for issues that may so far have been overlooked.
Risk cannot be avoided completely, even for very conservative designs. Indeed, good designs will have some reasonable risk or they will be too conservative, too costly, too heavy, and so on. However, designers cannot count on good luck; there are many more ways for a design to fail than there are for it to succeed. Thus, risks must be sought out, faced, and evaluated. Table 3 lists questions intended to reveal risks related to part configurations.

Table 3 Questions for revealing part configuration design risks

Factor	Questions
What are the most likely ways the part might fail in service?	
Excessive stress	Can the part be dimensioned to keep stresses below yield failure levels? Add ribs? Use stronger material?
Fatigue	If there will be cyclic loads, can the configuration be dimensioned so as to keep the internal stresses below the fatigue limit?
Stress concentrations	Can the part be dimensioned to keep local stress concentrations low?
Buckling	If buckling is a possibility, can the configuration be dimensioned to prevent it?
Unexpected shocks or loads	What unexpected dynamic loads might be encountered in service or in assembly? Can these be handled by the configuration?
What are the most likely ways the part might not meet its expected functionality?	
Tolerances	Is the configuration such that functionality will be especially sensitive to the actual tolerances that

Factor	Questions
	can be expected in a production situation? Are too many special (tight) tolerances required to make the part work well?
Creep	If creep is a possibility, will it result in loss of functionality?
Strain and deformation	If functional performance is sensitive to retention of size and shape, can the configuration be dimensioned to preserve the required integrity?
Thermal deformations	Might thermal expansion or contraction cause the configuration to deform so that function will be impaired?
Handling and assembly	Might there be unforeseen difficulties with handling and assembly?
Dimensions	Might the part end up being dimensioned so that assumptions about assembleability become invalid?
Tangling	Might the parts tangle if dimensioned in some way?
Will the available production machines be able to make the part?	
Production runs	Are the desired production runs consistent with the machines and expected costs?
Tooling wear	Is tooling wear or maintenance a possible problem that will impact part cost or performance?
Weld lines[(a)]	If the process is a flow process, can weld lines be located appropriately?
Other design and materials factors	
Geometric compatibility	Is the part geometrically compatible with its adjoining parts? What could go wrong is this regard? If there is a small change in this part, or in an adjoining part, can the configuration accommodate the change without major redesign? What about the effects of tolerances of the *adjoining* parts? Or on the assembly as a whole?
Materials	Is the material selected compatible with the configuration and the manufacturing process? Is surface finish properly accounted for? Will standard raw material supplies be of adequate quality? Has the material been thoroughly investigated for its use in *this* particular application? Are there previous uses in similar applications? Have experts on the properties and processing of the material been consulted? Is the material compatible with the rest of the product?
Designer and design team knowledge	Has every possible, unfortunate, unlikely, unlucky, even stupid "What if …" situation been considered? Are there aspects of the part design where the designer or design team is working without adequate knowledge? Where is the design based on insufficient knowledge of materials, forces, flows, temperatures, environment, etc.? Where are there guesses, hopes, fears, and assumptions instead of knowledge: Materials? Stresses? Fastening methods? Manufacturing process? Tolerances? Costs? Adjoining parts? Environmental conditions?

(a) A weld line is formed when a material flow must divide—say around a hole—and then rejoin. The weld lines tend to be weaker and more subject to fatigue failures. Source: Ref 4

Risk and Hazard Analysis. One tool used in the evaluation of risk is the process of risk and hazard analysis, which helps identify the level of risk and to pinpoint the parts of the system that represent the greatest risk for failure. If the analysis is used properly, steps can be taken to eliminate the cause or reduce the risk to an acceptable minimum. Some hardware systems approaching a "failure-free" condition may be produced when actions are taken at all levels that are based on:

- Attention to past experiences with similar systems
- Availability of risk information for all project personnel
- A sound, aggressive risk and hazard analysis during all design phases
- Development of suitable corrective action and safety programs based on the analysis
- A continuous and searching review of all phases of the program efforts

Rigorous applications of risk and hazard analysis have made difficult technological feats, such as landing on the moon, relatively accident-free.

The various analysis techniques of risk assessment have grown out of the search for system reliability. Consequently, the approach is hardware-oriented, with the emphasis on ensuring that hardware is able to perform its intended function. Backup systems and redundancies are also used to reduce such risks. Through cost/benefit analysis, the performance of the system will have a computable value that can be compared to the cost of accomplishing the objectives desired for a product or system.

Risk and hazard analysis tools have been developed to ensure system reliability in critical applications. With the increased emphasis on safety, reliability, and achieving performance objectives, design teams must incorporate risk/hazard considerations in their designs. Figure 4 (Ref 7) is a flow chart that shows the integration of risk and hazard analysis in the overall design process. Even if designers or design managers are not directly responsible for carrying out these analyses, they must be familiar with the methodology, so that they understand how they are carried out and how they can

respond in terms of design or system changes. Most efforts are best carried out during early design phases, and they can be effectively used during design reviews to provide valuable feedback to the design to avoid failures. More information on the principal methods of risk/hazard analysis is presented in Ref 7.

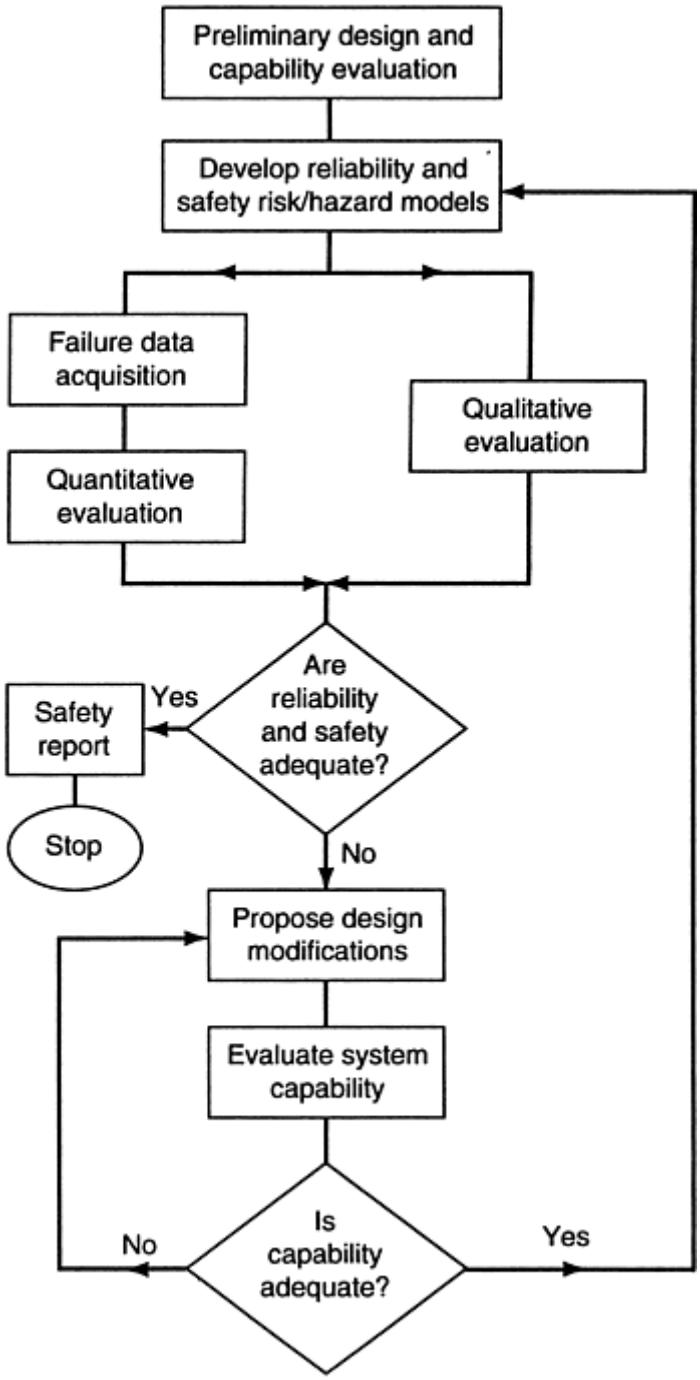


Fig. 4 Flow chart showing the integration of risk and hazard analysis into the design process. Source: Ref 7

References cited in this section

1. H.W. Stoll, *Product Design Methods and Practices,* Marcel Dekker, 1999, p 40, 148
2. J.R. Dixon and C. Poli, *Engineering Design and Design for Manufacturing,* Field Stone Publishers, 1995
3. C.O. Smith, Human Factors in Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997

4. J.R. Dixon, Conceptual and Configuration Design of Parts, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 33–38

5. G. Boothroyd and P. Dewhurst, *Product Design for Assembly,* Boothroyd Dewhurst, Inc., 1989

6. S. Pugh, *Total Design: Integrating Methods for Successful Product Engineering,* Addison-Wesley, 1991

7. G. Kardos, Risk and Hazard Analysis in Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997

Materials Selection for Failure Prevention

Brett A. Miller, Stork Technimet, Inc.

Materials Selection in Design

One of the chief concerns of any design or engineering effort is materials selection. Materials selection is a process whereby the function and desired final properties of a component are evaluated during all the various stages of design in order to identify suitable materials of construction. During every stage of the design process (i.e., conceptual, configuration, and parametric), some level of materials selection must be made in order to proceed with the design. This is one reason why, as previously noted, integrated product development (IPD) teams have been used.

Moreover, the options in materials selection have proliferated. The number of materials currently available for designers has grown, as shown by the timeline in Fig. 5 (Ref 8). This trend will probably continue, thus making the function of materials selection more difficult than it was many years ago. The manufacturing processes available to designers have also grown substantially. These are additional reasons why IPD teams are used, although perhaps on a limited or temporary basis.

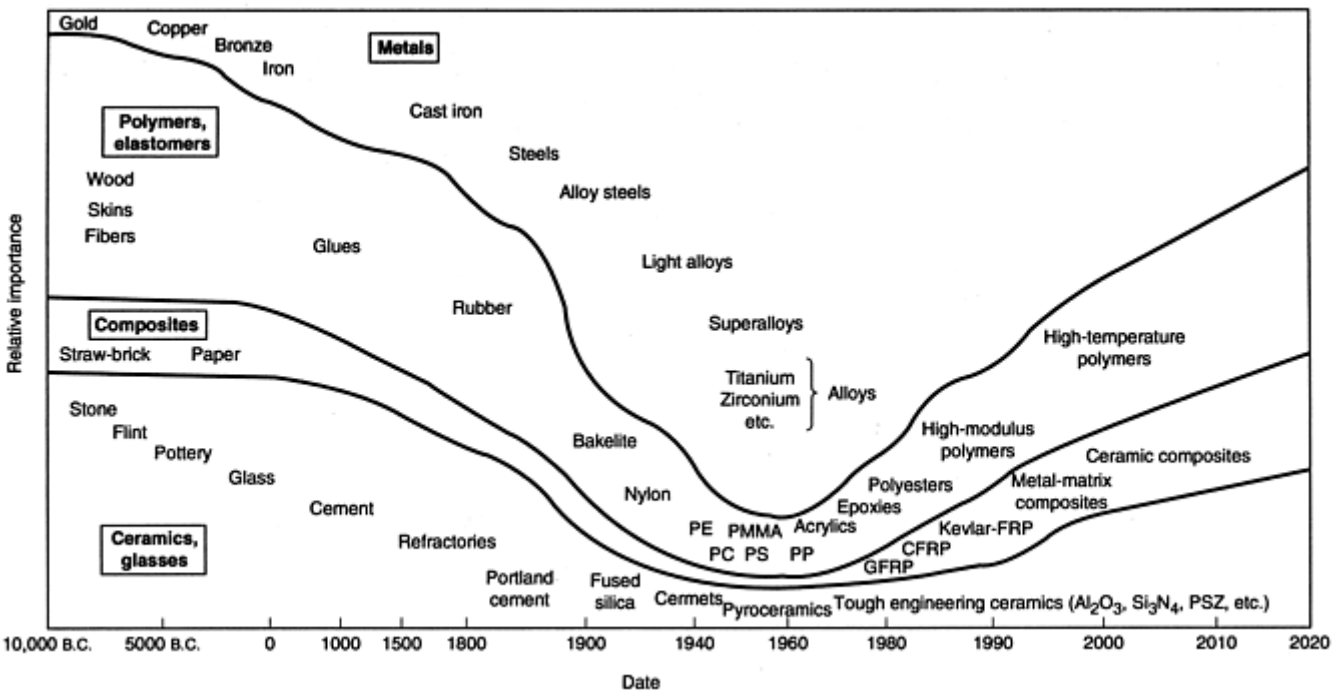


Fig. 5 The evolution of engineering materials through history. PE, polyethylene; PMMA, polymethylmethacrylate; PC, polycarbonate; PS, polystyrene; PP, polypropylene; CFRP, carbon-fiber-reinforced plastic; GFRP, graphite-fiber-reinforced plastic; PSZ, partially stabilized zirconia. Source: Ref 8

Until the IPD approach is in common use, and as an alternative to the traditional “materials-last” approach, another method of materials selection is the “materials-first” approach. The materials-first approach depends on a thorough understanding of the service environment and advocates choices based on properties that satisfy those performance needs. After the component is envisioned, the environment is evaluated, and the constraints are applied, the design concept (including preliminary selection of materials, processes, and product form) can be developed further during configuration design. This materials-first approach involves the selection of a general class of materials during conceptual design and further refines the alternatives through the iterative process of design. The selection process may involve a large set of performance and property criteria, which must be refined and developed during design. However, there are four elementary topics in materials selection that must be addressed during all stages of design. They are posed as simple questions:

- What is it?
- What is its environment?
- What cannot it be?
- What must it be?

These questions can be viewed as conceptual constraints or factors that influence materials selection in the stage of conceptual design. Then, of course, more detailed materials selection criteria and properties are used during the stages of configuration and parametric design.

The process of materials selection varies substantially for different purposes, because steps are typically amended to suit specific applications, and this process is sometimes formalized to ensure all of the steps have been thoroughly performed. This process is dynamic, because changes in design have to be considered in an ongoing fashion. Proper materials selection is a dynamic function that must accompany all design activities while remaining sufficiently flexible to accommodate inevitable engineering design changes. It is an aspect of design that is every bit as crucial as part dimensions and geometry. Proper materials selection is necessary during new design and for the improvement of existing designs that are found to be marginally or completely unsuitable.

Identification of the materials selection as a constant contributor to the design of a component also includes the necessity of incorporating requisite materials information on engineering drawings. Because a complete part drawing is a metric for establishing whether a fabricated component is acceptable, all pertinent characteristics must be identified. The specificity of materials and processing information required for thorough identification on the engineering drawing is a function of the complexity of the component. Newly created drawings are typically electronic computer-aided design constructs. These drawings require less storage space but can be as prone to human errors as the time-honored autographic drafting methods.

Many companies cross-reference stand-alone materials specifications that contain much more detailed purchasing and processing information than could be conveniently placed on the part drawings. This practice has an additional benefit in that a large number of drawings using the same material can be upgraded simultaneously without having to revise each drawing individually. With respect to materials and process information on drawings, some standardized symbols have been created. The symbols in current use include welding joint design and instructions, dimensional tolerances, and finish symbols. These symbols provide a great amount of information without appearing as additional written process descriptions and hence present information more clearly and effectively on drawings.

Materials Selection during Concept Design

At the concept level of design, materials and processes are considered rather broadly. The decision is to determine whether each design concept will be made from metal, plastics, ceramic, composite, or wood and to narrow it to a group of materials. The precision of property data needed is rather low. If an innovative choice of material is to be made, it should be done at the conceptual design step. Materials selection at this stage of design may use tools such as material property charts or general performance indices (e.g., see the articles “Material Property Charts” and “Performance Indices” in *Materials Selection and Design* Volume 20, *ASM Handbook*.)

The four fundamental questions of materials selection should also begin during the stage of conceptual design, where all the functional physical conditions (including any major economic and nontechnical conditions) are imposed. Simply, the component material is defined by what it is, by its intended environment, by what it cannot be, and by what it must do. Often, these criteria, although oversimplified here, can be a good acid test in narrowing the possible alternatives of material classes during conceptual design.

The Design Objective—What It Is. This objective is typically a simple component description before all of the requirements are identified. The designer must ensure separation of what something is from what it does. In other words, the physical embodiment (or configuration) may be related to the intended function, but other configurations may perform the same function. It is relatively easy to lose sight of the very basic utility of a part or structure during design by

inadvertently insinuating arbitrary constraints. This objectivity may be confusing, but incorrect assumptions during design conception can be difficult to surmount later.

The Design Environment—What Its Environment Is. A structure or component can be affected by a wide variety of service environments that may include:

- Temperature extremes
- Temperature fluctuation
- Alkalinity
- Acidity
- Pressure
- Oxygen content
- Flammability
- Flame impingement
- Humidity (wet/dry cycles)
- Galvanic differences
- Moisture
- Liquid metal
- Flow/flow rate
- Erosion
- Cavitation
- Hydrogen content
- Biological agents

This list shows many of the environmental factors that must be kept in mind during the design process, before any thought is given to properties. When the service conditions are not adequately understood (as in the early days of the space program), design can become a costly, iterative process requiring extensive trial-and-error bench testing.

The environment also must be evaluated prior to identifying necessary properties, because mechanical and physical properties can be severely altered by environmental factors. Very specific environments are often considered, including extremely corrosive high-temperature or high-pressure applications.

Design Constraints—What It Cannot Be. Design constraints are industry-specific or self-imposed restrictions on the materials or processes that may be considered in the design process. These constraints come from a variety of sources. Sometimes material constraints are applied by the end user, which may dictate exact materials and processes as a contractual obligation.

Constraints can act as an aid in the design process, because they obviate the consideration of certain prohibited materials or processes, narrowing down the possibilities. It must be kept in mind that in some cases, these constraints may not be realistic or well advised. It is not unusual to encounter over-restrictive or seemingly arbitrary constraints. Constraints can also be indirectly applied. If an entire assembly has a certain constraint, such as total weight, it can become rather complicated to balance the necessary weights and properties of the individual components.

Cost Constraints. Financial constraints accompany each engineering design, except for unusual critical applications when properties are far more important than relative material expense. Design choices can be severely limited by economic factors, particularly in the manufacture of highly competitive consumer items. It is not unusual to be financially constrained to using essentially the same material as the competition, especially in a marketplace without real or perceived product differentiation.

Quantity Constraints. In the case of a single component or structure being designed, it will not be necessary to tool an assembly line or create a manufacturing process capable of making them by the millions. Production of a few components can sometimes be given more personal attention, permitting the inherent labor costs to be a higher percentage of the total price. On the other hand, if a large number of identical or similar items are to be produced, an assembly line approach is mandated, with an accompanying reduction in the relative labor costs. The in-between cases are the most typical, hence the rise of small, medium, and large job shops capable of sufficient flexibility to produce a variety of parts on a short-term contract basis.

Size and Weight Constraints. The rough size and weight of a finished design must be approximated early in the design process, because they may constrain the subsequent design options to a great extent. Maximum sizes and weights can restrict the amount of margin that is possible.

Material Property and Processing Constraints. Many property constraints are placed on materials by the very nature of the item being designed. Restrictions to manufacture can also be present as the willingness to use only those processes and fabrication techniques for which the equipment is already on hand. There are certainly financial advantages to be gained by maximizing utilization of existing facility and equipment capabilities. For example, if a manufacturer has a captive heat treatment department, they may be prone to exploring heat treatment as a preferred processing option. Complicated

and highly technical processing steps are usually best addressed by specialists. Many codes and specifications allow a broad range of materials selections, whereas other codes are very specific and allow few substitutions.

What It Must Be. After the component is envisioned, the environment is evaluated, and the constraints are applied, the design concept, including materials selection, can be developed. A large set of performance and property criteria may be developed in order to define the function and surrounding conditions of a part. Primary and secondary criteria should also be identified. Primary, or absolute, requirements are essential to proper service and cannot be subordinated, whereas secondary requirements are those where judicious compromises can be made. For example, life-cycle considerations (such as recycling or environmental impact) may be a primary or secondary criterion, depending on the product objectives.

Selection Criteria during Detailed Design

The stage of detailed design (Fig. 3) includes the embodiment or configuration level of design and parametric design. During configuration design, the emphasis is on determining the shape and approximate size of a part using engineering methods of analysis, which can be based on methods of qualitative physical reasoning. During parametric design, quantitative methods are used to refine the design further.

Materials selection during configuration design requires the evaluation of a range of material (e.g., a range of carbon steel, low-alloy steel, stainless steel, age-hardening aluminum alloys, etc.), its general product form (e.g., wrought, cast, powder metallurgy, etc.), and the processing method (e.g., forged, die cast, injection molded, etc.). All of these factors must be considered when the shape of a part is defined during configuration design. Material properties during configuration design must also be known to a greater level of precision than in conceptual design, at least to allow qualitative comparison of the alternatives for the possible choices of material type, product, and processing method.

At the detail or parametric design level, the materials selection is narrowed further to a specific grade of material and manufacturing processes. Here, the emphasis will be on quantitative evaluation of allowable variations in material properties, critical tolerances, and any other performance parameters of the design (including the best manufacturing process using quality engineering and cost-modeling methodologies). Depending on the criticality of the part, material properties may need to be known to a high level of precision, with quantitative evaluation of variations in properties or performance. For example, anisotropic variations in the properties of worked products, or the effects of surface finish after machining, are quantitative factors that must be considered during parametric design. At this extreme, the development of a detailed property database or an extensive materials-testing program may be required.

Detailed evaluation of the size, shape, processing, fabrication, and material properties of an engineered part requires communication between designers, materials or manufacturing engineers, quality assurance, and purchasing agents. It can be a relatively simple or complex task, depending on the criteria for materials selection. Examples of materials information required during detailed design are listed in Table 4 (Ref 9). It also includes experience and application history, such as failure analysis reports. During design, it is necessary to identify primary and secondary materials selection criteria, and the following list contains a number of typical materials selection criteria that would be identified during the creation of a new component (Ref 10):

- Size
- Shape
- Weight
- Strength
- Wear resistance
- Environmental resistance
- Loading capabilities
- Life expectancies
- Fabricability
- Quantity
- Availability
- Cost
- Specifications
- Recycling
- Scrap value
- Standardization
- Safety

Table 4 Examples of materials information required during detail design

Material identification
Material class (metal, plastic, ceramic composite)
Material subclass
Material industry designation
Material product form
Material condition designation (temper, heat treatment, etc.)
Material specification
Material alternative names
Material component designations (composite/assembly)
Material production history
Manufacturability strengths and limitations
Material composition(s)
Material condition (fabrication)
Material assembly technology
Constitutive equations relating to properties
Material properties and test procedures
Density
Specific heat
Coefficient of thermal expansion
Thermal conductivity
Tensile strength
Yield strength
Elongation
Reduction of area
Moduli of elasticity
Stress-strain curve or equation
Hardness
Fatigue strength (define test methods, load, and environment)
Temperature (cryogenic-elevated)
Tensile strength, yield strength
Creep rates, rupture life at elevated temperatures
Relaxation at elevated temperatures
Toughness
Damage tolerance (if applicable)
Fracture toughness (define test)
Fatigue crack growth rates (define environment, and load)
Temperature effects
Environmental stability
Compatibility data
General corrosion resistance
Stress-corrosion cracking resistance
Toxicity (at all stages of production and operation)
Recyclability/disposal
Material design properties
Tension
Compression
Shear
Bearing
Controlled strain fatigue life
Processability information
Finishing characteristics
Weldability/joining technologies
Suitability for forging, extrusion, and rolling
Formability (finished product)

Castability
Repairability
Flammability
Joining technology applicable
Fusion
Adhesive bonding
Fasteners
Welding parameters
Finishing technology applicable
Impregnation
Painting
Stability of color
Application history/experience
Successful uses
Unsuccessful uses
Applications to be avoided
Failure analysis reports
Maximum life service
Availability
Multisource? Vendors?
Sizes
Forms
Cost/cost factors
Raw material
Finished product or require added processing
Special finishing/protection
Special tooling/tooling costs
Quality control/assurance issues
Inspectability
Repair
Repeatability

Source: Ref 9

This list contains the most-used criteria for materials selection but is by no means exhaustive (as suggested by the examples in Table 4). Selection criteria can vary as much as the items being designed. These concerns also are not entirely independent, but they are described individually. It is logical to assume that complicated service requirements will result in more stringent selection criteria. More restrictive selection criteria will invariably result in fewer materials that will likely satisfy the design requirements. Even so, engineering materials selection is very rarely a question of a single, suitable material. As in engineering design, materials selection can be an iterative process that compares alternatives during both conceptual and detailed design.

Size Considerations. The size of a designed component can often dictate the form of the material to be used. Very large parts may need to be fabricated from structural shapes, castings, or forgings. Welded fabrication may be necessitated. Extremely small components may need to be created by powder metallurgy, metal injection molding, or other fine forming techniques.

Shape Considerations. The shape and geometrical complexity of a component must also be considered in materials selection. Intricate shapes may not fill during casting or may not be possible to form by other methods, such as extrusion or forging. Many types of material forms are available, such as castings, forgings, extrusions, rolled shapes, wire and rod, plate and sheet, and many hybrid forms. In considering shape during materials selection, it is always best to match the form to the function during configuration design. This practice reduces scrap and promotes the optimal use of material with the desired properties.

Several material forms may be determined to be suitable, and additional factors must be assessed to determine the optimal shape. In the most general sense, increasing complexity narrows the range of processes and increases cost. A cardinal rule of design is, therefore, to keep the shape as simple as possible. This rule may, however, be broken if a more complex shape allows consolidation of several parts and/or elimination of one or more manufacturing steps. Limitations on shape are also imposed by properties of the material and by interactions with the production tooling. For example, minimum wall or section thickness of the web form shown in Fig. 6 (Ref 11) is a function of manufacturing process and material. The aim is, generally, to produce a net shape part ready for assembly. If this is not feasible, a near-net shape part that will need only minor finishing, usually by machining, is desirable.

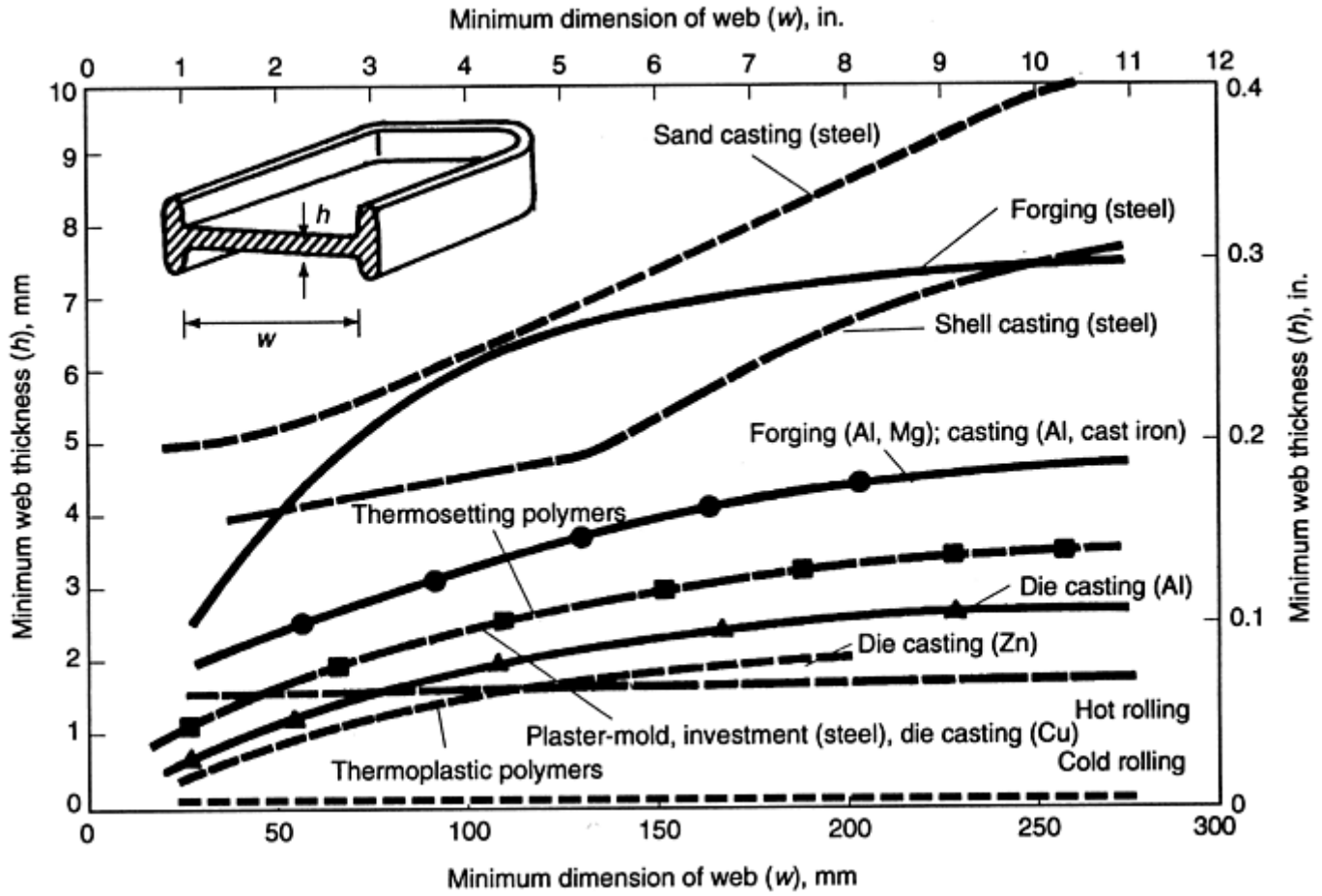


Fig. 6 Example of minimum web thickness for different materials and manufacturing processes. Source: Ref 11

Weight Considerations. There are very few applications of manufactured products where weight is not a consideration. Material weight is of vital importance in aerospace and automotive applications, where incremental weight reduction can be directly measurable in fuel savings and payload capacity. Strength-to-weight ratio, often called specific strength, is a hybrid consideration typically used in vehicle design. Aside from vehicles, weight considerations are also important if the materials are to be transported during manufacture or during service, or if the part moves during service, as in rotating or reciprocating parts of an engine or machinery.

Material Properties. A list of many of the properties to be considered during materials selection is shown in Table 5. Each primary or secondary property attribute must be carefully considered. Probably the most fundamental tenet of materials science is that properties are a function of structure and structure is a function of processing. The properties can never be considered separately from the processing, because processing decisions or steps can affect the nominal value and variability of a property within the geometry of a part. For example, austenitic stainless steel bar can be processed to high strength by drawing but can also be annealed to lower strength and improve ductility. Another example is the variability of properties within the geometry of a part (i.e., anisotropy), depending on the nature of the manufacturing process.

Table 5 Typical material properties used for selection

Tensile strength
Yield strength
Elongation
Compressive strength
Shear strength
Fatigue strength
Fracture toughness
Impact strength
Transition temperature
Modulus of elasticity

Wear resistance
Hardness
Lubricity
Density
Porosity
Melting point
Thermal stability
Thermal expansion
Thermal conductivity
Electrical conductivity
Magnetic characteristics
Galvanic character
Corrosion resistance
Optical characteristics
Fabrication characteristics
Welding characteristics
Finishing characteristics
Hardenability
Aesthetics

It is also important to understand relationships between the mechanical and physical properties. Some of the typical material property relationships are illustrated in Fig. 7 (Ref 12). This diagram shows the general inverse relationships between desirable and possibly undesirable characteristics. Each desired property will likely have attendant properties that may not be desirable. Therefore, materials selection will always contain a level of educated compromise, similar to the design process in its entirety.

Fig. 7 General relationships of different mechanical behavior. Rigidity and strength are generally inversely related to flexibility and ductility. Source: Ref 12

Wear Resistance is a very important property for all materials that come into repeated or non-stationary contact with other materials. The three principal types of wear are adhesive wear, abrasive wear, and corrosive wear. Unfortunately, the measure of wear resistance is problematic, because there are so many variables involved, including friction factors, lubrication factors, surface finish, and so on. As a result, tables of comparative wear resistance, even ostensibly employing the same standard testing methodology, have limited usefulness in materials selection. Optimal wear properties are usually obtained by prototype testing in an actual intended-service situation.

Knowledge of Operating Environment. This portion of the materials selection differs slightly from the original environmental constraints placed on the design during the conceptual stage. The environmental resistances and behaviors of individual materials are of interest, although very often the conceptual design requirements are the same as those for the configuration of individual parts. However, during configuration design, the surrounding environment may involve galvanic differences and potential deleterious interactions within a complex assembly or structure. Parametric design must account for corrosion rates, material replacement rates, and other life factors.

Types of Loading. The type and magnitude of applied loading are crucial aspects of materials selection. While considering potential materials with other requisite properties, the following general loading types must be addressed:

- Constant, sustained loading
- Cyclic, repetitive loading
- Rapid, shock loading
- Slow loading
- Distributed loading
- Concentrated loading
- Variable loading

The load-carrying attributes of a material can be altered substantially through different types of processing. Any processes that alter mechanical properties, such as heat treatment, rolling, welding, and grinding, can affect the load-bearing characteristics by enhancing or reducing the resistance to specific loading types. Combinations of loading types make the materials selection process more difficult.

Life Requirements. The life requirements influence the materials selection, because longer service duration can often necessitate more sophisticated materials. Some components and assemblies are single use and do not require prolonged capabilities. Many structures are intended to survive fifty or a hundred years before demolition and replacement. Short-term design can often be of a disposable nature, whereas long-term design may permit substantial repair or refurbishment during service.

Fabricability (Design for Manufacturing). Designing for effective and efficient manufacturing can be rather involved but is important in cost-effective designs. Fabrication and manufacturing characteristics may be difficult to quantify, because it may be a composite of many subjective measures, such as formability, machinability, and weldability. Every process that is applied to a material must be evaluated to determine if the process and material are compatible. The first step may be a qualitative comparison during conceptual or configuration design, based on compatibility charts such as the one in Table 6. The necessary processing for a material also might be so cumbersome and costly, or impossible, that an otherwise suitable material would be logically removed from consideration. Materials selections may impose additional inspection, heat treatment, welding, machining, and finishing requirements during manufacture.

Table 6 Compatibility between materials and manufacturing processes

Process	Cast iron	Carbon steel	Alloy steel	Stainless steel	Aluminum and aluminum alloys	Copper and copper alloys	Zinc and zinc alloys	Magnesium and magnesium alloys	Titanium and titanium alloys	Nickel and nickel alloys	Refractory metals	Thermoplastics	Thermoset plastics
Casting/molding													
Sand casting	•	•	•	•	•	•	—	•	—	•	—	X	X
Investment casting	—	•	•	•	•	•	—	—	—	•	—	X	X
Die casting	X	X	X	X	•	—	•	•	X	X	X	X	X
Injection molding	X	X	X	X	X	X	X	X	X	X	X	•	—
Structural foam molding	X	X	X	X	X	X	X	X	X	X	X	•	X
Blow molding (extrusion)	X	X	X	X	X	X	X	X	X	X	X	•	X
Blow molding (injection)	X	X	X	X	X	X	X	X	X	X	X	•	X
Rotational molding	X	X	X	X	X	X	X	X	X	X	X	•	X
Forging/bulk forming													
Impact extrusion	X	•	•	—	•	•	•	—	X	X	X	X	X
Cold heading	X	•	•	•	•	•	—	—	X	—	X	X	X
Closed die forging	X	•	•	•	•	•	X	•	•	—	—	X	X
Pressing and sintering (P/M)	X	•	•	•	•	•	X	•	—	•	•	X	X
Hot extrusion	X	•	—	—	•	•	X	•	—	—	—	X	X
Rotary swaging	X	•	•	•	•	—	—	•	X	•	•	X	X
Machining													
Machining from stock	•	•	•	•	•	•	•	•	—	—	—	—	—
Electrochemical machining	•	•	•	•	—	—	—	—	•	•	—	X	X
Electrical discharge machining (EDM)	X	•	•	•	•	•	—	—	—	•	—	X	X

Wire EDM	X	•	•	•	•	•	—	—	—	•	—	•	X
Forming													
Sheet metal forming	X	•	•	•	•	•	—	—	—	—	X	X	X
Thermoforming	X	X	X	X	X	X	X	X	X	X	X	•	X
Metal spinning	X	•	—	•	•	•	•	—	—	—	—	X	X

•, normal practice; —, less-common practice; X, not applicable; P/M, powder metallurgy. Source: Ref 9

During parametric design, quantitative evaluation of tolerances, tooling, and production costs would be required. It is often necessary to design the manufacturing tooling concurrently with the end product to be made on that tooling. This is especially true for near-net shape processes such as molding, casting, and forging. Sometimes, the designed components must be altered to permit manufacture; hence, the manufacturing functions need to be involved in the design from the beginning.

Quantity Requirements. The quantity of a component to be designed may also influence the material and processing options that are feasible. The manufacture of high volumes of parts may necessitate mass-production methodologies. High volumes may allow use of forming and production techniques that require expensive tooling and dies that would be financially unfeasible when only a few parts are to be produced. These high-production methods can be very cost-intensive, inflexible, and slow to become profitable. Lower quantities of components can often allow more individual attention to the quality and characteristics of each produced part. Many production and processing methods are not applicable to low quantities of parts.

Availability. Materials, as a result of their popularity and relative natural scarcity, may not always be available as production may require. Even abundant materials in unusual forms can become difficult to procure. Design of long-term projects or continuous production includes an implied assumption that the selected material will remain an obtainable and economical choice in the future.

Lower quantities of material types, forms, and shapes can be below that level which mills will supply directly, and these would need to be purchased from a service center or distributor. The uniqueness of the material may be problematic, because small buyers cannot singly affect what mills will produce. Reduced demand from other manufacturers may make desired materials no longer available.

Both raw material and alloying elements are not uniformly available. Foreign sources may be hostile or inconsistent, and general availability may severely restrict supply. Special consideration must be given to using any base materials or processing materials that may not have the requisite availability due to factors that cannot be controlled. These materials are often called strategic materials, and they can become a great concern during wartime, when necessary materials may become of short supply.

Cost. Determination of costs accompanying potential materials selections is not as straightforward as it may seem. In many applications, the material cost was traditionally dictated by the cheapest material that was available that satisfied the previously determined mechanical and physical property requirements. Modern design practices incorporate material and processing costs almost as a property of that material, to be a direct comparison factor. Artificial constraints to using only the subjectively least-expensive material available ignores additional potential benefits of more expensive materials, such as reduced maintenance, longer life, and better reliability. Often a value-in-use approach is employed to better evaluate the costs of potential materials. In this methodology, the additional benefits of better performance can be quantitatively considered along with the basic material cost. In some instances, the additional processing costs for cheaper materials may result in greater total expense than those materials that are traditionally more expensive.

The amount of total component or structure cost that is included in material and subsequent processing varies widely in different industries. Large structures use great volumes of typically lower-cost engineering materials, resulting in the material being a relatively low percentage of the overall project costs. Aerospace and electronic components are typically smaller, specialized items, where the material costs can be higher than the processing and installation costs. In extreme cases, the cost is no object, within reason. The designer must be certain that the accompanying production costs, maintenance costs, potential repair, and downtime costs are considered for prospective materials.

The costs of ordering and warehousing engineering materials are also a consideration. Depending on the amount of material needed, there is likely an economic order quantity that best suits the production requirements and minimizes material costs. This is also a dynamic function, because materials and material forms have unpredictably mutable costs and availability that can alter future purchasing requirements. In general, all costs regarding purchasing, receiving inspection, and storage are reduced, on a per pound basis, by the purchase of large amounts.

Existing Specifications and Codes. In many industries, applicable standards provide materials prohibitions or requirements above those applicable as original design constraints. Standards can restrict material form, heat treatment, welding, and other processing variables. Purchasing material grades and alloys to uniform, popular standards can result in greater availability of materials, due to a greater number of potential suppliers.

Standards can be industry consensus standards, domestic and foreign federal regulations, and customer-supplied engineering specifications. Reliance on these codified requirements is often precarious, because they are sometimes vague and can be interpreted in many ways. Many specifications still require producer and client agreement on crucial processing variables. Some statutes (e.g., the Federal Child Safety Act) may also mandate possible materials and a given design procedure.

Feasibility of Recycling. The potential recycling of manufacturing and process scrap can be an important selection parameter. The expense of a material may be easier to justify if all removed material may be recycled for remuneration via return to the supplier rather than requiring landfill or hazardous material disposal expenditures. Identification of recyclability or other end-of-life considerations as a primary or secondary selection attribute is well advised, even if the

remainder of the design is not part of a life-cycle design effort. This ability considers the ease of sorting and separation and the fluctuating cost of recycled materials compared to newly extracted materials.

Scrap Value. The consideration of whether a designed component may be profitably scrapped at the end of its useful life is an important part of materials selection. Reuse of nondegraded components or rework and refurbishment is often an inexpensive alternative to new purchases.

Standardization of Designs. Standardization of materials selection within organizations that have extensive and continuous design functions must be considered. It is possible to design identical components that will fit in multiple assemblies. This practice of employing analogous designs and materials can prevent costly, redundant design projects or "reinvention of the wheel."

Safety. Perhaps the most important factor in design of a structure or component is safety. A design successfully reaching the manufacturing stage is entirely dependent on critical review and scrutiny about whether the necessary safety factors are satisfied.

Safety also must be included as a selection consideration during manufacture and processing. Materials and processes exhibit potential safety concerns, such as toxicity, flammability, inhalation of fine particles, autoignition, and contact hazards as well as long-term effects such as carcinogenic and pathogenic characteristics. Manufacturers are compelled to make products safer, due to the ethical imperative and economic self-interest.

Regulatory bodies are continually assessing the potential health hazards of relatively newly developed materials. This is due to the greater number of lesser-known and more rare metals and nonmetals that are being incorporated into engineering design. The Occupational Safety and Health Administration (OSHA) and other federal regulatory bodies are constantly collecting greater knowledge of potential health concerns regarding various materials, during both their manufacture and service. Avoiding the use of materials or processes under scrutiny for health concerns would generally be prudent.

References cited in this section

8. R.W. Heckel, Introduction to the Effects of Composition, Processing, and Structure on Materials Properties, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 333

9. G. Dieter, Overview of the Materials Selection Process, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 243–254

10. H. Boyer, Introduction: Selection Criteria, *Selection of Materials for Component Design,* ASM International, 1986

11. J.A. Schey, *Introduction to Manufacturing Processes,* 2nd ed., McGraw-Hill, 1987

12. R. Gunia, Ed., *Source Book on Materials Selection,* Vol 2, American Society for Metals, 1977

Materials Selection for Failure Prevention

Brett A. Miller, Stork Technimet, Inc.

Materials Selection for Failure Prevention

The use of inappropriate materials and processes accounts for a significant number of failed parts. Table 7, for example, itemizes the general causes of failure, with the frequency of occurrence determined from a survey (Ref 13). In this survey, materials selection is the most frequent cause of failure for engineered components. In the case of aircraft components, however, the survey did not identify any failures caused by improper materials. This difference illustrates the important point of how different design methods may influence the process of materials selection. In aerospace, for example, design methods should involve more critical evaluations of material alternatives, because the hazards of failure can be severe.

Table 7 Frequency of causes for failure

Cause	Percentage of failures	
	Engineering components	Aircraft components
Improper materials selection	38	…

Fabrication imperfections	15	17
Faulty heat treatment	15	…
Design errors	11	16
Unanticipated service conditions	8	10
Uncontrolled environmental conditions	6	…
Inadequate inspection/quality control	5	…
Material mix	2	…
Inadequate maintenance	…	44
Defective material	…	7
Unknown	…	6

Source: Ref 13

The selection of materials to prevent failure is typically a structured approach including thorough and diligent research into suitable materials. There is no universal guide that will automatically identify the best material for any service, because the number of interrelated input variables can be difficult to manage by a formal decision structure. Executive decision trees and computer expert systems have been developed to simplify materials selection, and these systems can identify candidate materials from very large databases of engineering materials, with cross-referenced mechanical and physical properties. However, these systems are narrowly applicable and are usually industry or company specific.

Effective materials selection is aided by access to materials property information and acquired engineering knowledge of all engineers participating on the design. No generalizations can be made that will be valid for all materials-selection problems, because each design problem must be considered individually or on the basis of closely related experience. Table 8 (Ref 14), however, provides some general guidance to the criteria that are typically significant in selecting materials in relation to possible failure mechanisms, types of stress, and operating temperatures.

Table 8 Guide to criteria generally useful for selection of material in relation to possible failure mechanisms, types of loading, types of stress, and intended operating temperatures

Failure mechanisms	Types of loading			Types of stress			Operating temperatures			Criteria generally useful for selection of material
	Static	Repeated	Impact	Tension	Compression	Shear	Low	Room	High	
Brittle fracture	X	X	X	X	…	…	X	X	…	Charpy V-notch transition temperature. Notch toughness. K_{Ic} toughness measurements
Ductile fracture[(a)]	X	…	…	X	…	X	…	X	X	Tensile strength. Shearing yield strength
High-cycle fatigue[(b)]	…	X	…	X	…	X	X	X	X	Fatigue strength for expected life, with typical stress raisers present
Low-cycle fatigue	…	X	…	X	…	X	X	X	X	Static ductility available and the peak cyclic plastic strain expected at stress raisers during prescribed life
Corrosion fatigue	…	X	…	X	…	X	…	X	X	Corrosion-fatigue strength

Failure mechanisms	Types of loading			Types of stress			Operating temperatures			Criteria generally useful for selection of material
	Static	Repeated	Impact	Tension	Compression	Shear	Low	Room	High	
										for the metal and contaminant and for similar time[(c)]
Buckling	X	…	X	…	X	…	X	X	X	Modulus of elasticity and compressive yield strength
Gross yielding[(a)]	X	…	…	X	X	X	X	X	X	Yield strength
Creep	X	…	…	X	X	X	…	…	X	Creep rate or sustained stress-rupture strength for the temperature and expected life[(c)]
Caustic or hydrogen embrittlement	X	…	…	X	…	…	…	X	X	Stability under simultaneous stress and hydrogen or other chemical environment[(c)]
Stress-corrosion cracking	X	…	…	X	…	X	…	X	X	Residual or imposed stress and corrosion resistance to the environment. K_{ISCC} measurements[(c)]

K_{Ic}, plane-strain fracture toughness; K_{ISCC}, threshold stress intensity to produce stress-corrosion cracking.
(a) Applies to ductile metals only.
(b) Millions of cycles.
(c) Items strongly dependent on elapsed time.
Source: Ref 14

Perhaps one of the most troublesome areas of materials selection relates to the change (or variation) in properties and performance. Property variations can occur within the part geometry from processing and fabrication, or changes in properties can occur over time from factors such as:

- Wear
- Temperature extremes or changes
- Corrosion
- Fatigue

These application conditions require a great deal of judgment in interpreting laboratory test data into design and extrapolating properties and performance over extended periods of time. Often, simulated service testing may be required. An important role of the materials engineer is to assist the designer in making meaningful connections between materials properties and the performance of the part or system being designed. For most mechanical systems, performance is limited not by a single property but by a combination of them. For example, the materials with the best thermal shock resistance are those with the largest values of $\sigma_f/E\alpha$, where σ_f is the failure stress, E is Young's modulus, and α is the thermal coefficient of expansion. These types of performance indices (i.e., groupings of material properties that, when maximized, maximize some aspect of performance) can be useful to compare materials.

Understanding the connection between properties and the failure modes is also important. Figure 8 is a chart of relationships between common failure modes and material properties (Ref 15). For most modes of failure, two or more material properties act to control the material behavior. However, it is also important to understand how property data should be interpreted. For example, even though most standard specifications require tensile-test data, these data are only partially indicative of mechanical performance in specific conditions. The purpose of tensile testing is often to monitor relative quality of different lots, not necessarily for design. Moreover, except in those conditions where ductile fracture or gross yielding may be the limiting condition for failure (Fig. 8), tensile strength and yield strength may be inadequate criteria for avoiding failure. A high tensile strength, for example, might be indicative of lower ductility and toughness, and thus a part with severe stress raisers might be prone to failure.

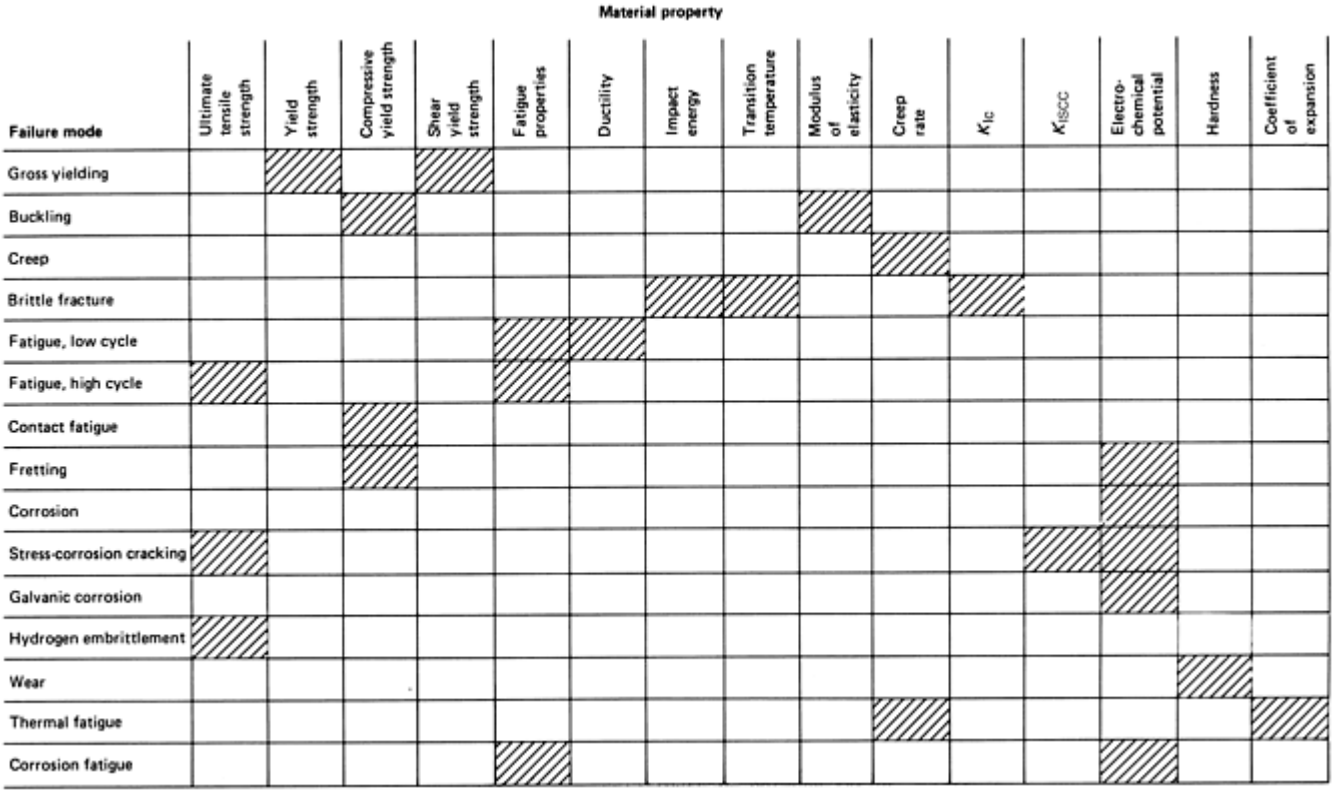


Fig. 8 General relationships between failure modes and material properties. Shaded blocks indicate properties that are influential in controlling a particular failure mode. K_{Ic}, plane-strain fracture toughness; K_{ISCC}, threshold stress intensity for stress-corrosion cracking. Source: Ref 15

References cited in this section

13. G.J. Davies, *Performance in Service, Essential Metallurgy for Engineers,* E.J. Bradbury, Ed., Van Nostrand Reinhold, London, 1985, p 126–155

14. Fundamental Sources of Failure, *Failure Analysis and Prevention,* Vol 10, *Metals Handbook,* 8th ed., American Society for Metals, 1975, p 4

15. C.O. Smith and B.E. Boardman, Concepts and Criteria in Materials Engineering, *Properties and Selection: Stainless Steels, Tool Materials, and Special Purpose Metals,* Vol 3, *Metals Handbook,* 9th ed., American Society for Metals, 1980, p 825–834

Brett A. Miller, Stork Technimet, Inc.

Materials Selection and Failure Analysis

Case histories of failure investigations provide an indispensable tool not only for design but also in the education and training of engineers. Even historical case studies for obsolete materials and technology can still offer insight in identifying root causes and preventing failures in new designs. The investigative process can also identify secondary contributory causes so that they may be accounted for in a preemptive manner rather than by repetitive trial and error. In this way, case studies can be very important to the overall design process and materials selection.

Conversely, the materials selection process is of great importance to the failure analyst. A working knowledge of materials selection is a prerequisite for all engineers performing failure analysis. Throughout the failure analysis process, the investigator needs to consider the appropriateness of the selected material and processing. Inadvertent material substitutions and processing mistakes are often encountered. At other times, the physical evidence suggests that slightly unusual or wholly inappropriate materials and processes were employed. Scientific failure analysis may reveal that the original design was inadequate or had not considered all of the relevant service factors.

The composition and grade identification are typically ascertained during a failure investigation. The results are compared to the specified material or compared to standard grades in cases where no specifications were provided. Subtle deviation from the required composition is not necessarily the cause of a failure, as is sometimes erroneously assumed. The investigator should also endeavor to determine the likely processing methods used on the part or structure. This will be an aid, because certain manufacturing methods exhibit characteristic propensities for certain flaw and defect types. Processes such as heat treatment, welding, and machining can be addressed on a postmortem basis by evaluation of the strength and microstructure. Standard materials laboratory tests and scanning electron microscope fractography provide the remaining observations and data necessary to identify the failure mode and causative factors. The material and processing history thereby gleaned can be compared to the known service conditions to deduce the design concept. This reasoning can then be used to improve the design or address other materials or processing inadequacies.

Similar to design, failure analysis is somewhat influenced by the technical background and experience of the analyst. The basic weaknesses of single-discipline failure analysis can be analogous to design. Different engineering disciplines tend to approach failure analyses primarily within their area of specialization. A failure analysis team approach will likely provide the most beneficial corrective actions.

Some Questions the Failure Analyst Should Ask

A large variety of questions would be asked and answered during the course of a failure investigation. Education and experience will indicate to the analyst what questions should necessarily be addressed. The compound questions listed subsequently are among the many an analyst would address during an investigation:

- Should the complex part be an assembly of several parts rather than one?
- How was the component loaded, and was anisotropy considered?
- Is the material capable of being produced with the required properties, in the form used?
- Can any available material meet the specifications?
- Did the strength requirements preclude toughness or corrosion-resistance needs?
- Was the wear resistance adequate for the materials in contact?
- Were the desired properties compromised by the use of low-cost materials or processes?
- Did the materials and processing comply with the applicable codes and standards?
- Was the product made with unique materials and processes?
- Were proprietary or obsolete materials and processes employed?
- Were the manufacturing processes used to create the desired shape appropriate?
- Did the individual processing methods make sense?
- Should it have been preheated prior to heat treatment or welding?
- Did the fabricability requirements compromise the desired mechanical or physical properties?
- Were the manufacturing methods appropriate for the quantity produced?
- Were the operating conditions and maintenance as intended?
- Were the service conditions easy to anticipate?
- Does the material possess adequate durability in the service environment?
- How did the scrap value contribute to repair and maintenance decisions in service?

Examples of Improper Materials Selection

The consequences of improper materials selection can range from simply aggravating to catastrophic. The causes for failures due to materials and processing are many and varied, because the design process involves the balancing of part

function with manufacturing, cost, and service conditions. Failure analysis will typically indicate whether a material was suitable, marginally unsuitable, or drastically incompatible. The following examples describe some failures that suggest a questionable choice of material.

Example 1: Failure of a Steel Lifting Eye. A steel lifting eye that had fractured during service is shown in Fig. 9. No additional service-related information was provided. The eye was reportedly manufactured from a grade 1144 steel and should exhibit a minimum tensile strength of 689 MPa (100 ksi).

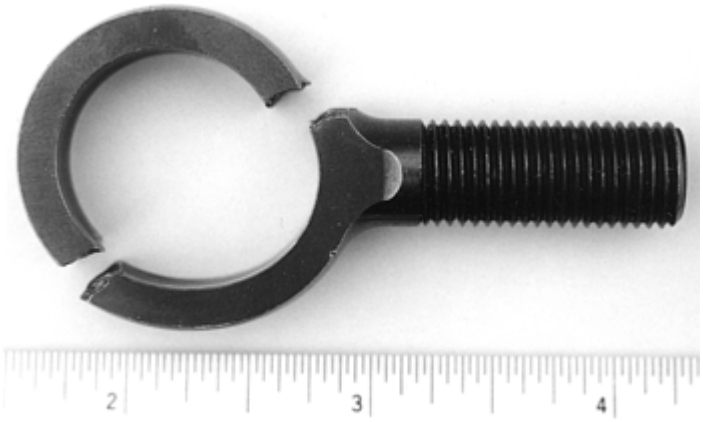


Fig. 9 Steel eye that had fractured in two locations during service

The eye was approximately 70 mm (2.75 in.) long and appeared to be machined. Fracture occurred in two locations: adjacent to the threaded shank and diametrically opposite to this region. The circular eye was deformed longitudinally, and the fracture surfaces exhibited an angular orientation.

Chemical analysis confirmed that the eye was similar to a resulfurized and rephosphorized grade 1144 steel. The sulfur content was slightly below the normal limits, and the phosphorus content was slightly above the typical range.

Scanning electron microscope (SEM) examination of the fracture surfaces revealed woody overload features, typical for resulfurized steels. The morphology was identified as a mixed fracture mode of cleavage and ductile rupture, and the directionality of the features was suggestive of shear overload. A typical region of the fracture nearest the shank, which was the likely origin, is shown in Fig. 10. Fracture preferentially followed the nonmetallic inclusions. Tensile testing could not be performed on the eye, but the hardness was found to be 32 HRC. This is roughly equivalent to 1,000 MPa (145 ksi), which exceeded the drawing specification.

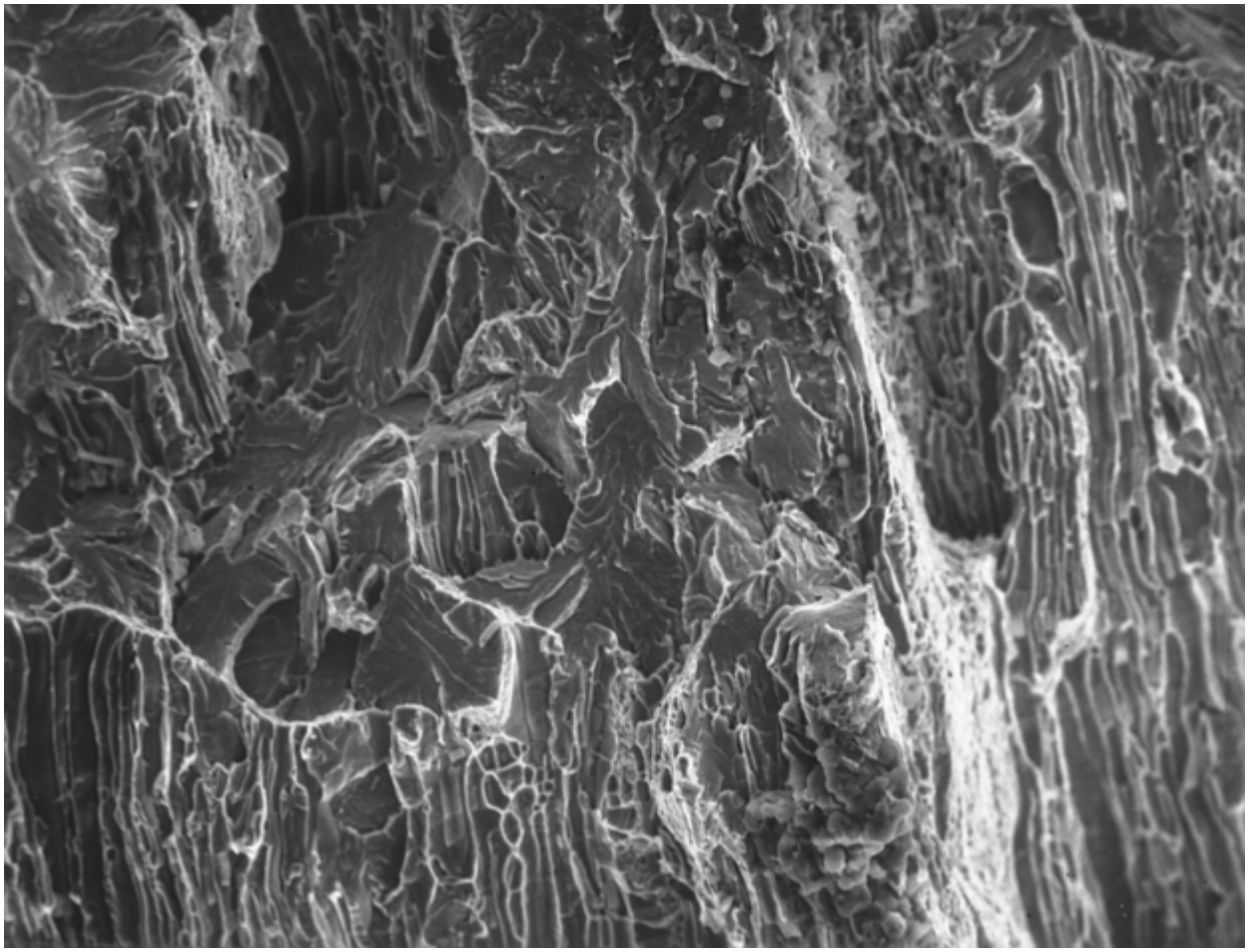


Fig. 10 Scanning electron microscope micrograph of typical eye fracture morphology consisting of woody, ductile features. 500×

Metallographic examination was performed through the fracture surfaces, and the fracture profile of the fracture surface near the shank is shown in Fig. 11 The fracture was parallel to the direction of the manganese sulfide stringer inclusions. Etching revealed the presence of significant banding of the ferrite and pearlite microstructure. The fracture is primarily along the inclusions and through bands of ferrite, as shown in Fig. 12

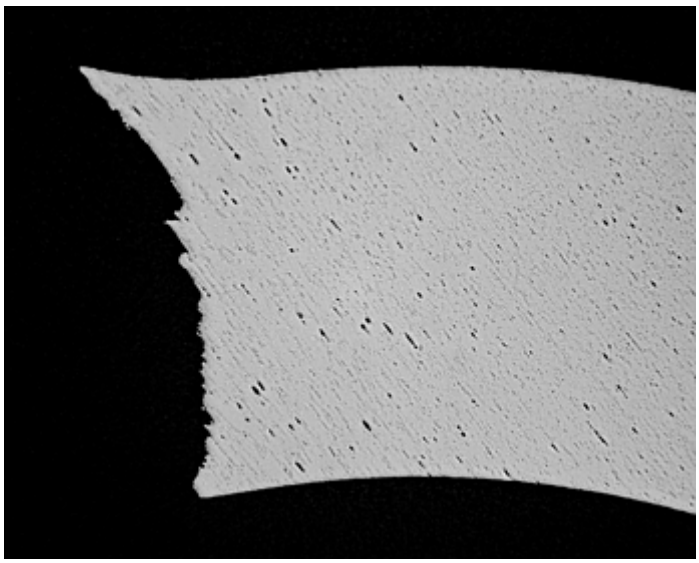


Fig. 11 Cross section through the eye showing cracking through the aligned stringer inclusions. Unetched. 2×

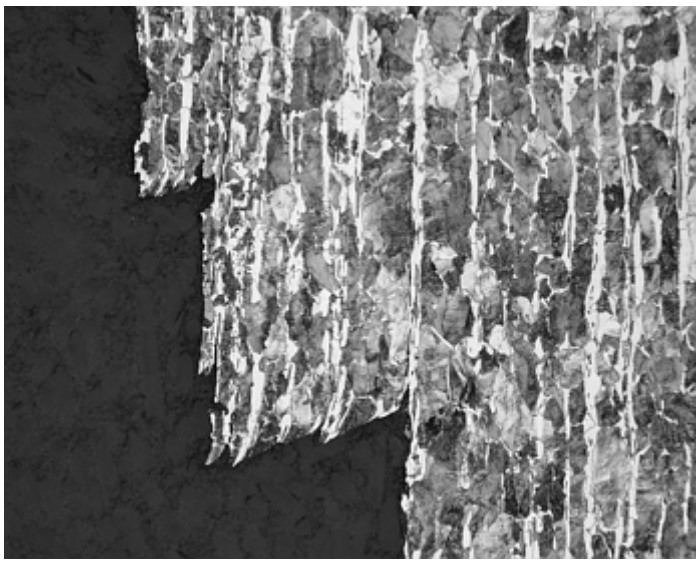


Fig. 12 High-magnification view of the eye fracture surface showing fracture through the sulfide inclusions and a banded microstructure. 2% nital etchant. 28×

It was concluded that the lifting eye failed as a result of overload. Fracture occurred parallel to the rolling direction, through manganese sulfide stringers and ferrite bands in the base metal matrix. The eye was machined from grade 1144 steel that was heavily cold rolled for strength. This material is very anisotropic, exhibiting substantially poorer long and short transverse mechanical properties than the longitudinal properties, which were likely used for design. It is likely that the materials selection process did not properly account for this anisotropy. The selection of a rolled product may also be questionable here. It may be better to use a forged product in this case because of resulting “grain flow” and inclusion orientation.

Example 2: Failure of a Tank Coupling. Leakage was identified around a coupling welded into a stainless steel holding tank. The tank had been in service for several years, storing condensate water with low impurity content. The tank and fitting were manufactured from type 304 stainless steel. The fitting was fillet welded to the tank wall, and the tank was covered with insulation in service.

A diagram of the failed tank section is shown in Fig. 13. The coupling joint consisted of an internal groove weld and an external fillet weld. Cracking was apparent on the tank surface, adjacent to the coupling weld. Some reddish rust was present on the surface, but no gross mechanical damage, yielding, or weld defects were evident.

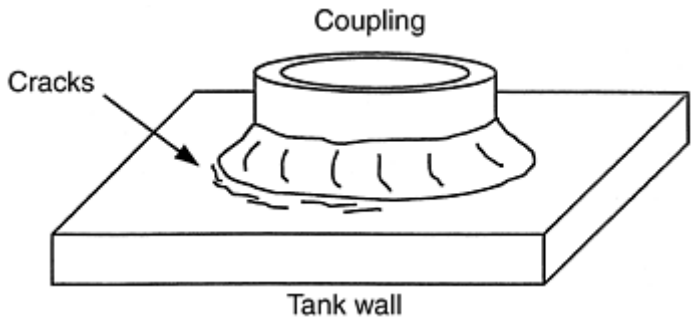


Fig. 13 Diagram of a tank coupling region that leaked during service

The chemical composition of the plate was consistent with a type 304 austenitic stainless steel. No compositional anomalies were detected. Energy-dispersive x-ray spectrometric analysis of the corrosion product on the crack surfaces revealed chlorine, carbon, and oxygen in addition to the base metal elements. The amount of corrosion present at the primary crack prevented SEM examination for morphological identification. Hardness testing of the plate revealed a hardness level considered typical for annealed stainless steel plate.

A metallographic cross section through the most severe cracking is shown in Fig. 14. A great number of secondary, branching cracks are evident in the weld, heat-affected zone (HAZ), and base metal. A typical crack is shown at higher magnification in Fig. 15. Branching, transgranular cracking is evident, emanating primarily from the exterior of the tank. Examination of the HAZ microstructure did not reveal evidence of substantial sensitization.

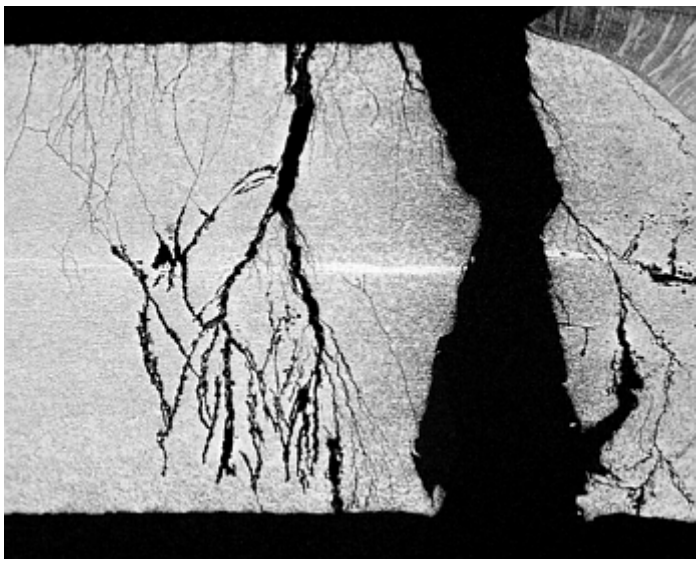


Fig. 14 Metallographic cross section through the cracked region of the coupling, showing branching cracks from the exterior (top). 10% oxalic acid, electrolytic etch. 1.75×

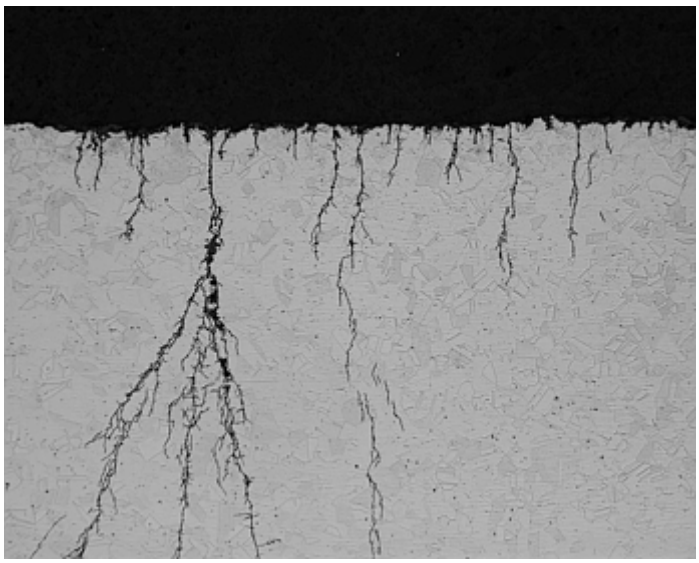


Fig. 15 High-magnification view at the exterior tank surface showing branching, transgranular stress-corrosion cracking. 10% oxalic acid, electrolytic etch. 14×

The analytical investigation concluded that the tank failed as a result of stress-corrosion cracking (SCC) that initiated at the exterior surface. Contaminant material containing chlorine, which may have leached from the insulation, combined with the inherent susceptibility of the base material and residual stresses from fabrication and welding. These factors combined synergistically to result in cracking. Aqueous chlorides, especially within an acidic environment, have been shown to cause SCC in austenitic stainless steels under tensile stress. The use of a type 304 stainless steel to prevent internal corrosion damage did not adequately foresee the potential for corrosion damage from external contamination.

Example 3: Localized Corrosion of Type 303 Stainless Steel Exposed to Acidic Soft Drinks (Ref 16). This example from *Metals Handbook,* 8th edition, (Ref 16) illustrates how a secondary selection factor (machinability) was not adequately evaluated in terms of a particular service environment and function. In this case, the failure is related to the selection of type 303 stainless steels for ease of machining instead of type 304 stainless steel.

After about two years in service, a valve in contact with soft drink in a vending machine occasionally dispensed discolored drink with a sulfide odor. The soft drink in question was one of the more strongly acidic, containing citric and phosphoric acids with a pH of 2.4 to 2.5, according to the laboratory at the bottling plant.

Manufacturing specifications for the valve called for type 303 stainless steel, a free-machining grade chosen because of the substantial amount of machining required for the part. Other parts in contact with the drink were made from type 304 stainless or inert plastics. In this application, type 303 stainless steel had only marginal corrosion resistance because of the size and distribution of sulfide stringers in some lots. When the machine was unused overnight or over the weekend, there

was occasionally enough attack on exposed sulfide stringers to make the adjacent liquid unpalatable. Specification of type 304, which is suitable for this application, was thus recommended.

Reference cited in this section

16. *Failure Analysis and Prevention,* Vol 10, *Metals Handbook,* 8th ed., 1975, p 179

Materials Selection for Failure Prevention

Brett A. Miller, Stork Technimet, Inc.

References

1. H.W. Stoll, *Product Design Methods and Practices,* Marcel Dekker, 1999, p 40, 148

2. J.R. Dixon and C. Poli, *Engineering Design and Design for Manufacturing,* Field Stone Publishers, 1995

3. C.O. Smith, Human Factors in Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997

4. J.R. Dixon, Conceptual and Configuration Design of Parts, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 33–38

5. G. Boothroyd and P. Dewhurst, *Product Design for Assembly,* Boothroyd Dewhurst, Inc., 1989

6. S. Pugh, *Total Design: Integrating Methods for Successful Product Engineering,* Addison-Wesley, 1991

7. G. Kardos, Risk and Hazard Analysis in Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997

8. R.W. Heckel, Introduction to the Effects of Composition, Processing, and Structure on Materials Properties, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 333

9. G. Dieter, Overview of the Materials Selection Process, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 243–254

10. H. Boyer, Introduction: Selection Criteria, *Selection of Materials for Component Design,* ASM International, 1986

11. J.A. Schey, *Introduction to Manufacturing Processes,* 2nd ed., McGraw-Hill, 1987

12. R. Gunia, Ed., *Source Book on Materials Selection,* Vol 2, American Society for Metals, 1977

13. G.J. Davies, *Performance in Service, Essential Metallurgy for Engineers,* E.J. Bradbury, Ed., Van Nostrand Reinhold, London, 1985, p 126–155

14. Fundamental Sources of Failure, *Failure Analysis and Prevention,* Vol 10, *Metals Handbook,* 8th ed., American Society for Metals, 1975, p 4

15. C.O. Smith and B.E. Boardman, Concepts and Criteria in Materials Engineering, *Properties and Selection: Stainless Steels, Tool Materials, and Special Purpose Metals,* Vol 3, *Metals Handbook,* 9th ed., American Society for Metals, 1980, p 825–834

16. *Failure Analysis and Prevention,* Vol 10, *Metals Handbook,* 8th ed., 1975, p 179

Brett A. Miller, Stork Technimet, Inc.

Selected References

- D. Askeland, *The Science and Engineering of Materials,* 2nd ed., PWS-KENT Publishing Company, Boston, 1989
- ASM Committee on Material Requirements for Service Conditions, Material Requirements for Service Conditions, *Welding, Brazing, and Soldering,* Vol 6, *ASM Handbook,* ASM International, 1993, p 373–375
- B.P. Bardes and L.J. Korb, Guidelines for Selection of Material, *Properties and Selection: Stainless Steels, Tool Materials, Special Purpose Metals* Vol 3, *ASM Handbook* 9th ed., American Society for Metals, 1980
- J.A. Charles and F.A.A. Crane, *Selection and Use of Engineering Materials,* 2nd ed., Butterworths, London, 1989
- *Failure Analysis and Prevention,* Vol 11, *Metals Handbook,* 9th ed., American Society for Metals, 1986
- M.M. Farag, *Materials and Process Selection in Engineering,* Applied Science Publishers, Ltd., London, 1979
- R. Gunia, Ed., *Source Book on Materials Selection,* Vol 1, American Society for Metals, 1977
- R. Gunia, Ed., *Source Book on Materials Selection,* Vol 2, American Society for Metals, 1977
- R. Kern and M. Suess, *Steel Selection—A Guide for Improving Performance and Profits,* John Wiley and Sons, New York, 1979
- C. MacDermott and A. Shenoy, *Selecting Thermoplastics for Engineering Applications,* 2nd ed., Marcel Dekker, Inc., New York, 1997
- *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997
- F.K. Naumann, *Failure Analysis, Case Histories and Methodology,* Dr. Riederer-Verlag GmbH, Stuttgart, Germany, 1983
- J.B. Park and R.S. Lakes, *Biomaterials: An Introduction,* 2nd ed., Plenum Press, New York, 1992
- H. Petroski, *To Engineer is Human,* Vintage Books/Random House, Inc., New York, 1985
- C.O. Smith, and B.E. Boardman, Concepts and Criteria in Materials Engineering, *Properties and Selection: Stainless Steels, Tool Materials, Special Purpose Metals,* Vol 3, *Metals Handbook,* 9th ed., American Society for Metals, 1980
- G. Vander Voort, Use of Failure Analysis in Materials Selection, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 322–328

Design Review for Failure Analysis and Prevention

Crispin Hales and Cheryl Pattin, Triodyne Inc.

Introduction

THE AIM OF THIS ARTICLE is to assist the failure analyst in broadening the initial scope of the investigation of a physical engineering failure in order to identify the root cause of the problem. Analysis methods for determining the loading pattern and other factors responsible for a service failure are described elsewhere in this Volume. However, for effective measures to be taken in preventing future similar failures, it is essential to identify the root cause of the problem, as distinct from the immediate cause of the physical failure itself. The design process that led up to the physical failure is of great importance when trying to get back to the root cause because it is during the design process that all the original contributing factors are brought together to create the components that are assembled into the complete system. The intention here is simply to set the physical failure within a broader context and then to focus on assessing the design process evidence available within that context. The article purposely does not address issues such as manufacturing, operation, and maintenance, except as they relate to the design process, because these other important aspects are covered in detail within other sections of the current Volume. Neither is it intended to provide prescriptive guidance on carrying out the design process, which is covered in detail within *Materials Selection and Design*, Volume 20 of the *ASM Handbook*. Indeed, if the guidelines and systematic design methods described in Volume 20 are appropriately followed, then the likelihood of a failure due to faulty design is minimized in the first place.

What Is an Engineering Failure?

When a piece of material breaks, cracks, corrodes, or otherwise “fails” in service, it is only natural to look at the “failure” with an initial assumption that it should not have happened. While the assumption may be valid in some cases, in many others it is misleading. The “failure” is simply the physical result of a set of preexisting circumstances, a sequence of events, or a developing situation, and it must be considered in its appropriate context.

It is most important that the initial investigation is approached from a broad perspective, rather than from a specialist viewpoint. Many times the perceived problem is circumscribed too early, and the entire investigation becomes focused on one particular aspect, rather than establishing the root cause. The result is that biased conclusions become accepted, and any actions taken for future improvements are not soundly based. To approach the investigation of an engineering failure from a broad perspective means analyzing the design process that created the product, equipment, or system, as well as the physical failure itself. By using a systematic approach to the analysis, it is possible to review basic design issues and work toward the details in a progressive fashion. This helps to ensure that key points are not missed.

For example, a large number of electric appliance motors failed by seizure when elastomeric components, which had been added to reduce noise and vibration, rapidly disintegrated. A large claim was filed against each of the companies in the component supply chain for providing parts manufactured from substandard material. As it happened, some of the elastomeric material supplied to the component manufacturer was “an equivalent” to the material actually ordered and lacked the oil-resisting capabilities of the material actually ordered by the component manufacturer. To the motor manufacturer the cause of the failure appeared clear: Substandard material had been made into components that disintegrated in service, causing premature motor failures. However, an engineering investigation of the matter revealed quite a different cause of failure. Despite the fact that the company was International Organization for Standardization (ISO) 9001 registered, the additional components had been purchased and fitted to the motors at the sole bidding of an employee engineer whose focus was on reducing vibration and noise in the product and who had seen a similar solution applied by another company. The problem in the motors was never actually defined, no alternative concepts were considered, and the elastomeric material for the additional components was improperly specified on both drawings and purchase orders. Instead of listing an appropriate set of elastomeric material requirements, the

documents simply stated a specific manufacturer's product number. It might have been expected that a design review would have caught these deficiencies before the additional components were made and fitted to the motors on the production line, but there was no such design review. This is because the ISO 9001 requirements for design reviews are essentially customer driven. They are intended to ensure that the customer gets what the customer wants. They do not address the situation in which an employee simply decides that the product could be improved by the adding of components and personally implements the change from within the company. The root cause of the failure in this example was a flawed design process with three major deficiencies. The inevitable failure situation was then exacerbated by a material supplier providing “an equivalent” material without the required notification or approval.

The important point here is that for several years afterward the forensic investigation by each of the companies involved was focused solely on the failure of the specific material in service, and an enormous amount of effort was spent in taking depositions, amassing documents, and life testing motors to try and prove one point or another. However, once the root cause of the failure had been identified, the disputed issues changed and the matter was soon resolved. This is typical of a failure investigation in which the initial focus on the material failure itself leads away from the root cause and masks the true issues. Without addressing the design process issues, the quality of future products would still have been at risk.

The Context of an Engineering Failure

Before windowing in on the design process itself, it is helpful to try to visualize how a particular engineering failure fits into the wider context under the circumstances. This is not easy to do because most likely the failure results from a complex sequence of events taking place at different levels of resolution and seen from different perspectives. Nevertheless, ways of mapping such a context diagrammatically have been developed over the years to help in carrying out design projects, such as shown in Fig. 1. Design processes within this context are discussed in “Overview of The Design Process” by Dixon (Ref 2), and “Conceptual and Configuration Design of Products and Assemblies” by Otto and Wood (Ref 3) in Volume 20 of the *ASM Handbook*. The idea in this article is to set the engineering failure in context so as to help in retracing a path both backward in time and outward in perspective. This is necessary in order to identify possible contributing factors to the engineering failure, which at first may not seem related to the physical failure at all.

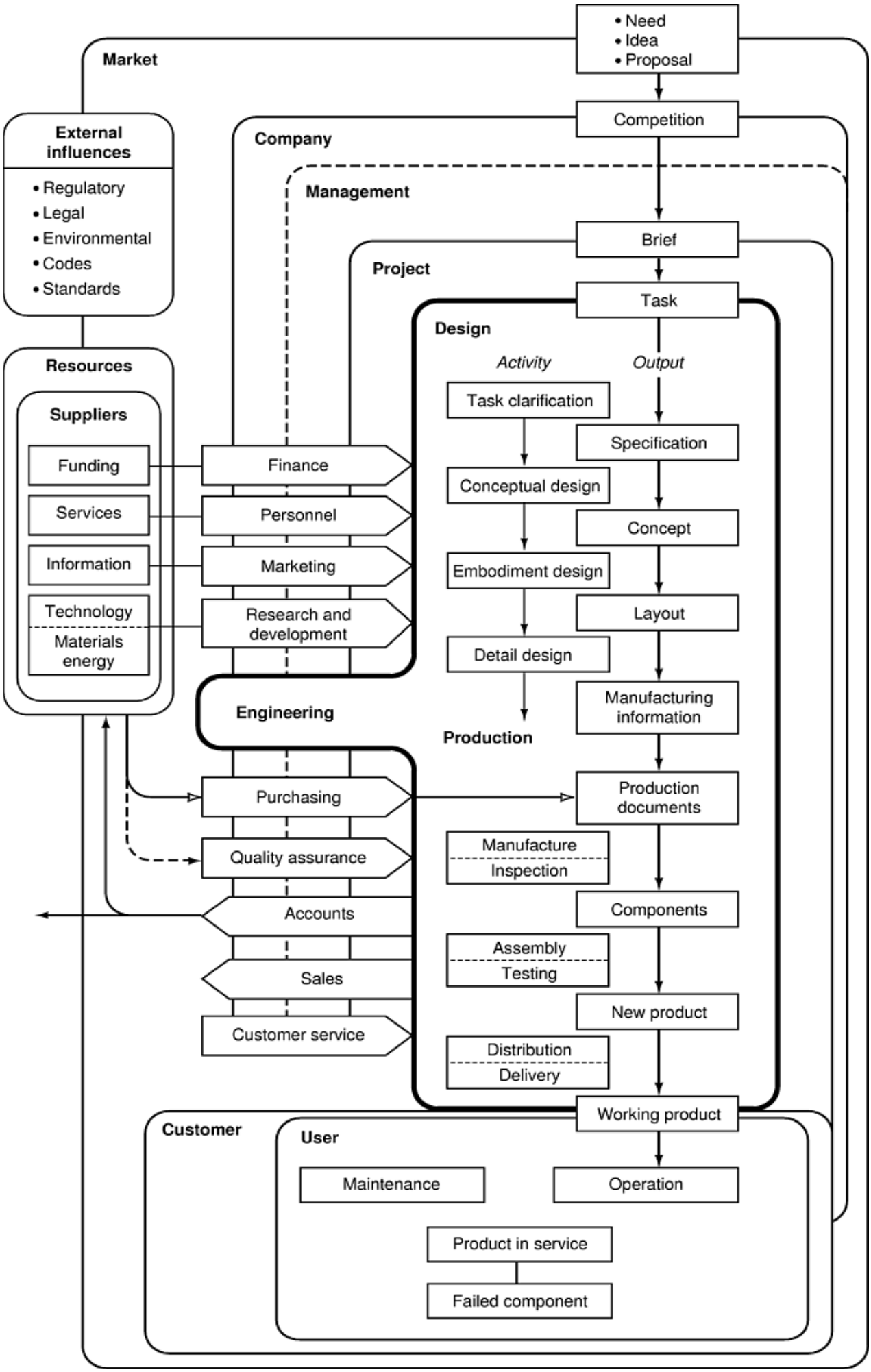


Fig. 1 Engineering failure set in context. Source: Ref 1

The diagram or map in Fig. 1 shows a component as part of a product (or system or structure) being operated by a user and having been purchased as a working product (or system or structure) by a customer through a distribution and shipping process. Tracking backward up the map reveals that before the sale of the product comes the manufacturing process, which may well include the purchase of component parts or semi-finished products from other suppliers. Prior to that there is some kind of design process, which provides the specifications, instructions, and information for manufacture; and initiating the design process in the first place is a commercial process arising from a need, idea, or market force. All of this is part of what may be considered an economic loop. New developments and products are generated by companies tracking around the full loop or by short-circuiting it and modifying existing products without going through the full design process. Also represented on the map is the notion of *levels of resolution*, shown more specifically in Fig. 2. In this simplified diagram, the design and manufacturing processes, integrated together, can be visualized as forming just one of a number of processes (such as finance or marketing) taking place within a project, within a management system, within a company, within a market, within an external environment. At each level of resolution there are influences that impinge on the levels below and thereby affect the outcome of the project and its resulting product (or system or structure). The remainder of this article is concerned only with the design process at the personal and project levels, but takes into consideration the effects of some higher level influences and interfaces that are often found to contribute to engineering failures.

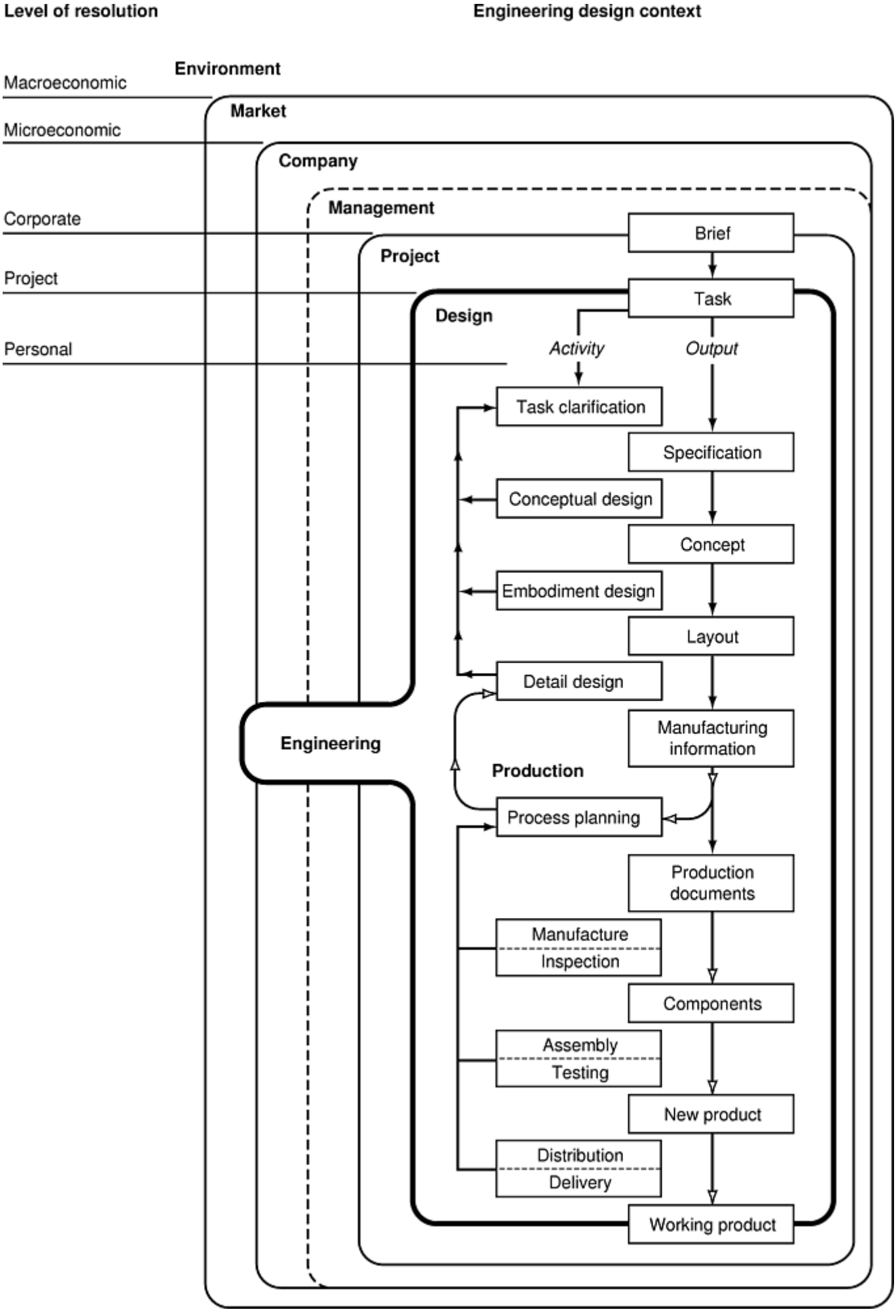


Fig. 2 Levels of resolution related to the engineering design process. Source: Ref 1

References cited in this section

1. C. Hales, *Managing Engineering Design,* Longman Scientific & Technical, 1993; C. Hales and S. Gooch, 2nd ed., Springer-Verlag London Limited, to be published 2003

2. J.R. Dixon, Overview of the Design Process, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 7–14

3. K.N. Otto and K.L. Wood, Conceptual and Configuration Design of Products and Assemblies, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 15–32

The Engineering Design Process

In theory, the basic engineering design process is usually described as a sequence of phases, such as shown in Fig. 2, beginning with a perceived need and finishing with the detailed description of a particular technical system or product (Ref 1, 4, 5, 6, 7). As represented in Fig. 2, each phase involves activities intended to result in particular outcomes. Depending on the product, system, or structure being designed, the phases may be labeled in different ways and will often be carried out in parallel with the design of the manufacturing process (Ref 8). Each phase may be considered as a sub-design process in itself, consisting of an iterative set of steps and commonly summarized as follows:

Step	Purpose	Output
Task clarification	Definition of the problem	Design specification
Conceptual design	Generation, selection, and evaluation of solutions	Design concept
Embodiment design	Development of the concept	Final layout
Detail design	Definition of every component in shape and form	Manufacturing information

In practice it is unlikely for the design process to progress through the sequential set of project phases exactly as outlined, but the sequencing is less important than the existence, nature, and effectiveness of the actual design activities implied within each phase. For example, there must be some kind of design specification as a starting point, and there must be some kind of concept from which a final design evolved. The concept must be developed to a greater or lesser degree so as to result in a practicable overall design, and the details of every component must be defined to the point at which the product or system can be manufactured. A combination of human activities is required to reach each of these end points or outputs, as described by Smith in his article “Cross-Functional Design Teams” (Ref 9) in Volume 20 of the *ASM Handbook.* The activities of the design team, the influences on the design team, and the consequent output from the design team are the focus when analyzing the design process (Ref 1).

In order to analyze the design process factors that may have contributed to an engineering failure, it is necessary to review the phases of the design process in the context of the original project history, its management, prevailing commercial pressures, and external influencing factors. Figure 3 shows a general sequence for such a review. By starting with the failed component and working backward through its history, collecting evidence on how it was designed, developed, and produced, it is often possible to identify and characterize weaknesses that became contributing factors to the failure.

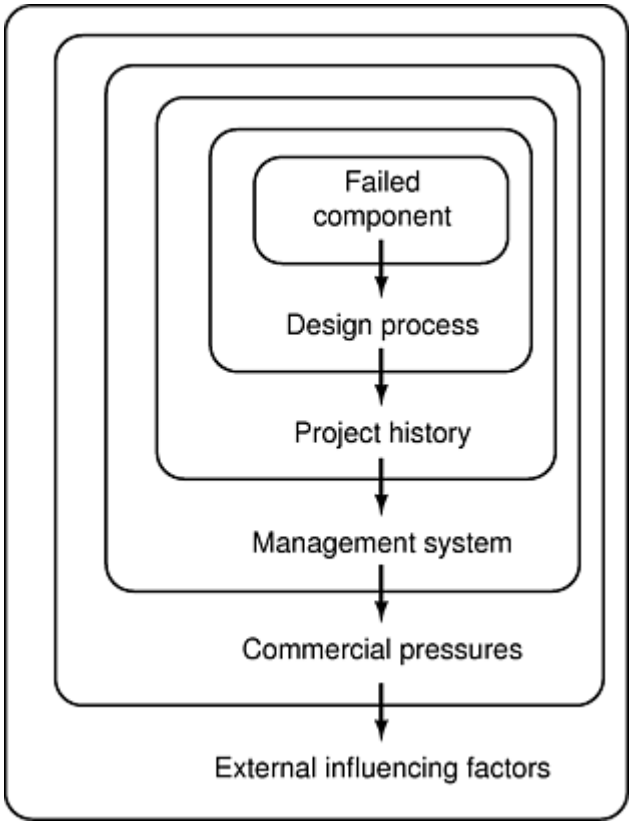


Fig. 3 General sequence of failure investigation from design point of view

References cited in this section

1. C. Hales, *Managing Engineering Design,* Longman Scientific & Technical, 1993; C. Hales and S. Gooch, 2nd ed., Springer-Verlag London Limited, to be published 2003

4. G. Pahl and W. Beitz, *Engineering Design,* K.M. Wallace, Ed., The Design Council, 1984

5. E. Frankenberger, P. Badke-Schaub, and H. Birkhofer, Ed., *Designers—The Key to Successful Product Development,* Springer-Verlag, 1998

6. V. Hubka, *Principles of Engineering Design,* Butterworth Scientific, 1982

7. K.M. Wallace, A Systematic Approach to Engineering Design, *Design Management: A Handbook of Issues and Methods,* M. Oakley, Ed., Basil Blackwell Ltd., 1990

8. M.M. Andreasen and L. Hein, *Integrated Product Development,* IFS (Publications) Ltd. and Springer-Verlag, 1987

9. P.G. Smith, Cross-Functional Design Teams, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 49–53

Preliminary Investigation

When the components in the system that either failed or were associated with the failure are identified, the first step is to request and gather all the available evidence pertaining to the failure. For example, the failed parts themselves, drawings, supply documents, change orders, reports, photographs, videotapes, sales literature, technical information, statements, interview transcripts, deposition transcripts, installation, operating and maintenance manuals, relevant standards and codes—all are potential sources of information that need to be explored. A detailed review of all this material helps to put the failure in perspective and enables a preliminary timeline or outline sequence of events to be compiled.

Analysis of the Engineering Design Process

It is possible to arrive at a design in many different ways, ranging from informal experimentation through to a highly organized formal procedure (Ref 1). When it comes to analyzing any particular case, however, it is essential to have a clear and structured approach to mapping whatever evidence is available, no matter what type of design process is used. The more it can be broken down into measurable components, the less subjective the conclusions will be and the better the chance of developing defensible opinions is. For the purposes of this article, the schematic shown in Fig. 2 represents the basic design and manufacturing process adequately, and the following sections are structured in sequence according to this diagram.

Reference cited in this section

1. C. Hales, *Managing Engineering Design,* Longman Scientific & Technical, 1993; C. Hales and S. Gooch, 2nd ed., Springer-Verlag London Limited, to be published 2003

Task Clarification—Defining the Problem

In order for a design project to be carried out, two things need to be established at the outset, each being a complementary result of the "task clarification" phase of the design process:

- A clear statement of the problem to be solved, for which solutions will be sought
- A set of requirements and constraints against which to evaluate the proposed solutions

The first is termed a *problem statement,* or *definition of the problem,* while the second is termed a *specification,* a *target specification,* or more correctly a *design specification.* Both of these are essential if a solution to the problem that satisfies all parties is to be found. Considerable effort (and possibly some preliminary design work) may be needed to help establish what the real problem is, but it must be done. Finding solutions to the wrong problem is unacceptable design practice. Once the problem is defined, the criteria for selecting an appropriate concept must be established in the form of a design specification that lists all the requirements to be met by any solution to the problem. Here again, if the requirements are inaccurate or incomplete, then the design process is flawed from the start.

Frequently it is found that the real design problem is never clearly defined, is incorrectly defined, or the wrong problem is identified, as described in the introductory example. This can be investigated by reference to the original bid documents or equipment quotations and the sequence of communications between supplier and customer that led up to them. The wording of all this needs to be reviewed in detail and with a good understanding of the equipment, as it is easy to miss specification deficiencies that had consequential ill effects on the final design.

During the review it is useful to ask a series of questions concerning the design task, the team that carried out the job, and the details of the team's activities during each phase of the design process. For example:

- Was the overall design problem understood and clearly defined prior to conceptual design work?
- Was a realistic project plan prepared, acceptable to all parties?
- Was a comprehensive design specification compiled?
- Is the design specification independent of solutions, or does it include fictitious constraints?
- Was the design specification circulated to all those involved for comment and approval?
- Was the design specification formally approved before work proceeded on conceptual design?
- Where is the design specification?
- Who developed it?
- Who approved it?
- What changes were made? Who made them? Why? When? How?

Progressing through a checklist of questions for each phase of the design process allows problems and weaknesses to start to emerge. A properly structured design specification provides the maximum design freedom within the given constraints. The degree to which this is carried out may be checked by reference to

standard specifications for the particular equipment involved and checklists such as the general ones offered by Pahl & Beitz (Ref 4) and Hales (Ref 1).

Defining the design problem and listing the requirements in the form of a design specification provides a proper foundation for the project to proceed through the conceptual, embodiment, and detail design phases. Solutions to the defined problem may then be sought by the use of conceptual design methods, and the resulting concepts may be evaluated against the design specification.

The following example illustrates the pitfall of introducing fictitious constraints into the requirements. The result is that the concept selected will be over-constrained and therefore not an optimal solution to the problem.

A large number of custom-designed vertical lift conveyors were required for use in a series of new automated U.S. mail-sorting facilities. These facilities comprise essentially a series of code-reading sorting units at ground level with a large array of horizontal roller conveyors above. Vertical lift conveyors are used to transport plastic trays of sorted mail from ground level up to the horizontal conveyor level, and the reverse. The vertical lifting concept was developed to overcome the slippage problem encountered with inclined roller conveyors when plastic trays superseded the cardboard trays used in earlier facilities.

The general contractor had prepared a voluminous design specification for the post office. This was to apply to all facilities, but for each particular facility there was also a detailed set of special requirements. Within these documents were embedded the requirements for the conveying systems in general, and within those the requirements for the vertical lift units in particular. Prototype vertical lift units that used twin in-running belts to deliver the trays to three-fingered lift platforms already had been developed and tested off-line to the point of acceptable performance for this application. The post office design specification for the lifting units was compiled with the prototype units in mind, even to the point of requiring in-running entry and out-running exit conveyors having two or more belts, and lift platforms with three or more fingers. In such a way the specification unnecessarily constrained the design to a specific concept that was known to have inherent operational problems. The many required units were designed, built, tested, delivered, and installed at a cost of more than $500,000. While some design weaknesses and manufacturing problems were evident, which detracted from the performance of the units, they did pass the acceptance tests laid down in the design specifications, right up to the final production trials. At this point, however, multiple component failures and jams were encountered because the operators were using the vertical lift units on a start-stop basis instead of continuously as intended. It was claimed that the units did not meet the specification, and they were all removed from service and scrapped without payment to the manufacturer. At the same time, another supplier was contracted to provide quite different replacement units. Investigation revealed that a modified design specification had been issued to alternative manufacturers who were invited to bid on supplying the replacement units, and that the units finally selected were based on a concept that could never have met the requirements of the original design specification. In the modified design specification, the requirements for belt conveyors and for the lift platforms to have three or more fingers had been deleted. This removed the fictitious constraints on the design and allowed the use of an articulated slat conveyor, far superior in concept to the suspended tray type of conveyor for this particular application.

The root cause of this failure was a deficient design specification, not the component failures themselves. The result was an enormous waste of effort, money, and materials, as well as the bankruptcy of the original vertical lift unit manufacturer.

References cited in this section

1. C. Hales, *Managing Engineering Design,* Longman Scientific & Technical, 1993; C. Hales and S. Gooch, 2nd ed., Springer-Verlag London Limited, to be published 2003

4. G. Pahl and W. Beitz, *Engineering Design,* K.M. Wallace, Ed., The Design Council, 1984

Conceptual Design

Once the problem is defined, it is possible to start generating ideas leading to concepts that will solve the problem (Ref 10, 11). Conceptual design is also discussed in the article “Creative Concept Development” (Ref 12) in Volume 20 of the *ASM Handbook.* This is not the same as invention. What is meant by *conceptual design* is the conscious activity of generating numerous ideas leading to specific concepts that are then selected and

evaluated according to how well they meet the requirements of the design specification. Of course a design concept may become patented as an invention in intellectual property terms, but an invention is not necessarily a design. There is a fundamental difference between an inventor happening to have an inspiration on Thursday and a design engineer producing an acceptable design concept within time, budget, and specification constraints by Thursday. An important part of the necessary concept evaluation activity is searching for weak spots. For example, if an otherwise excellent concept has an inherent reliability problem, it may have to be rejected in favor of perhaps a less innovative concept, but one that is inherently more reliable.

Many times it will be revealed during the investigation that the concept for a design has either simply been assumed, or else has been adapted from a different product, without a genuine search for alternatives or formal assessment against the design specification. Although on the surface it may appear to perform adequately, a non-optimal concept creates secondary problems that then must be addressed by further concepts, which themselves may lead to an unexpected or premature failure. When analyzing the design process, it is necessary to isolate and visualize the principal concept within the context of the project, and to assess how well it meets the requirements of the design specification. Again, the checklist approach may be used for this (Ref 1), but often a simple set of questions will suffice, such as:

- What alternative concepts were produced?
- How were the various alternatives assessed?
- Were all alternatives considered on an equal basis?
- Who decided what concept should be used and how was the decision made?
- Were the weak points of the chosen concept adequately evaluated?
- Why were alternative concepts rejected?

In addition to the example presented in the introduction, the following example illustrates the kind of problem that can occur from the selection of a wrong concept for the application:

A post office in Michigan was fitted with heavy doors that could swing more than 90° from their closed position in either direction. The doors had alloy frames with large glass panels and were fitted with mechanical door closers that always returned the door to its closed position with a “damped spring” action after use. In order to save maintenance and inventory costs, the original architect had specified the same doors throughout the building. Although physically the door could be used as an exterior door, a number of features such as the lack of positive sealing indicated that it was primarily a door for internal use. However, there was no requirement to this effect, and the interchangeability of the doors meant that if, for example, the glass in an external door was ever broken, one of the internal doors could immediately be moved to replace it so that the building security would not be compromised. The external door closest to the public parking area was the most heavily used door in the building. It was exposed to all weathers and the automatic device for damping the door movement and closing it after use required frequent maintenance.

One windy Saturday, a woman was about to enter the post office through this door when a gust of wind blew it first inward and then outward, evidently without any damper control. The door swung outward and then right around, until it hit the edge of the building wall close to its axis of rotation. This caused it to lever itself from its mountings and fall to the ground, killing the woman as it fell. Improper maintenance leading to premature wear of components was claimed.

A detailed investigation revealed that the door closer was not working properly at the time of the accident, allowing the door to swing freely in either direction. In addition, the door had a specially designed hinge system to facilitate the interchange of doors. Mounted on the bottom threshold was a roller that fitted into a fixed socket located within the door, forming the bottom “hinge.” At the top, the door closer unit itself acted as the hinge, mounted on the lintel overhead with its torsion arm connected to the top of the door by means of a spring-loaded quick release mechanism. Heavy usage had resulted in distortion of the door closer torsion arm, sufficient to displace the spring-loaded plunger arrangement almost to the point of release. Uncontrolled swinging, combined with a leverage action when the door hit the wall, generated enough force to separate the door completely from the door closer at the time of the accident.

The design concept of using this particular spring-loaded plunger mechanism as part of the door hinge system was inappropriate in that there was no positive connection between the door and its top “hinge.” A later model of this door design had been installed elsewhere inside the post office, with a bolted connection replacing the quick-release mechanism. If this had been fitted to the door in question, there would have been no accident.

References cited in this section

1. C. Hales, *Managing Engineering Design,* Longman Scientific & Technical, 1993; C. Hales and S. Gooch, 2nd ed., Springer-Verlag London Limited, to be published 2003

10. S. Pugh, *Total Design—Integrated Methods for Successful Product Engineering,* Addison-Wesley Publishing Company, 1990

11. M. Khan and D.G. Smith, Overcoming Conceptual Barriers—by Systematic Design, C377/079, WDK-18, *Proc. ICED-89,* I.Mech.E., 1989

12. B.L. Tuttle, Creative Concept Development, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 39–48

Embodiment Design

Embodiment Design Issues. During the phase following conceptual design, the selected concept must be developed into a practical, reliable, and safe design. This phase is often called *embodiment design,* as there are systematic guidelines available for the progressive development of a concept (Ref 1, 4, 5) quite different in nature from those used in the detail design of individual components. These guidelines may be used in the form of a checklist to help identify common deficiencies in developing a particular design. For example, the following issues should be reviewed:

- Simplicity
- Clarity of function
- Safety
- Selection of materials, products, and parts
- Transportation

Simplicity. It is important for a design to be as simple as possible while still meeting the requirements of the design specification. This may be achieved, for example, by reducing the number of components, making components do more than one function, and promoting the use of near net shape manufacturing techniques. It is all too easy to become enamored of exciting new technology or a complex way of doing something, to the detriment of the final result.

For example, some types of vegetables in supermarket display cabinets are misted with water to keep them fresh. This poses design problems such as control of the spray and how to avoid startling customers who are selecting produce as the water jets spray without warning. One innovative design was fitted with an enclosed stainless steel trough, level controls, four ultrasonic misting devices, oscillator circuits, a blower, and electromechanical controls. It was indeed a complex mechatronic marvel, the workings of which could be neither seen nor understood without disassembly. Although it functioned adequately, it was very difficult to clean. The result was that it was left alone, and it began to spray harmful bacteria as well as the misted water. Finally a series of customers fell victim to Legionnaire's Disease, which was traced back to this mechatronic misting device. A survey of designs by other manufacturers showed that the misting could be done just as effectively with simple valves, timers, and sprays, without the risk of harboring bacteria.

Clarity of Function. Clarity in design (Ref 1, 4) means making sure that the design itself explains how it is to be put together, what the load paths are, what the function of each component is, and how the parts move relative to each other. Specific guidelines are available for addressing this issue in a systematic fashion (Ref 4), and in conjunction with other embodiment design guidelines, many types of failure can be avoided. Lack of clarity in design creates ambiguities and design weaknesses that may not be immediately obvious. This is illustrated by the following example.

Heavy trucks usually have a beam axle front suspension with a kingpin mounted near vertically through the eye at each end of the axle. The yoke of each front wheel stub axle fits over the corresponding eye of the beam axle and around the kingpin, thereby forming the axis about which the stub axle can turn in order to steer the truck. Kingpins are normally simple cylindrical hardened shafts extending out of the top and bottom of the axle eye

into the top and bottom bearings of the stub axle yoke. A thrust bearing is fitted between the bottom of the axle eye and the bottom bearings in the yoke to accept the vehicle weight from the axle eye. The forces are thus transmitted from axle eye to yoke without undue friction when steering action takes place. The kingpin is locked in place by means of a cross bolt arrangement through the axle eye.

In Ohio, a heavy dump truck was coming around a right curve in the road when the front left wheel assembly suddenly collapsed and separated from the truck. As the steering box was linked to this particular stub axle, the driver immediately lost all steering control. The vehicle went straight ahead, colliding head-on with a car going the opposite way and killing the passenger in the car.

Investigation revealed that this particular model of dump truck is fitted with a tapered kingpin inserted from the bottom. It is held in place by friction against the tapered hole, together with a nut and washer arrangement at the top. The nut had come loose and the kingpin had dropped progressively as the nut turned. When the nut unscrewed completely, the kingpin fell right out of the axle eye, causing the whole wheel assembly to separate from the truck. In conventional arrangements the vertical component of the truck weight is always transmitted from springs to beam axle and from beam axle to stub axle through the thrust bearing below the axle eye. Even if the thrust bearing were to disintegrate, the load path would still pass through the same components. With the dump truck, however, as wear takes place in the thrust bearing, the clearance between the nut plus washer at the top of the kingpin and the top of the stub axle yoke decreases to the point where the load path changes. The weight of the truck transfers from the thrust bearing to the nut and washer. It is only a matter of time before the torsional friction forces are sufficient to shear the pin that locks the nut, and then to undo the nut in a progressive manner.

In this case the embodiment design was deficient in that it allowed an unacceptable load path change with wear or failure of the thrust bearing. This would have been prevented by the use of accepted embodiment design guidelines, and the accident would not have happened.

Safety. Safety is something to consider throughout the design process, but there are specific embodiment design guidelines such as the safety hierarchy (Ref 13) to follow when developing a design concept. Techniques such as safe-life design, fail-safe design, redundant design, and hazard analysis (Ref 4, 5) must be considered as integral parts of the product development, and not simply as “add-ons” at the end of the project. Safety considerations and hazard analysis in design are discussed further in the articles “Safety in Design” (Ref 14) and “Risk and Hazard Analysis in Design” (Ref 15) in Volume 20 of the *ASM Handbook*. The focus of product liability lawsuits is on the safety aspects of design (Ref 16). It is often claimed that the litigation has the effect of “improving safety” (Ref 17). However, it is probably more accurate to say that such lawsuits are a very expensive way of addressing safety, and that the results are unpredictable (Ref 18). Sometimes safety is improved; sometimes it is unchanged; and sometimes it is compromised. For example, the automobile airbag raises several issues concerning safety in design:

A man was driving at night when a deer leapt into the road right in front of his car. On impact the front of the car knocked the deer's legs from under him and as he rolled through the windshield, the driver's airbag was activated, apparently causing the man's hands to be forced upward into the path of the deer's antlers. An antler caught one of the man's thumbs and tore it off. A lawsuit was filed against the manufacturer, claiming that if the airbag had not deployed, the man would not have lost his thumb.

Would the man have been injured less or more if there had been no airbag? What if the airbag had been there but had not deployed? What if the man had lost control after the airbag deployed and had driven into a tree without any airbag protection? Did the airbag actually protect the man from more serious injuries upon impact by the deer?

If a failure appears to involve safety issues, then a simple starting point in the analysis is to work through the safety hierarchy in the form of the following questions:

- Were the hazards/risks associated with the device eliminated by design as far as practicable?
- Were appropriate protective systems incorporated for remaining hazards and/or risks?
- Were appropriate warnings provided for residual hazards and/or risks?
- Was there provision for instruction and training?
- Was personal protection prescribed (if necessary and as a last resort)?

With regard to large equipment and commercial systems, a hazard analysis, failure modes and effects analysis (FMEA), or other type of formal safety analysis is often required as part of the design process. While such

analyses do help to identify potential problems, and help to reduce the risk of a crippling failure or accident, they are very dependent on the initial assumptions. Many times a catastrophic failure occurs, despite all the elaborate analysis, simply because an assumption was made that *no one would ever do such a thing.* For example, if the absolute minimum allowable brake rotor thickness and the minimum allowable brake pad thickness are both defined in the maintenance manual for an automatic passenger transit system, is it appropriate to assume that the professional maintenance staff, dedicated to ensuring the safe operation of the vehicles, would replace the pads and/or rotors rather than let every pad and every rotor on the vehicle wear to below the absolute minimum? If it is assumed that these specific instructions will not be followed, then what are all the other instructions that must then fall under the same assumption? It may sound ridiculous, but this is exactly what happened, and it was claimed that the ensuing crash caused by pads no longer touching the rotors during braking was caused by a faulty FMEA!

In recent years the legal community has so overreached in cases involving claims over “lack of warnings” that the situation has become untenable for many products. Plastering complex warning notices over every conceivable surface of a product in letters so small that a microscope is required to read them contributes nothing toward safety. However, carefully designed warnings, applied in appropriate circumstances, can be an effective means of preventing accidents and failures.

Selection of Materials, Products, and Parts. Another critical aspect of design, which needs to be addressed formally in the embodiment phase, is the selection of materials, semi-finished products, and standard components. Materials selection in mechanical design is discussed extensively in the article “Overview of the Materials Selection Process” in Volume 20 of the *ASM Handbook* (Ref 19). Detailed methods are provided in Ashby's *Materials Selection in Mechanical Design* (Ref 20). Material selection should be carefully reviewed in light of all intended service conditions to foresee any conflicts, as discussed in the article “Materials Selection for Failure Prevention” in this Volume (Ref 21). For example, a seal might be expected to perform in an operating environment at 150 °C (302 °F). Butyl rubber is often an excellent material choice for seals, but in this case the extensive creep of butyl rubber at high temperature would rule it out.

Sometimes a given part is chosen in order to improve performance in one area while other aspects of the service environment are ignored. In the example described at the beginning of this article, elastomeric components were selected for their capability to reduce noise and vibration in motors, but they degraded rapidly in service due to insufficient oil resistance.

There are numerous aids available for determining material behavior. These resources include graphical materials selection charts (Ref 20), which are appropriate for the selection of a broad class of materials for a given application (such as polymers). Materials selection at this level is appropriate for conceptual design. Detailed web-based software, such as the Cambridge Engineering Selector (Ref 22) or vendor databases, allow the designer to obtain properties of individual materials, often useful in the detail design phase of a project.

It is obvious that if an incorrect selection is made during design, then the risk of failure will be high. What is less obvious is the increasing problem of substandard materials and copied or counterfeit parts. Some industries, such as the aircraft and petroleum industries, have procedures to ensure that supplied components are in fact what they are claimed to be. However, many others have a long way to go in addressing this potential cause for failure, as shown by the following example:

A replacement strainer was fitted into a process steam line in a chicken feed plant. Soon afterward, and luckily during a weekend, the cast iron wall of the strainer failed. As a result of the explosion, the entire plant had to be cleaned and repainted from the steam damage, and of course there were claims and counterclaims over who was at fault. Thickness measurements at the rupture site on the wall of the imported strainer showed the wall thickness to be well below the minimum required by applicable American National Standards. The sales drawing (cut sheet) for the strainer indicated “ISO 9002 Quality Assurance,” yet there was no evidence of any International Organization for Standardization (ISO) 9000 registration nor even that the company staff knew what the term meant. This imported product was held out to be equivalent to those produced by U.S. manufacturers. It demonstrated clearly that it was not.

Transportation Issues. Although at first it may not seem appropriate to review transportation issues when analyzing possible flaws in the engineering design process, it is surprising how many accidents and failures stem from the lack of attention to transportation issues at the design stage. During transportation, a material, component subassembly, or complete product is subjected to conditions completely different from those to be expected during manufacture or subsequent operation in service. The environmental conditions are different; the loading conditions are different; and even the perception of the item being transported is different. For

example, it is common for machines to have a center of gravity that is biased toward one side, or that may be above the midpoint of the height. However, when crated the machine can no longer be seen, and to the shipper it is considered simply a box of a certain size and weight to be conveyed from pickup point to delivery point using agreed modes of transportation. Unless there are clear warnings and instructions with regard to handling of "the box," the shipper may not realize the high risk of overturning and damage due to load shifting. Similarly, when vehicles being shipped are tied down to prevent movement, the rolling elements of axle bearings are held in a fixed position under load and may experience impact loads high enough to cause damage from brinelling. This type of issue should be addressed no later than the embodiment design phase of the design process in order to avoid problems during delivery and afterwards.

References cited in this section

1. C. Hales, *Managing Engineering Design,* Longman Scientific & Technical, 1993; C. Hales and S. Gooch, 2nd ed., Springer-Verlag London Limited, to be published 2003

4. G. Pahl and W. Beitz, *Engineering Design,* K.M. Wallace, Ed., The Design Council, 1984

5. E. Frankenberger, P. Badke-Schaub, and H. Birkhofer, Ed., *Designers—The Key to Successful Product Development,* Springer-Verlag, 1998

13. R.L. Barnett and W.G. Switalski, Principles of Human Safety, Safety Brief, Vol 5 (No. 1), Triodyne Inc., 1988

14. C.O. Smith, Safety in Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 139–145

15. G. Kardos, Risk and Hazard Analysis in Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 117–125

16. T. Willis, M.P. Kaplan, and M.B. Kane, Safety in Design: an American Experience, C377/110, *Proc. of the Institution of Mechanical Engineers International Conference on Engineering Design (ICED-89),* Institution of Mechanical Engineers, 1989

17. C. Hales, Legal Threats to Innovation in Design, *Proc. ICED 99: International Conference on Engineering Design,* Institution of Mechanical Engineers, 1999

18. P.W. Huber, *Liability: The Legal Revolution and Its Consequences,* Basic Books, Inc., 1988

19. G.E. Dieter (chairperson), Overview of the Materials Selection Process, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 241–328

20. M.F. Ashby, *Materials Selection in Mechanical Design,* 2nd ed., Butterworth-Heinmann, 1999

21. B.A. Miller, Materials Selection for Failure Prevention, *Failure Analysis,* Vol 11, *ASM Handbook,* ASM International, 2002

22. M.F. Ashby and D. Cebon, Cambridge Engineering Selector, CES Software Version 3.2, Granta Design Limited (www.grantadesign.com), 2001

Detail Design

Once a design has passed into the detail design phase, it used to be common for it to be *given to the draftsman* to finish. It is a fatal mistake to think that detail design is unimportant and needs less attention than other phases. While it is true that excellent detail design cannot compensate for a bad concept, it is equally true that

poor detail design can ruin a good concept. Detail design is critical and it needs to be understood and investigated when tracking the root cause of an engineering failure. As shown by the following example, this is not just a matter of considering individual components, but can also involve the interaction of various materials in an assembly or with the environment.

An articulated tractor and low-loading trailer was being driven through some hilly country in Missouri after delivery of a bulldozer to a construction site. The driver noticed that his rear trailer brakes had started to smoke, and he found that the brakes were partially applied even when he wasn't using them. He assumed that the brakes were running hotter because he was driving faster through the hills without the load of the bulldozer, and that this heat was causing his brakes to drag. Using a wrench he backed off the brake shoes from each wheel to the point at which he thought they were correctly adjusted. He telephoned his dispatcher from the next rest area, explained what had happened, and was told that he should continue with his journey. A few miles later the weather changed to a misty drizzle, and far ahead he saw that a minor rear-end accident had just occurred at an intersection. He applied his brakes but found that there were now no brakes on his heavy trailer. People started running about on the road and shoulder ahead of him, so he could not steer around. He decided all he could do was to apply his brakes as hard as he could. The tractor-trailer jackknifed completely and collided head-on with a vehicle coming the other way. One person was killed and several were seriously injured.

Large articulated trucks generally have air-operated braking systems, and certain types of low-loading trailers require the use of a particular relay valve because of their physical layout. The relay valve directs and releases air, depending on signals from the driver's brake pedal, trailer brake controls, and the integrity of the air system. An inspection of the valve from the accident trailer showed that a small steel valve head with a 5 mm threaded stem had come loose. The valve head hangs down from a piston, and when the threaded connection came loose it progressively unscrewed to the point that the valve head blocked an air exhaust port, thus causing the brakes to remain partially applied. Without knowing it, the driver had backed off his brake shoes to the point that they could not come into contact with the drums when the brakes were applied. It was later found that the detail design of this particular threaded connection had been revised and tested numerous times by the manufacturer and that several thread-locking methods had been put into production over the years. The thread involved in the accident had not been locked in place, and the loose thread tolerance allowed complete unscrewing to occur.

The detail design was flawed in that a critical component was hung vertically inside a valve by a loose 5 mm thread. The unscrewing of the thread resulted in the death of one person and serious injuries to several others. It was a known problem, which had been brought to the attention of management, and if it had been addressed properly the accident would not have happened.

Management Influences

Management Control. One of the most frustrating things for design engineers is the way projects are manipulated by those who have very little to do with the design process itself. It is critical for the design manager or team leader not only to be aware of the impact of various influences, but also to exercise control over those that can be controlled and compensate for those that cannot, in the best interests of the customer, the project, and the design team.

Design team activities must be directed and monitored for performance. The design output must be assessed against the specification requirements continually. The effect of influencing factors must be actively predicted, monitored, and controlled when possible. Management involvement in these issues is crucial to the development of high-quality and cost-competitive products (Ref 1, 23, 24, 25). From the design management point of view, the ultimate goal is to produce the highest quality product, meeting the user's expectations for the lowest cost in the shortest time.

A particular challenge in the management of engineering design is to be able to cope with issues that range from “hard” to “soft”—for example, from the dimensional tolerance on a single component to the user's satisfaction with a product in service. Another challenge is that the critical issues must be considered at different levels of resolution and from different points of view. The design manager must be able to see the overall picture while rapidly windowing in on the details and understanding the effect that even tiny details might have on the overall project. A lack of management skill in this area has contributed to many engineering disasters, such as the failure of the solid rocket booster (SRB) on the space shuttle Challenger. Again, simple sets of questions, based on fundamental design principles and asked at the appropriate time by a manager with

adequate technical understanding, can highlight design weaknesses long before a disaster becomes inevitable (Ref 26).

While such issues are not often the focus of a forensic engineer's investigation, it is important to be aware of the influence of the management factor and to be able to recognize situations in which it contributes to the failure. For example, it turned out that a bearing failure in a large industrial dynamometer was the result of the withholding of design information from a new employee by one who felt that his position had been usurped. What started as a routine bearing and material failure analysis ended up as a seminar for the company on project management.

Design Team. Nowadays it is common for design teams to be created and staffed on a project-by-project basis, and specifically for the duration of the project. Increasingly it is common for the work to be carried out in multiple locations by means of electronic communication. Bringing together and orchestrating a team to produce a quality design in a timely fashion is not easy and must be recognized as a critical task. The article "Cross-Functional Design Teams" (Ref 9) in Volume 20 of the *ASM Handbook* addresses this aspect of engineering design.

People have a functional role in the team, using their particular technical expertise and experience, and obviously this has to be matched to the work at hand. They also have a team role, using their particular character traits to help make the team operate effectively as a team (Ref 24). If the set of team roles is not well balanced, then the output suffers badly, no matter how good the balance of functional roles is. A team may be adequate in a functional sense, having the right expertise and experience, yet may not have the right balance of personalities to be productive. Teams need a mix of personalities covering basic "team roles," with the addition of "specialist" roles in technical situations (Ref 1).

The negotiating ability and the negotiating power of the team are critical to the design process. To be successful, a design team needs to be good at negotiating, and it needs to negotiate from a position of power. Often design teams have more power than they realize, for without their input the company would fail. The more incisively the design team can present its case, the better it is able to control the things that matter. This is illustrated by the failure of the space shuttle Challenger. The failure of management to comprehend the importance of detail design, coupled with the failure of the design team to get the message across, set the scene for the failure. If the design team had understood and learned how to use its latent power effectively, Challenger would not have been launched. It is interesting to note that the great engineers of the world, such as Eiffel, Brunel, and Ford, were not only excellent technically, but also were persuasive, entertaining, and politically involved individuals.

Communication Problems. The importance of effectively communicating the results of the design process is discussed in the article "Documenting and Communicating the Design" (Ref 27) in Volume 20 of the *ASM Handbook.* In fact, effective communication has become even more important as the idea of *global design teams* gains in popularity. The recent failure of the Mars Orbiter mission due to mixed imperial and metric units highlights this type of problem, which becomes compounded by differences in language and culture. What used to be a matter of simple transmittal of information is now a matter of ascertaining what someone actually interpreted from the information provided. The following example shows how easy it for a failure to occur when communication is ineffective during the design process.

To upgrade its production facilities, a European manufacturer ordered two huge custom-built metalworking presses, high on the scales of novelty, complexity, and cost, with a short time frame for delivery. The concept involved a unique automatic transfer system for progressing parts longitudinally through a series of operating stations, so the design work was subcontracted to an experienced North American company with special knowledge of the particular type of machine required. As the time frame was so short, the documentation of customer requirements was perfunctory, and there was some confusion as to the responsibilities of the various parties. Under pressure to meet unrealistic deadlines, the design was rushed through to detailing with insufficient time spent either on the concept evaluation or on development of the concept to meet the design specification. Some 4000 drawings were produced by hand, involving translation from one language to another and changes from imperial to metric units. The machines were built according to the translated drawings with almost no communication between the design company and the manufacturer. Apparently it was assumed that the knowledge of the design company would be fully imparted to the manufacturer through the medium of the drawings. The time frame was too short to allow for testing and commissioning, so the machines were built and put into service without normal shakedown procedures. Although slow-speed operation was achieved, the machines could not be run at the speed specified by the customer and agreed to by the manufacturer. Severe

mechanical damage to components was caused in the attempts to reach full speed. The design work was never paid for, and the claims for damages far exceeded the total cost of design and manufacture of the machines. Inadequate communication, a deficient design specification, and the separation of design from manufacture all contributed to the inevitable failure.

Time and Money Constraints. It is an unfortunate fact that the design process takes time and costs money at an early stage in a project, long before there can be any return on investment. Even still, because the outcome of the design process sets all future costs on a project, common sense would dictate that sufficient time and money should be invested in the design process to ensure that the optimal design is produced to minimize downstream costs. In practice it is more often the case that time and money constraints force the design team to take shortcuts, resulting in costly mistakes or design weaknesses. By breaking the design process down into phases for the purpose of analysis, it is possible to check the amount of effort (ideally in hours) that is spent on design activities during the course of a project and thereby come to a conclusion as to whether the time and money constraints contributed to the failure. For example, in a case where a company claimed to have developed its own technology without copying from others, an analysis of the time spent by each different team member between the start and finish dates showed that the job could not have been done in that time except by copying an existing design.

References cited in this section

1. C. Hales, *Managing Engineering Design,* Longman Scientific & Technical, 1993; C. Hales and S. Gooch, 2nd ed., Springer-Verlag London Limited, to be published 2003

9. P.G. Smith, Cross-Functional Design Teams, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 49–53

23. "Guide to Managing Product Design," *BS 7000:1989,* British Standards Institution, 1989

24. R. Stetter and D.G. Ullman, Team-Roles in Mechanical Design, *8th International Conference on Design Theory and Methodology, 96-DETC/DTM-1508,* J. Cagan and K.L. Wood, Ed., American Society of Mechanical Engineers, 1996, p 1–8

25. E.E. Rothschild, *Product Development Management,* T. Wilson Publishing Company, 1987

26. C. Hales, Analysis of an Engineering Design—the Space Shuttle Challenger, *Engineering Design and Manufacturing Management,* A.E. Samuel, Ed., Elsevier, 1989

27. G. Vrsek, Documenting and Communicating the Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 222–230

External Influences

In coming up with a design that meets the user's needs in the best way possible, trade-offs must be made among the requirements of function, safety, timeliness, cost, ergonomics, the environment, and aesthetics. Often, for example, safety systems can be improved, but at additional cost. In this case the question becomes one of *how safe is safe enough.* Established standards and codes provide the design engineer with a basis for making such judgments in a professionally acceptable manner, and regulations help to enforce accepted levels of practice. The effects of codes and standards on design are discussed in the article "Designing to Codes and Standards" (Ref 28) in Volume 20 of the *ASM Handbook.* In general, standards are more concerned with setting a level of performance, quality, or safety by the definition of criteria, while codes are more concerned with ensuring a level of performance, quality, or safety through adherence to a set of rules or guidelines. The variety of each is enormous; the requirements vary from area to area; and commonly there are inconsistencies that are complicated to resolve. It may be very difficult for the design engineer even to determine which standards or codes apply under particular circumstances, let alone to interpret the details of the fine print. Some are in ISO units; others are in Imperial units. Some are international; others are national; others are regional, and others are

local. Some are specific to a particular product, while others are more generic. Some deal with the minute details of a material composition, while others deal with the testing of whole assemblies or lay down safety procedures. Even the terminology used varies from one document to another, and subtle differences in meaning can sometimes lead to expensive misunderstandings.

When analyzing a failure from the design point of view, it is often important to identify which standards and codes applied to the product at the time of design, as well as at the time of the failure. It is then necessary to determine whether there has been a violation and whether the requirements are legislated and mandatory, voluntary, or just accepted professional engineering practice. There are now standards for carrying out and managing the design process itself, such as VDI Guideline 2221 (Ref 29) and British Standard BS 7000 (Ref 23); with regard to the broader issue of product quality, the series of *ISO 9000 International Standards for Quality Management* is a key reference.

The most controversial standards are often those concerned with safety. They are of concern to the design manager because they strongly influence the design, operation, and maintenance of technical systems and products. A safety standard is a document that attempts to specify components and practices that will result in predictable and acceptable levels of safety. The concept of what is safe needs more careful definition in the design context than in general usage, and its definition is dependent on a set of related ones, as shown by the following definitions adapted from Hebert and Uzgiris (Ref 30):

accident.
: An undesirable event or failure that results in *harm*

harm.
: An adverse effect that occurs in an *accident*

hazard.
: A condition or situation exhibiting the potential for causing *harm*

risk.
: A measure of the probability and severity of *harm:* the potential of a *hazard* to cause *harm*

safe.
: A characterization of a machine, product, process, or practice whose attendant *risks* are judged to be acceptable

safety.
: A state or condition wherein people and property are exposed to a level of *risk* that is judged to be acceptable

safety standard.
: A set of criteria or means for achieving a level of *risk* that is judged to be acceptable by the body formulating the *safety standard.*

In terms of these definitions, the task of the design engineer is to design a safe system by identifying the hazards and controlling the associated risks to within acceptable limits. The criteria for what is acceptable are set, in part, by safety standards. Although compliance with applicable safety standards is generally understood to be a necessary condition for safe design, it may not be a sufficient condition (Ref 31). The standard may not have kept pace with industry or new developments, and it may not address all the hazards involved. It is up to the design engineer to identify hazards, regardless or whether they are described in the standard, and to make sure that the issues are adequately addressed. For example, the American Society of Mechanical Engineers (ASME) Boiler and Pressure Vessel Code provides design rules for the pressure-containing components of a boiler, but not the burner system. The burner system is covered by various other codes and standards, depending on the type of fuel and style of burner. This leaves the mechanical interface between burner and boiler not covered by any code or standard. It so happened that an old but serviceable industrial boiler was upgraded by fitting a new and more efficient burner. The frame-mounted fan and ducting to the windbox were replaced by an in-line burner, fan, and motor unit, which was connected directly to the end of the boiler casing by means of a custom-made adapter ring. As there was no code requirement for how to make such a connection, it was simply done on-site by the boilermaker without any engineering review or inspection. Five years later the connection to the boiler casing gave way, allowing the burner to tilt and the flame to impinge directly on the wall of the first pass Morrison tube. This caused overheating of the steel and collapsing of the tube under normal internal steam pressure. The boiler exploded, killing several workers and injuring many more.

References cited in this section

23. "Guide to Managing Product Design," *BS 7000:1989,* British Standards Institution, 1989

28. T.A. Hunter, Designing to Codes and Standards, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 66–71

29. Systematic Approach to the Design of Technical Systems and Products (translation), *VDI Guideline 2221:1987* Dusseldorf: Verein Deutscher Ingenieure, 1987

30. J. Hebert, and S.C. Uzgiris, The Role of Safety Standards in the Design Process, *ASME 89-DE-3* (New York), American Society of Mechanical Engineers, 1989

31. M.A. Dilich and D.F. Rudny, Compliance with Safety Standards: A Necessary but Not Sufficient Condition, *ASME 89-DE-1* (New York), American Society of Mechanical Engineers, April 1989

Design Life-Cycle Issues

When reviewing the design process that led ultimately to a failed component, it is sometimes necessary to consider life-cycle issues such as the expected service conditions, design for maintenance, design life, and design for recycling or disposal. Many types of equipment are now used far beyond their original design life, and often there is no guidance on how to predict what failures might occur, when they might occur, or where they might occur. This has long been an issue with regard to aircraft, especially with their stressed lightweight alloy structures, but it is now becoming critical with less obvious items such as sectional cast iron boilers more than 30 years old. The inspection procedures mandated for such boilers did not envisage the need to check the wall thickness of the cast iron sections in order to determine the annual metal loss due to corrosion, and it is possible to get a catastrophic failure of the wall during operation even though the boiler has been routinely inspected according to the prevailing regulations.

Acknowledgments

The authors acknowledge and thank Robert Koutny, Triodyne Graphic Communications, for his help in preparing the figures.

References

1. C. Hales, *Managing Engineering Design,* Longman Scientific & Technical, 1993; C. Hales and S. Gooch, 2nd ed., Springer-Verlag London Limited, to be published 2003

2. J.R. Dixon, Overview of the Design Process, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 7–14

3. K.N. Otto and K.L. Wood, Conceptual and Configuration Design of Products and Assemblies, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 15–32

4. G. Pahl and W. Beitz, *Engineering Design,* K.M. Wallace, Ed., The Design Council, 1984

5. E. Frankenberger, P. Badke-Schaub, and H. Birkhofer, Ed., *Designers—The Key to Successful Product Development,* Springer-Verlag, 1998

6. V. Hubka, *Principles of Engineering Design,* Butterworth Scientific, 1982

7. K.M. Wallace, A Systematic Approach to Engineering Design, *Design Management: A Handbook of Issues and Methods,* M. Oakley, Ed., Basil Blackwell Ltd., 1990

8. M.M. Andreasen and L. Hein, *Integrated Product Development,* IFS (Publications) Ltd. and Springer-Verlag, 1987

9. P.G. Smith, Cross-Functional Design Teams, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 49–53

10. S. Pugh, *Total Design—Integrated Methods for Successful Product Engineering,* Addison-Wesley Publishing Company, 1990

11. M. Khan and D.G. Smith, Overcoming Conceptual Barriers—by Systematic Design, C377/079, WDK-18, *Proc. ICED-89,* I.Mech.E., 1989

12. B.L. Tuttle, Creative Concept Development, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 39–48

13. R.L. Barnett and W.G. Switalski, Principles of Human Safety, Safety Brief, Vol 5 (No. 1), Triodyne Inc., 1988

14. C.O. Smith, Safety in Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 139–145

15. G. Kardos, Risk and Hazard Analysis in Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 117–125

16. T. Willis, M.P. Kaplan, and M.B. Kane, Safety in Design: an American Experience, C377/110, *Proc. of the Institution of Mechanical Engineers International Conference on Engineering Design (ICED-89),* Institution of Mechanical Engineers, 1989

17. C. Hales, Legal Threats to Innovation in Design, *Proc. ICED 99: International Conference on Engineering Design,* Institution of Mechanical Engineers, 1999

18. P.W. Huber, *Liability: The Legal Revolution and Its Consequences,* Basic Books, Inc., 1988

19. G.E. Dieter (chairperson), Overview of the Materials Selection Process, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 241–328

20. M.F. Ashby, *Materials Selection in Mechanical Design,* 2nd ed., Butterworth-Heinmann, 1999

21. B.A. Miller, Materials Selection for Failure Prevention, *Failure Analysis,* Vol 11, *ASM Handbook,* ASM International, 2002

22. M.F. Ashby and D. Cebon, Cambridge Engineering Selector, CES Software Version 3.2, Granta Design Limited (www.grantadesign.com), 2001

23. "Guide to Managing Product Design," *BS 7000:1989,* British Standards Institution, 1989

24. R. Stetter and D.G. Ullman, Team-Roles in Mechanical Design, *8th International Conference on Design Theory and Methodology, 96-DETC/DTM-1508,* J. Cagan and K.L. Wood, Ed., American Society of Mechanical Engineers, 1996, p 1–8

25. E.E. Rothschild, *Product Development Management,* T. Wilson Publishing Company, 1987

26. C. Hales, Analysis of an Engineering Design—the Space Shuttle Challenger, *Engineering Design and Manufacturing Management,* A.E. Samuel, Ed., Elsevier, 1989

27. G. Vrsek, Documenting and Communicating the Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 222–230

28. T.A. Hunter, Designing to Codes and Standards, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 66–71

29. Systematic Approach to the Design of Technical Systems and Products (translation), *VDI Guideline 2221:1987* Dusseldorf: Verein Deutscher Ingenieure, 1987

30. J. Hebert, and S.C. Uzgiris, The Role of Safety Standards in the Design Process, *ASME 89-DE-3* (New York), American Society of Mechanical Engineers, 1989

31. M.A. Dilich and D.F. Rudny, Compliance with Safety Standards: A Necessary but Not Sufficient Condition, *ASME 89-DE-1* (New York), American Society of Mechanical Engineers, April 1989

Failure Modes and Effects Analysis

John B. Bowles, Computer Science and Engineering, University of South Carolina

Introduction

FAILURE MODES AND EFFECTS ANALYSIS (FMEA) has evolved into a powerful tool that can be used by design engineers during all phases of product development to enhance product safety and reliability by eliminating or mitigating the potential effects of item failures. FMEA consists of examining the modes and causes of item failures and determining the product response to the failures. Steps can then be taken to change the design in order to eliminate the failure, mitigate its effects, or develop compensating provisions in case the failure should occur. A structured approach to the FMEA ensures that all appropriate failure modes are analyzed, that the system satisfies its fault mitigation requirements, and that these requirements are properly allocated. The FMEA methodology can be usefully employed throughout the design cycle from conceptual design to production and deployment. Tools have been developed to reduce the amount of labor required for the analysis and to evaluate hardware-, software-, material-, and process-related causes of failure. Significant progress is also being made in automated tools to facilitate the analysis.

This article describes the methodology for performing an FMEA. The overview section describes the process with the specific example of a hot water heater, followed by a discussion of the role of FMEA in the design process. The second section describes the analysis procedures and shows how proper planning, along with functional, interface, and detailed fault analyses, makes FMEA a process that can contribute to the design throughout the product development cycle. The third section describes the use of fault equivalence to reduce the amount of labor required by the analysis. The next section shows how fault trees are used to unify the analysis of failure modes caused by design errors, manufacturing and maintenance processes, materials, and so on, and to assess the probability of the failure mode occurring. The last section describes some of the approaches to automating some of the analysis.

Overview of FMEA

As the name suggests, FMEA is a procedure that examines each item in a system, considers how that item can fail, and then determines how that failure will affect the operation of the system. It is a structured, logical, and systematic analysis. Identifying possible component failure modes and determining their effects on the system operation helps the analyst to develop a deeper understanding of the relationships among the system components and, ultimately, to improve the system design by making changes to either eliminate or mitigate the undesirable effects of a failure.

Although designers have always had to be concerned with the possible effects of item failures, FMEA developed as a formal methodology during the 1950s at Grumman Aircraft Engineering Corporation, where it was used to analyze the safety of flight control systems for naval aircraft (Ref 1, 2). The first article to describe the FMEA process in detail was published in 1960 by Lomas, who was an engineer employed by the U.S. Navy (Ref 2), but the procedure he described was similar to one required by Mil-F-18372 (Aer.) (Ref 3), so some significant earlier work must have been done. Another early description of a procedure for performing a “failure mode effects analysis” was given by Coutinho (also at Grumman Aircraft Engineering Corporation) at the New York Academy of Sciences in 1964 (Ref 4). These first reports described a quite modern approach to doing a FMEA. Lomas included the probability of the failure mode occurrence in the analysis, and he introduced the tabular FMEA worksheet, which initiated the paradigm of doing the analysis by “filling out the form.” Coutinho recommended that FMEA be used as part of a formal design review by nondesigners of the system; he also classified the failures in terms of their consequences. During the 1970s and '80s, various military and professional society standards were written to define the analysis methodology (Ref 5, 6, 7, 8). MIL-STD 1629 (ships), “Procedures For Performing a Failure Mode Effects and Criticality Analysis” (Ref 9) was published in 1974, and, through several revisions, became the basic approach for analyzing a system.

Initially, FMEAs were used primarily as a safety analysis on the system hardware after the design was nearly complete. This application meant that any problems uncovered by the analysis were likely to be extremely expensive to fix. Recent refinements in the methodology have expanded the types of failures that can be analyzed to include functional failures in a functional representation of the system (Ref 10, 11, 12), failures of software components (Ref 11, 13, 14), and failures in the processes through which a product is built or maintained (Ref 15, 16). Along with these extensions in the

methodology, tools have also been developed to reduce the amount of labor required for the analysis (Ref 17), analyze the cause of the failure (Ref 18), and even automate much of the work required for the analysis (Ref 19, 20, 21, 22). Thus, over the last half-century, FMEA has become a "traditional" reliability analysis technique and has evolved into a highly effective tool that can be used throughout the product development process to improve the design of the system to which it is applied.

The term *FMEA* is often used almost interchangeably with the term *FMECA*—failure modes, effects, and criticality analysis. Some authors (Ref 1) take the view that FMEA is limited to an analysis of the effects of item failure modes, while FMECA includes the reliability function of assessing the probability of the occurrence of a failure mode along with its effect on the system. Others (Ref 23, 24) take the view that FMECA extends the analysis to include a ranking of the failure modes based on both a combination of the failure mode's probability of occurrence and the severity of its effect. Still others use the term *FMEA* even when a ranking is included (Ref 15, 16). This article adopts the terminology in Ref 11 and uses *FMEA* as the general term to include assessing the failure mode probability of occurrence under its scope. In the case of software- and process-related failure modes, it often is not possible to determine a meaningful probability of occurrence. This article does not include ranking procedures.

FMEA Example: Analysis of a Domestic Hot Water Heater. Figure 1 shows a schematic of a gas hot water heater. Functionally, the hot water heater takes cold water and gas as inputs and produces hot water as an output; flue gases and heat leakage are also produced as waste outputs. The water temperature is regulated by the controller opening and closing the main gas valve (labeled stop valve) when the temperature of the water in the tank goes outside the preset limits of 60 to 82 °C (140 to 180 °F). The pilot light is always on and the gas valve operates the main burner in full-on/full-off modes. The controller is operated by the temperature measuring and comparing device. The check valve in the water inlet pipe prevents reverse flow due to overpressure in the hot water system, and the relief valve opens if the system pressure exceeds 100 psig.

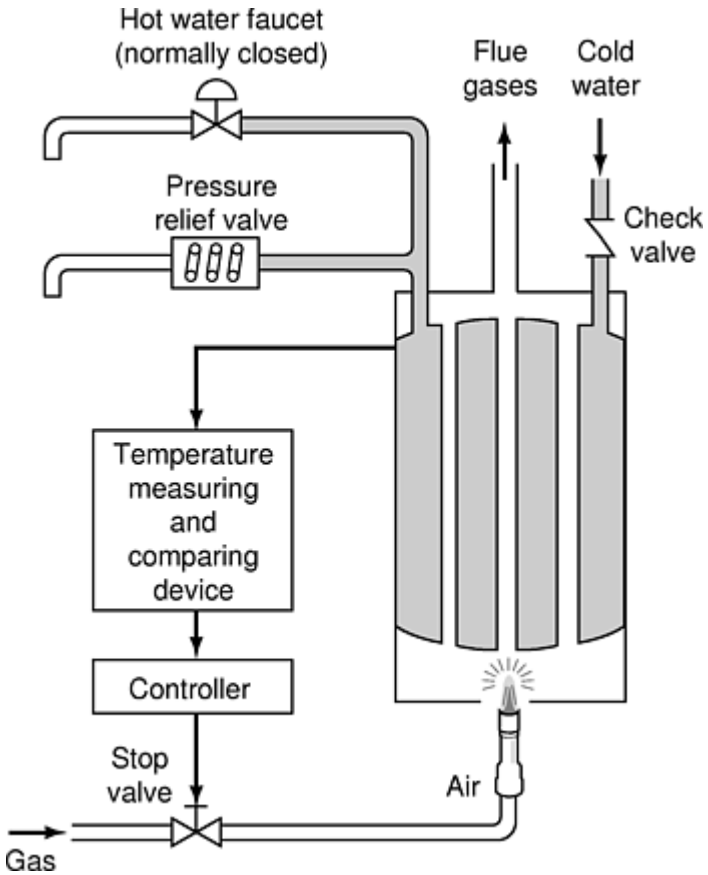


Fig. 1 Schematic for a gas hot water heater (Ref 25)

Several models provide useful representations of the system for analysis. First, the elements of the hot water heater can be represented hierarchically as shown in Fig. 2. This type of hierarchical model reflects the way in which the system design develops by first defining the system functions, then subdividing the system into smaller subsystems. The levels in the model are often called *levels of indenture*—a result of documentation techniques in which the descriptions of subsystem functions were indented relative to the description of the subsystem of which they were a part. Numbering conventions such as 1.23.12, which identify the system (1), subsystem (23), and so on, are often used to identify the system components.

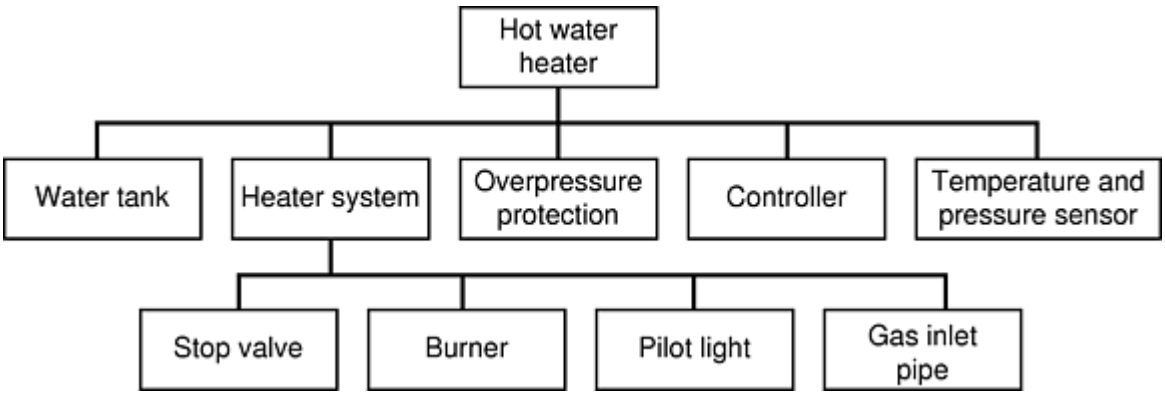


Fig. 2 Hierarchical representation of the hot water heater and its subsystems

The representation in Fig. 2 shows five subsystems, each of which can be further subdivided. For example, the heater subsystem can be subdivided into the stop valve, burner, pilot light, and gas inlet pipe.
The functions of the components shown in Fig 1 are:

Component	**Function**
Stop valve	Controls gas flow; full-on/full-off (controlled by the controller)
Controller	Opens and closes stop valve (responds to temperature sensor)
Temperature/pressure sensor	Senses water temperature and pressure
Check valve	Prevents reverse flow if overpressure
Relief valve	Opens when pressure >100 psig
Pilot light	Lights burner (always on)
Burner	Heats water (operated by stop valve)
Tank	Holds water (safe up to 100 psig)
Faucet	Releases water when needed

The functional relationships between the different system components are shown in a functional block diagram. Figure 3 shows a functional model for the hot water heater. The heater system, water tank, and overpressure protection form the basic system for producing the hot water; the temperature and pressure sensor and the controller provide a feedback loop for monitoring the production of the hot water and regulating its temperature. The heater system can be decomposed into the stop valve, gas inlet pipe, pilot light, and burner.

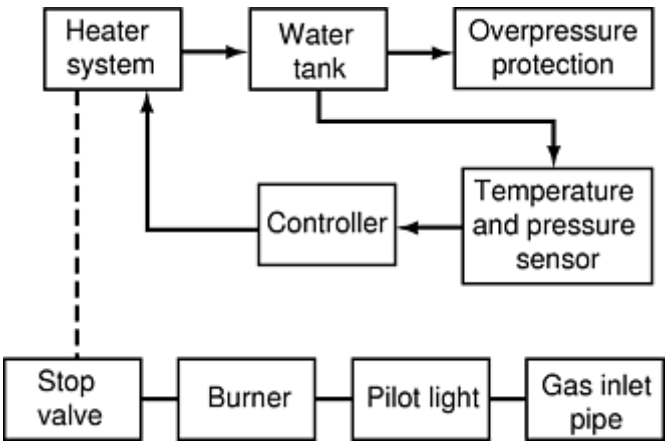


Fig. 3 Functional block diagram of hot water heater

Finally, from a reliability point of view, the hot water heater is a series system in which all components must operate properly for the system to operate. This is illustrated by the series reliability logic diagram in Fig. 4. Again, each system component can be further decomposed into its constituent subcomponents.

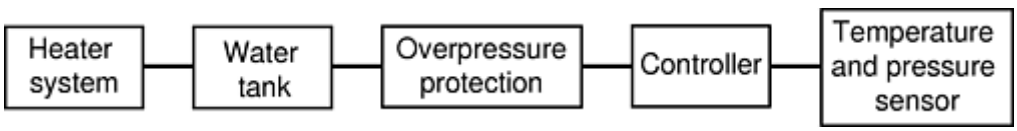


Fig. 4 Reliability logic diagram for hot water heater

The hierarchical decomposition of the system, the functional block diagram, and the reliability logic diagram help the analyst to understand the relationships between the system components. The next step is to identify all the ways in which each item can fail and the effect that each of those failures will have on the system. Effects are determined at each level of the system hierarchy—the effect on the module containing the failed component (local), the effect on every subsystem of which the component is a part, and the effect on the system. This analysis is illustrated by considering the stop valve. Table 1 illustrates part of the FMEA for the hot water heater showing the component failure modes and their effects at the local and system levels. Degraded operation (leaks and partial opening or closing of the stop valve), rather than outright failure, is one of the most common types of component failure modes.

Table 1 FMEA analysis of the stop valve for a hot water heater

Failure mode	**Effect**	
	Local	**System**
1. Fails closed	Burner off	No hot water
2. Fails open	Burner will not shut off	Overheats, release valve releases pressure, may get scalded
3. Does not open fully	Burner not fully on	Water heats slowly or does not reach desired temperature
4. Does not respond to controller—stays open	(Same as 2)	(Same as 2)
5. Does not respond to controller—stays closed	(Same as 1)	(Same as 1)
6. Gas leaks through valve	Burner will not shut off; burns at low level	Water overheats (possibly)
7. Gas leaks around valve	Gas leaks into room	Possible explosion, fire, or gas asphyxiation

Source: Ref 23

Generally, the analyst also identifies the possible causes of the failure and attempts to either eliminate them from the design or, if that is not possible, mitigate their effects in some way. For example, the valve could fail to open or close properly due to corrosion. This in turn could be caused by electrolysis between different metals or by a reaction of the valve material with the gas. Careful consideration of the valve materials would eliminate this failure mode. (At least it would eliminate this cause of the failure mode; there could be other causes.)

An analysis of the pilot light would reveal the hazard if it fails (i.e., it is not lit) and the gas is on. This failure mode cannot readily be eliminated, but it can be mitigated by adding an interlock to prevent the gas valve from opening if the pilot light is not on. Many gas hot water heaters have this type of interlock. If the interlock fails, the burner will not be able to heat the water even if the pilot light is on.

Role of FMEA in the Design Process. Figure 5 shows a typical product development cycle, beginning with conceptual design and progressing to deployment in the field. As indicated in the figure, the activities that constitute the FMEA complement and add value at every stage of the development cycle. During the conceptual design and preliminary design phases, the FMEA serves primarily to verify the adequacy of the system requirements; during the detailed design phase it is used to verify compliance with the requirements. During the verification and validation phase, it helps to maintain the integrity of design changes. Finally, during the production, use, and support phase, it serves as a guide for collecting field failure data and for developing maintenance and troubleshooting procedures.

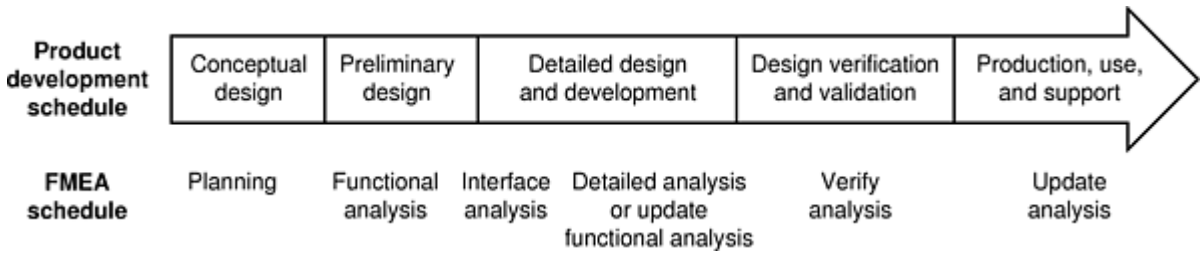


Fig. 5 Typical product development cycle and FMEA schedule (Ref 11)

As the design develops from concept to implementation, the system is subdivided into smaller subsystems. Once the subsystems are defined and their functions and interfaces are specified, the designer can focus on the design details of the subsystem, subdividing it into smaller subsystems until a solution is found. The conceptual subdividing of the system often parallels its subdivision into physical modules, and it is important that the subsystems be defined with a thorough knowledge of the technology (hardware, software, and materials) that will be used to construct each subsystem. The

FMEA helps to keep the designer, who is focusing on subsystem details, aware of the system-level effects of his or her design decisions. Like the design, the FMEA evolves from a functional to a detailed point of view (Fig. 6).

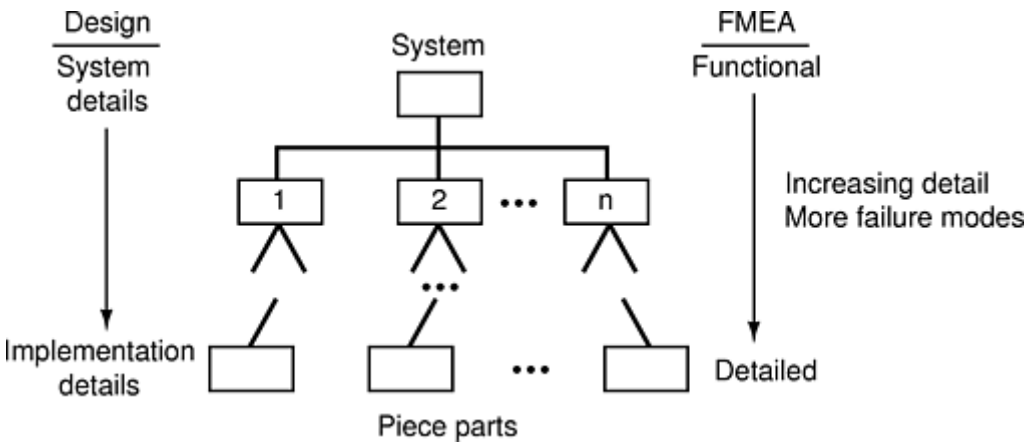


Fig. 6 FMEA tracks the design

Since the FMEA is concerned with the overall system behavior, it can serve as a unifying element for the different engineering groups involved with the product design. Such groups are often discipline specific and focus on their own areas of responsibility while ignoring other aspects of the design. For example, electrical engineers rarely address the design problems associated with vibration or heat transfer, and mechanical engineers sometimes forget effects such as cross talk in cabling. The FMEA provides a common communication tool for all the groups that must be involved with the design of the product: product designers, manufacturing engineers, test engineers, reliability and maintainability engineers, logistic support personnel, users, and others. It identifies potential single-point failure modes of system components and keeps critical items visible throughout the design process. It is useful for identifying the types of tests and testing environments needed to certify whether a design is suitable and as a basis for evaluating the adequacy of changes in product design, manufacturing process, or materials.

During the design process, many different users contribute to the FMEA, and it must meet their needs. The most important of these are discussed subsequently (Ref 23).

System Design. Design has overall responsibility for engineering, analysis, and the detailed design of the product. Because the fundamental reason for doing a FMEA is to improve the product design, the system designer is the most important FMEA user. He or she uses the FMEA to ensure that the results of previous “lessons learned” in the areas of product failure modes, causes of failures, and their effects have been addressed and then designs the product to remove or compensate for any unacceptable failure modes.

The designer contributes to the FMEA by providing detailed product descriptions to the analysis. A meaningful FMEA requires a close working cooperation between the reliability engineer, who has overall responsibility for coordinating all FMEA data, and the system designer, who has the detailed product knowledge to analyze the design and the ability to make the design changes indicated by the analysis.

Reliability. Reliability engineering is responsible for providing analyses to assess the probability that the product will successfully perform its intended function and mission. This organization usually has responsibility for developing the FMEA and maintaining the FMEA database. This entails coordinating the input from design engineering with that from other reliability, maintainability, and support functions. Failure modes developed from the product design specification must be analyzed along with those that are not directly under the designer's control—for example: failures resulting from improper maintenance, operation outside specification limits, or deviations in the manufacturing process.

The completed FMEA is a basis for future design and cost trade-off studies. Documentation supporting changes in design should compare the potential failure modes and compensating provisions of the new design to those of the baseline design.

System Safety. System safety engineers analyze how a product can be safely operated and maintained in its intended operating environment. They develop system hazard analyses identifying catastrophic and severe hazards and the component failures that can cause them. The FMEA is a source for identifying failure modes and causes that can lead to these hazards, and quantifying the hazard probabilities within the failure categories.

Maintainability. Maintainability engineering works with design to ensure that the product can be maintained by the customer in an efficient and cost-effective manner. This function requires analysis of part removal, replacement, teardown and buildup of the product in order to determine the time to do the operation, the skill levels required, the type of support equipment needed, and the documentation required. Maintainability engineers use the FMEA to identify potential maintenance tasks for which detailed task analyses are then developed. Potential preventive maintenance tasks, inspection schedules, and part-replacement schedules, can also be derived from the lists of compensating provisions and estimates of times to failure.

Logistics Support. Logistics support is responsible for ensuring that the product can be supported efficiently by the end user. This includes spare parts provisioning, support equipment, technical publications, and in some cases the provisioning of maintenance facilities, training of maintenance personnel, and packaging, handling, storage, and transportation of the product. The FMEA data help to identify maintenance tasks and other support activities, such as spares provisioning, support-equipment needs, and necessary publications.

The scope of the failure modes required by the logistics support function is considerably more extensive than just those produced by design engineering—they include failure modes induced as a result of the way the product is operated or maintained. (However, because the product design can strongly influence these types of failure modes, the product designer must also be made aware of them.)

Manufacturing. Manufacturing engineers design the sequence of operations through which the product will be produced. They also participate in defining the "critical item" list, considering items that have special producibility concerns. The FMEA provides manufacturing engineers with failure modes and effects which are used to refine the manufacturing processes to prevent manufacturing defects. In some cases, manufacturing may recommend design changes to reduce potential failure modes caused by manufacturing processes.

The manufacturing engineer identifies product-related, process failure modes, and assesses their system effects in what is often called a ***process FMEA.*** He or she also identifies both the manufacturing or assembly process causes of the failure mode and important process variables to monitor in order to reduce or detect the failure condition. Just as the design engineer has the detailed knowledge and ability to make ***design***-related changes to reduce or eliminate unacceptable failure modes found in the (product) FMEA, the manufacturing engineer has the detailed process knowledge and ability to make ***process*** changes, or impose process monitors, to eliminate unacceptable failure modes identified in the (process) FMEA.

Systems. Systems engineering uses the FMEA to facilitate trade-off studies involving the various disciplines that contribute to product development. In the early design phases, failure modes identified in the FMEA might require system-level design changes to control or mitigate their effects. For example, system-level changes in technology, such as using computer control instead of analog electronic or hydraulic control systems, might eliminate or reduce the system effects of some component failure modes; they may also introduce new failure modes, which must be analyzed. When systems built using "old technology" are upgraded, the potential failure modes and their effects identified in the FMEA are important for ensuring compatibility between the old system and its replacement.

Testability and Quality Assurance. The FMEA provides test engineers with information on anticipated failure modes and causes, against which to match fault detection and isolation capabilities. These are used to ensure that all failures can be detected, that appropriate tests are developed, and that built-in tests can detect and isolate all important failure modes. The FMEA can also provide information needed to develop both built-in and off-line diagnostic routines.

Project Management. The FMEA provides project managers with critical component lists and assurances that potential product reliability problems and safety risks have been identified and addressed in the product design. The FMEA data also indicate maintenance tasks that must be performed, and they substantiate the need for support equipment and other logistic support.

For some products, government agencies, such as the military and those involved in aviation and space exploration, are responsible for ensuring that users receive a system that is reliable, safe, cost-effective, and easy to maintain and use. FMEA is often one of the analyses required for readiness reviews and certification procedures to document that the designer has considered historical failure modes and causes, and has designed the product to minimize the effects of those failures.

In companies that use integrated product development teams to design their products, the design team generally includes members from most of these groups. In this case the FMEA facilitates communication among the team members and helps to ensure that a proper balance is maintained across all competing interests. The FMEA should include all the known component failure modes and their effects, including those related to the product's mechanical design, its electrical characteristics, its material composition, its production, its operation, and its support.

Throughout the design process, each failure mode is treated independently as its effects on the system reliability and safety are analyzed. Possible common mode failures in fault-tolerant systems should also be considered. Where redundant, or backup, subsystems have been proposed, the analysis should be broadened to include the failure conditions that result in the perceived need for such systems and how the failure condition will be detected, even with the redundancy or backup system. Possible design changes should be considered to eliminate or control the effects of all important failure modes. Single-point failures identified during the analysis should be uniquely identified so as to maintain their visibility throughout the design process.

References cited in this section

1. M. Campbell, "History and Background of FMECA," unpublished personal communication

2. W.R. Lomas, “Safety Considerations in the Design of Flight Control Systems for Navy Aircraft,” Paper 60-AV-34, American Society of Mechanical Engineers, 1960

3. Mil-F-18372 (Aer.), “General Specification for Design, Installations, and Test of Aircraft Flight Control Systems,” Bureau of Naval Weapons, Department of the Navy, Washington, D.C.

4. J.S. Coutinho, Failure-Effect Analysis, *Transactions New York Academy of Sciences,* Vol 26, 1964, p 564–585

5. “Fault/Failure Analysis Procedure,” ARP 926, Society of Automotive Engineers Aerospace Recommended Practice, 15 Sept 1967; ARP 926A, 15 Nov 1979

6. “Failure Mode and Effect Analyses,” Electronic Industries Association G-41 Committee on Reliability, Reliability Bulletin No. 9, Nov 1971

7. “Analysis Techniques for System Reliability—Procedure for Failure Mode and Effects Analysis (FMEA),” International Electrotechnical Commission, IEC Standard Publication 812, 1985

8. “Fault/Failure Analysis For Digital Systems and Equipment,” ARP 1834, Society of Automotive Engineers Aerospace Recommended Practice, Aug 1986

9. “Procedures for Performing a Failure Mode Effects and Criticality Analysis,” US MIL-STD-1629 (ships), 1 Nov 1974, US MIL-STD-1629A, 24 Nov 1980, US MIL-STD-1629A/Notice 2, 28 Nov 1984

10. C.S. Spangler, Systems Engineering—The Fault Analysis Process for Commercial Avionics Application, *Proc. of the Third Annual International Symposium of the National Council on Systems Engineering,* 1993, p 729–736

11. “Recommended Failure Mode and Effects Analysis (FMEA) Practices for Non-Automobile Applications,” ARP 5580, Society of Automotive Engineers Aerospace Recommended Practice, July 2001

12. J.B. Bowles, The New SAE FMECA Standard, *Proc. Annual Reliability and Maintainability Symposium,* 1998, p 48–53

13. P.L. Goddard, Validating the Safety of Real Time Control Systems Using FMEA, *Proc. Annual Reliability and Maintainability Symposium,* 1993, p 227–230

14. P.L. Goddard, Software FMEA Techniques, *Proc. Annual Reliability and Maintainability Symposium,* 2000, p 118–123

15. “Potential Failure Mode and Effects Analysis in Design (Design FMEA) and for Manufacturing and Assembly Processes (Process FMEA) Instruction Manual,” Ford Motor Company, Sept 1988

16. “Potential Failure Mode and Effects Analysis in Design (Design FMEA) and Potential Failure Mode and Effects Analysis in Manufacturing and Assembly Processes (Process FMEA) Reference Manual,” J1739, Society of Automotive Engineers, Surface Vehicle Recommended Practice, July 1994

17. C.S. Spangler, Equivalence Relations within the Failure Mode and Effects Analysis, *Proc. Annual Reliability and Maintainability Symposium,* 1999, p 352–357

18. M. Krasich, Use of Fault Tree Analysis for Evaluation of System-Reliability Improvements in Design Phase, *Proc. Annual Reliability and Maintainability Symposium,* 2000, p 1–7

19. T.A. Montgomery, D.R. Pugh, S.T. Leedham, and S.R. Twitchett, FMEA Automation for the Complete Design Process, *Proc. Annual Reliability and Maintainability Symposium,* 1996, p 30–36

20. C.J. Price, Effortless Incremental Design FMEA, *Proc. Annual Reliability and Maintainability Symposium,* 1996, p 30–36

21. C.J. Price, D.R. Pugh, M.S. Wilson, and N. Snooke, The Flame System: Automating Electrical Failure Mode and Effects Analysis (FMEA), *Proc. Annual Reliability and Maintainability Symposium,* 1995, p 90–95

22. D.J. Russomanno, R.D. Bonnell, and J.B. Bowles, Viewing Computer-Aided Failure Modes and Effects Analysis from an Artificial Intelligence Perspective, *Integrated Computer-Aided Engineering,* Vol. 1 (No. 3), Wiley, 1994, p 209–228

23. J.B. Bowles and R.D. Bonnell, Failure Mode, Effects, and Criticality Analysis (What It Is and How to Use It), *Topics in Reliability and Maintainability and Statistics, Annual Reliability and Maintainability Symposium tutorial notes,* 1999

24. "The FMECA Process in the Concurrent Engineering (CE) Environment," AIR 4845, Society of Automotive Engineers Aerospace Information Report, 18 June, 1993

25. H.E. Lambert, Report No. UCID-16238 Lawrence Livermore Laboratory, May 1973

The FMEA Process

The FMEA methodology is based on a hierarchical, inductive approach to analysis; the analyst must determine how every possible failure mode of every system component affects the system operation. The procedure consists of:

1. Identifying all item failure modes
2. Determining the effect of the failure for each failure mode, both locally and on the overall system being analyzed
3. Classifying the failure by its effects on the system operation and mission
4. Determining the failure's probability of occurrence
5. Identifying how the failure mode can be detected. (This is especially important for fault tolerant configurations.)
6. Identifying any compensating provisions or design changes to mitigate the failure effects.

The details of the FMEA analysis are captured on analysis worksheets. These worksheets provide a description of the failure modes and their consequences traceable to diagrams or other design documentation. Generally they include:

- Identification of the component being analyzed
- Its purpose or function
- The component failure mode
- The cause of the failure and how the failure is detected
- The local, subsystem, and system-level effects of the failure mode
- The severity classification and probability of occurrence of the failure mode

A FMEA normally analyzes each item failure as if it were the only failure within the system. When the failure is undetectable or latent or the item is redundant, the analysis may be extended to determine the effects of another failure, which in combination with the first failure could result in an undesirable condition. All single-point failures identified during the analysis that have undesirable consequences must be identified on the FMEA worksheets for proper disposition.

When analyzing failure effects, the analyst must also be concerned about possible failure cascades and common cause failures where a single event can lead to multiple failures. Such failures often result from the physical placement of the components rather than their operational functions. For example, a failure of a disk turbine in which the disk disintegrates and throws off pieces of broken metal could disable several independent hydraulic systems in an aircraft if they are all routed near the turbine.

In Fig. 5, the functional, interface, and detailed analyses provide tools for evaluating the design at each phase of the development cycle. The analysis iterates as the design evolves and expands to include more failure modes as more design details become available, until all the required equipment elements have been completely

defined, analyzed, and documented. Conducting the analysis in this manner enforces a disciplined review of the baseline design and allows timely feedback to the design process.

FMEA Planning. Careful planning for the FMEA tailors the scope of the analysis to the needs of the program and provides a process that efficiently identifies design deficiencies so that corrective actions or compensating provisions can be made in a timely manner. Proper planning requires that the system requirements—including the operating modes and functions of the system, required performance levels, environmental considerations, and safety or regulatory requirements—all be specified. During the planning process, data such as field reports, design rules, checklists, and other guidelines based on lessons learned, technology advances, and the history or analysis of similar systems are collected and studied. Models (often in the form of block or flow diagrams as in Fig. 2 through 4) are developed to illustrate the physical and functional relationships between system components and the interfaces within the system. These models are especially helpful for:

- Identifying the material and component technologies that are being proposed
- Identifying their characteristic failure modes
- Examining the effects of those types of failures on the system safety and operation
- Identifying potential compensating provisions in the design

The failures can then be assessed according to their effects on the operation and mission of the system, how the failures will be detected, and any compensating provisions or design changes needed to mitigate the effects of the failures.

The complexity of the analysis tasks makes it essential to establish a set of ground rules for the analysis as early as possible. These rules help to ensure the completeness, correctness, and consistency of the analysis. Ground rules identify assumptions, limitations, analysis approach, boundary conditions, failure criteria for fault models, and what constitutes a failure (in terms of performance criteria, success/failure criteria, or interface factors) (Ref 11). They also identify requirements to be verified, possible end-item support equipment (e.g., operational or ground support, maintenance support, special test equipment, etc.), lowest indenture level to be analyzed, assumed environmental conditions, possible mission objectives and modes of operation, risk factors defined by system safety analyses, and so on. Some of the ground rules for the analysis of the hot water heater included the assumption of a continuous supply of gas and water; that the water heater was intended for home use, and that it operated between 60 and 80°C (140 and 180°F).

It is also useful to define libraries with descriptions of failure modes and consequences. Such libraries help control the analysis process and ensure consistency in terminology, types of failure modes considered, and so on among all the analysts (including future analysts) contributing to the project. They also provide direction as to the level of detail for the analysis while ensuring a more consistent and uniform documentation. The following libraries should be developed for the FMEA:

- Functional, interface, and detailed failure modes for each item type
- Mission phases and operating modes
- Effects that each failure mode has on the overall system and on the next-higher indenture level above the postulated failure mode
- Descriptions with which to classify the severity of each failure mode's effect on the end-item
- Monitor descriptions that identify how a failure mode is detected

Functional Fault Analysis. A functional fault analysis is performed on the conceptual design to verify that provisions to compensate for component failures are both necessary and sufficient. The relief valve in the hot water heater is an example of such a compensating provision; it is intended to relieve excessive pressure that could potentially rupture the tank if the water overheats. A functional FMEA of the hot water heater would show that it is needed and that it provides the required protection. The relief valve also illustrates the concern about undetected failures: How does one know that the valve is operating properly? Thus, in addition to showing that the valve is needed, the functional FMEA might also result in a requirement for periodic inspections of the relief valve or a requirement for a monitor to detect a failure of the valve.

A functional analysis begins with a functional block diagram or equivalent system representation. The block diagram indicates the input/output transfer function, the flow of information, energy, force, fluid, and so on within the system, and the primary relationship between the items to be covered in the analysis. Functional

failure mode models are assigned to each block resulting in the list of postulated failure modes to be analyzed. Then each function is analytically failed in each of its failure modes to determine the effects and characteristic indications of failure mode in each applicable operating mode.

Typical functional failure modes are of the form “function fails to perform,” “function is continuously performed,” or “function is performed at the wrong time.” In the hot water heater example, the function of the stop valve is to control the flow of gas in full-on/full-off mode. Its functional failure modes are a failure to control the flow of the gas in that:

- Gas is on when it should be off.
- Gas is off when it should be on.
- Gas is not full-on or full-off.

Ideally, a functional fault analysis focuses on the functions that an item or group of items perform rather than the characteristics of the specific components used in their implementation. In practice, the types of failure modes considered for a function may depend on how the function will be implemented. When a functional analysis is applied to manufacturing processes, typical failure mode categories include manufacturing and assembly operations, receiving inspection, and testing. Process failure modes are described by process characteristics that can be corrected. For example, part misorientation, part hole off-center, binding, cracked, and so on. When the analysis is applied to software, typical functional failure modes are: (a) failure to execute; (b) incomplete execution; (c) execution at an incorrect time (early, late, or when it should not have been executed); (d) incorrect result. For some software, the effects of other failure modes may also have to be assessed. For example, the analysis of a real-time system may require an assessment of interrupt timing and priority assignments.

As the design details are developed, the functional block diagrams and analyses are expanded, and the analysis iterates until all the system elements have been completely defined and documented. Any undetected failures that cause loss of system-level functions are corrected by incorporating requirements for compensating provisions into the design and revising the functional fault analysis to reflect the modifications.

A major benefit of a functional FMEA is that the functional failure modes can be identified in the conceptual design before the detailed design has been developed. Thus the analysis is aimed at influencing the design before the construction of any hardware. Typical results of the analysis identify functional failure modes that need to be eliminated or mitigated by changing the functional design of the system (Ref 10, 11).

Interface Fault Analysis. The interface fault analysis focuses on determining the characteristics of failures in the interconnections between subsystem elements. Cables, plumbing, fiber-optic links, mechanical linkages, and other interconnections between subsystem modules provide the basis for the postulated failure modes. Each type of interconnection has its own set of potential failure modes.

The interface fault analysis begins by defining the specific failure modes of the interfaces between subsystem elements. Typical electrical failure modes are “signal fails in the open condition,” “signal fails in the short condition,” and “input or output shorted to ground.” Typical mechanical failure modes are “piping fails in the closed position,” and “hydraulic pressure low.” Software interface failure modes focus on failures affecting the interfaces between disparate software and hardware elements. The four failure modes most often applied to software interfaces are: (a) failure to update an interface value; (b) incomplete update of the interface value; (c) update to interface value occurs at an incorrect time (early or late); and (d) error in the values or message provided at the software interface. Other failure modes specific to the software or the interface hardware may also need to be considered. Process errors that could result in misalignment or improper connection of parts are an important failure mode for a process interface FMEA. These types of errors are often eliminated by designing interconnections that can be made in only one way or ones for which any orientation is valid.

Because the interface analysis involves the interfaces between subsystem elements, it is the responsibility of the system integrator to ensure that the analysis is complete. The subsystem designer is responsible for assessing the effects of all inputs to the subsystem. The integrator uses the results of these analyses to determine the effects of the interface failure modes on the subsystem and system.

The advantage of a separate interface fault analysis is that it can be performed before detailed module designs are available; it can begin as soon as the subsystem inputs, outputs, and their interconnections are defined. Typical results of this analysis are interface failure modes that need to be eliminated or mitigated by interface design changes.

Detailed Fault Analysis. A detailed fault analysis is used to verify that the design complies with the system requirements for: (a) failures that can cause the loss of system functions; (b) single-point failures; (c) fault detection capabilities; and (d) fault isolation. It uses component failure modes postulated from the individual components in the detailed design. This includes the physical devices in the design, software modules, and the processing steps to produce the item.

A detailed fault analysis on the system hardware has traditionally been done in a FMEA. Table 1 is an example of such an analysis. This type of analysis is sometimes called a *piece-part fault analysis* because it is done on the "piece parts" that compose the system. To perform the analysis, an established set of component failure modes and their corresponding occurrence ratios are especially useful. For example, failure modes normally considered for a capacitor are "open," "short," and "leaking"; for an integrated circuit they are "output pin stuck high" and "output pin stuck low." For a bearing, they are "binding or sticking," "excessive play," and "contaminated." The failure mode ratios allow the item failure probability to be apportioned among its failure modes to give the failure mode probability of occurrence. Failure mode ratios are best obtained from field data that are representative of the particular item application but when such data are not available, generic references such as *Failure Mode/Mechanism Distributions 1997 (FMD 97)* (Ref 26) can be used for guidance. Failure mode ratios for a particular component type may vary depending on the operating environment, manufacturer, application, and other factors.

A detailed software fault analysis is intended to assess the effect of failures in the software functionality or in the variables used in the software. Unlike hardware, the software errors found by the FMEA are not due to potential failure modes; rather, they already exist and will be activated under the "right" set of conditions. A software fault analysis is applied to the as-implemented code. The failure modes used in the analysis include errors in the code that implements the software function, such as algorithm singularities and failure modes for each software variable. The variable failure modes must form a logically complete set of the possible error states for the variable type. For example, a Boolean variable with validity flag would have the following failure modes: "value true when it should be false and validity flag set to valid," "value false when it should be true and validity flag set to valid," and "value is correct but validity flag is set to invalid." Extensive lists of failure modes for many common variable types are included in Ref 11, 13, and 14. Due to the usually large number of variables and possible states, and the associated cost of performing a detailed FMEA on even a small piece of software, the analysis is most usefully applied to small embedded systems that have little or no memory protection and computational error checking provided by the hardware.

Process-related failure modes are specific to the manufacturing or maintenance process. For example, a wax coating may be applied too thinly to provide the corrosion protection it is intended to provide.

One problem associated with a detailed fault analysis is that the level of detail required to do the analysis means that it cannot be initiated until the design has matured to the point that detailed schematics and parts lists are available. This means that any major errors found by the analysis are likely to be very expensive to fix. Conversely, even major errors in the design concept are relatively easy to fix in the early design phase when the functional and interface fault analyses are done.

Identify Failure Consequences. The consequences of a failure mode analyzed in a FMEA are its effects, a classification of the severity of the failure mode based on its system-level effects, and the probability of the failure mode occurrence. The analysis is conducted for all phases and modes of system operation including normal operating modes, contingency modes, and test modes, and with respect to the primary and secondary mission objectives. The local, next-higher, and end-level failure effects of each item failure mode must be determined, and corrective actions or compensating provisions must be identified within each applicable operating mode.

Assessment of the failure mode effects must identify the system conditions or operational modes that manifest the anomalous behavior. For example, failures in the landing gear of an airplane that would cause "loss of landing gear extension" will have significantly different effects if the failure occurs on the ground than if it occurs while attempting to land. Likewise, the discovery mechanism for detecting the failure may be different.

The analysis identifies the effects of each postulated failure mode in a bottom-up manner, beginning with the lowest-level items identified. The effects of each failure mode are evaluated with respect to the function of the item being analyzed. Because the item failure under consideration might impact the system at several levels of indenture, the failure effects are then related to the functions at the next-higher indenture level of the design, continuing progressively to the top or system-level functions.

The local effect(s) description gives a detailed accounting of the impact the failure has on the local operation or function of the item being analyzed. The fault condition is described in sufficient detail that it can be used with the next-level effects, end-effects, and detecting monitor(s) to identify and isolate the faulty equipment, thus providing a basis for evaluating compensating provisions and recommending corrective actions.

Next-level effects describe the effect the failure has on the next-higher level operation, function, or status. Descriptions of the next-level effects are normally compiled in a table for consistency of annotation. The failure effect at one level of indenture is the item failure mode of the next higher level of which the item is a component.

End-effects describe the effect the failure has on the ability of the system to operate and properly complete its mission. End-effects also provide a "go/no-go" assessment of system capability to perform its intended mission. The system-level, failure effect descriptions are best derived from the system requirements and compiled in a table for consistency of annotation.

Failure modes are usually classified by an assessment of the significance of the end-effect on the system operation and mission. The FMEA ground rules should provide a ranking and classification system, and appropriate criteria for assessing the severity of failures for the product being analyzed. Often a four-level classification system developed for military equipment with severity classifications ranging from "catastrophic" to "minor" is used (Ref 9). The automobile industry generally uses a ten-level classification system (Ref 16).

Classifying a failure mode and ranking the consequences of failure require knowledge of the system and its phases of operation. For example, in some situations a failed tire might result in nothing more than the inconvenience of having to change the tire. In other cases, a failed tire could lead to loss of control of the vehicle and much more serious consequences.

When items are redundant and there is no warning that a redundant item has failed, the severity should be assessed as if all of the redundant items have failed.

Corrective Action Recommendations. Corrective actions are needed for undetectable faults and for faults having significant consequences—for example, unsafe conditions, mission- or safety-critical single-point failures, adverse effects on operating capability, or high maintenance costs. Corrective actions may not be needed if the risks for the specific consequence(s) of a failure are acceptable based on a low enough probability of occurrence. Corrective actions generally take the form of changes in requirements, design, processes, procedures, or materials to eliminate the design deficiency.

Development of an appropriate corrective action usually requires understanding and eliminating the cause of the specific failure mode; conversely, careful analysis of the failure mode causes may suggest ways to eliminate the failure. Some examples of failure causes are:

- Incorrect material specification
- Overstressing of a component
- Insufficient lubrication
- Inadequate maintenance instructions
- Poor protection from the environment
- Incorrect algorithm (software)
- Software design errors, including software requirements errors

Special attention to the failure mode causes may be needed to ensure that proper materials are used when the operational environment is especially severe due to effects such as extreme temperature cycling, very high or very low operating temperatures, the presence of corrosive chemicals, and so on. Once a corrective action is implemented and validated, the affected fault analyses must be revised to reflect the new baseline configuration.

If a failure results in unsafe system operating conditions, warnings are necessary to alert the user and to ensure satisfactory system status before commencing operations. Monitors must be strategically located to cover all undetected catastrophic, hazardous, and single-point failures, based on the system requirements and intended uses. Once the necessary monitors have been identified, a subsequent FMEA iteration is conducted (in support of maintenance activities) to verify that any remaining undetected failure modes comply with the system fault detection requirements. Requirements for fault detection monitors are then derived to cover the remaining undetected failure modes. These monitoring requirements include operator procedures and human monitoring, as well as built-in test.

When design changes to correct a deficiency are not possible or feasible, compensating provisions must be identified to circumvent or mitigate the effect of the failure when it occurs. Such provisions are often in the form of design provisions or designated operator actions that allow continued safe operation when a failure occurs.

References cited in this section

9. "Procedures for Performing a Failure Mode Effects and Criticality Analysis," US MIL-STD-1629 (ships), 1 Nov 1974, US MIL-STD-1629A, 24 Nov 1980, US MIL-STD-1629A/Notice 2, 28 Nov 1984

10. C.S. Spangler, Systems Engineering—The Fault Analysis Process for Commercial Avionics Application, *Proc. of the Third Annual International Symposium of the National Council on Systems Engineering,* 1993, p 729–736

11. "Recommended Failure Mode and Effects Analysis (FMEA) Practices for Non-Automobile Applications," ARP 5580, Society of Automotive Engineers Aerospace Recommended Practice, July 2001

13. P.L. Goddard, Validating the Safety of Real Time Control Systems Using FMEA, *Proc. Annual Reliability and Maintainability Symposium,* 1993, p 227–230

14. P.L. Goddard, Software FMEA Techniques, *Proc. Annual Reliability and Maintainability Symposium,* 2000, p 118–123

16. "Potential Failure Mode and Effects Analysis in Design (Design FMEA) and Potential Failure Mode and Effects Analysis in Manufacturing and Assembly Processes (Process FMEA) Reference Manual," J1739, Society of Automotive Engineers, Surface Vehicle Recommended Practice, July 1994

26. "Failure Mode/Mechanism Distributions 1997," FMD-97, Reliability Analysis Center, IIT Research Institute, 1997

Fault Equivalence

Traditionally, FMEAs have been performed on a component-by-component basis using a tabular format much like that originally developed by Lomas. Each item is listed, followed by a list of its potential failure modes; then the consequences of each failure mode are determined by analytically simulating the operation of the system with the failed item. Worksheets, such as those in MIL-STD-1629 (Ref 9), SAE J1739 (Ref 16), and the one shown in Table 1, are organized to facilitate this type of analysis.

In a manual environment, this type of systematic analysis is important because it ensures that every potential failure mode of every component is analyzed. It also makes the analysis procedure quite laborious, leads to inconsistencies in describing identical failure effects, and has given FMEA its focus on "filling out the form," rather than focusing on what has been learned from the analysis.

When a database or other automated technology is used, much of the duplicative work associated with the FMEA can be eliminated by grouping the failure modes into equivalence groups consisting of all the failure modes that exhibit identical consequences (Ref 17). Such a group is illustrated in Fig. 7 where "no output at A," "open at B," and "open input at C" all have the same effect.

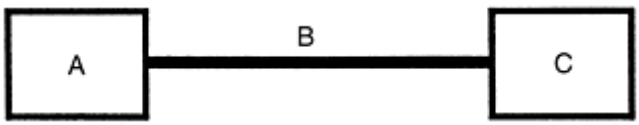


Fig. 7 Fault equivalent failure modes: A, output open; B, open; C, input open

The failure modes that exhibit identical consequences are called "fault equivalent failure modes" and grouped under a single identifying Fault (equivalent) Identifier Number (FIN).

The FMEA process starts with the functional fault analysis, progresses to the interface analysis, and concludes with the detailed analysis. Each analysis represents a more detailed iteration from the previous analyses. The common element

between the analysis types is the subsystem fault, which (along with its associated description of the local effect(s), next-level effect(s), end-effect(s), severity, compensating provisions, and detecting monitor(s)) is identified by the FIN. Because the FIN is generated on a functional basis, it allows previously generated fault information to be traced and used in subsequent analyses. For example, faults generated for the functional fault analysis are used during the interface failure mode analysis when the consequences of the interface failure modes are identical to those of the previously analyzed functional failure modes. In this case, the interface failure modes are assigned the FIN for the already recorded functional fault, thereby maintaining continuity. This process continues for the detailed analysis. Detailed failure modes that have identical consequences to previously analyzed functional and/or interface failure modes are assigned to the same equivalence group. Failure modes whose consequences are not identical to existing faults are assigned new FINs.

In the hot water heater example, the functional failure modes "gas is on when it should be off" and the piece part failure modes "stop valve fails open," and "stop valve does not respond to controller—stays open" all have identical consequences and hence would belong to the same fault equivalence group.

Organizing the analysis by fault equivalent groups allows the functional fault analysis to be integrated with the interface and piece part fault analyses. Identification of fault equivalence groups permits the analyst to manage failure consequences in place of individual failure modes. This reduces the magnitude of the analysis effort while improving the consistency of the results. Spangler cites one study in which 12,401 failure modes resulted in 1,759 equivalent fault conditions (Ref 17).

References cited in this section

9. "Procedures for Performing a Failure Mode Effects and Criticality Analysis," US MIL-STD-1629 (ships), 1 Nov 1974, US MIL-STD-1629A, 24 Nov 1980, US MIL-STD-1629A/Notice 2, 28 Nov 1984

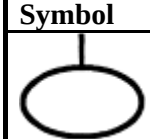
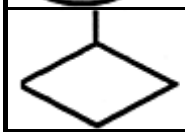
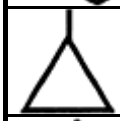
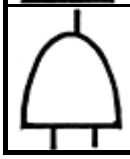
16. "Potential Failure Mode and Effects Analysis in Design (Design FMEA) and Potential Failure Mode and Effects Analysis in Manufacturing and Assembly Processes (Process FMEA) Reference Manual," J1739, Society of Automotive Engineers, Surface Vehicle Recommended Practice, July 1994

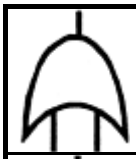
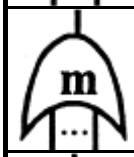
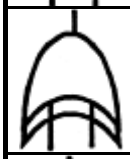
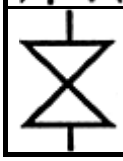
17. C.S. Spangler, Equivalence Relations within the Failure Mode and Effects Analysis, *Proc. Annual Reliability and Maintainability Symposium,* 1999, p 352–357

The Failure Cause Model

As noted in the hot water heater example, faults are postulated at a given level of the hierarchical system model, and their effects are propagated upward to ascertain the overall system effect of each item failure mode. A failure cause model is used to assess the causes of the failure mode and evaluate the failure mode probability of occurrence (Ref 18, 27). The failure cause model often takes the form of a fault tree whose top event is the failure mode of interest. The fault tree for a failure mode at a given level is built up from combinations of subsystem failures at levels lower than that at which the failure mode is postulated. If the failure mode can result from any of several lower-level events, it is represented logically as the OR of those events; if it can result only if all of several lower level events occur, it is the AND of those events. The fault tree is most often represented graphically using logic gates such as those shown in Table 2.

Table 2 Symbols used in the construction of a fault tree

Symbol	Name	Description	Reliability model
	Basic event	Event that is not further decomposed and for which reliability information is available	Component failure mode, or a failure mode cause
	Undeveloped event	A part of the system that has not yet been developed or defined	A contributor to the probability of failure but the structure of that system part has not been defined
	Transfer gate	A gate indicating that the corresponding part of the system fault tree is developed on another page or part of the diagram	A partial reliability block diagram is shown in another location of the overall system block diagram
	AND gate	The output event occurs only if all of the input events occur	Failure occurs if all of the parts of the system fail—redundant system

	OR gate	The output event occurs if any of its input events occur	Failure occurs if any of the parts of the system fail—series system
	Majority OR gate	The output event occurs if *m* of the input events occur	*k*-out-of-*n* module redundancy where $m = n - k + 1$.
	Exclusive OR gate	The output event occurs if one input event but not both input events takes place	Failure occurs only if one, but not both, of the two possible failures occurs
	NOT gate	The output event occurs only if the input events do not occur	Exclusive event or preventive measure does not take place.

From the analysis in Table 1 the probability of the stop valve not working is the logical OR of its failure modes. This is illustrated in Fig. 8, where one should observe that the fault trees for each failure mode are developed separately. The fault tree for the failure mode, "Valve does not respond to controller—stays open" is developed in Fig. 9. A similar fault tree can be developed for each of the other failure modes. Observe that this fault tree includes operational causes of failure such as debris getting in the actuator, manufacturing process failures such as a cold solder connection, and design errors such as using too brittle wire for the connector. Thus, the failure cause model for the failure mode unifies the various types of FMEA for product, process, software, and so on.

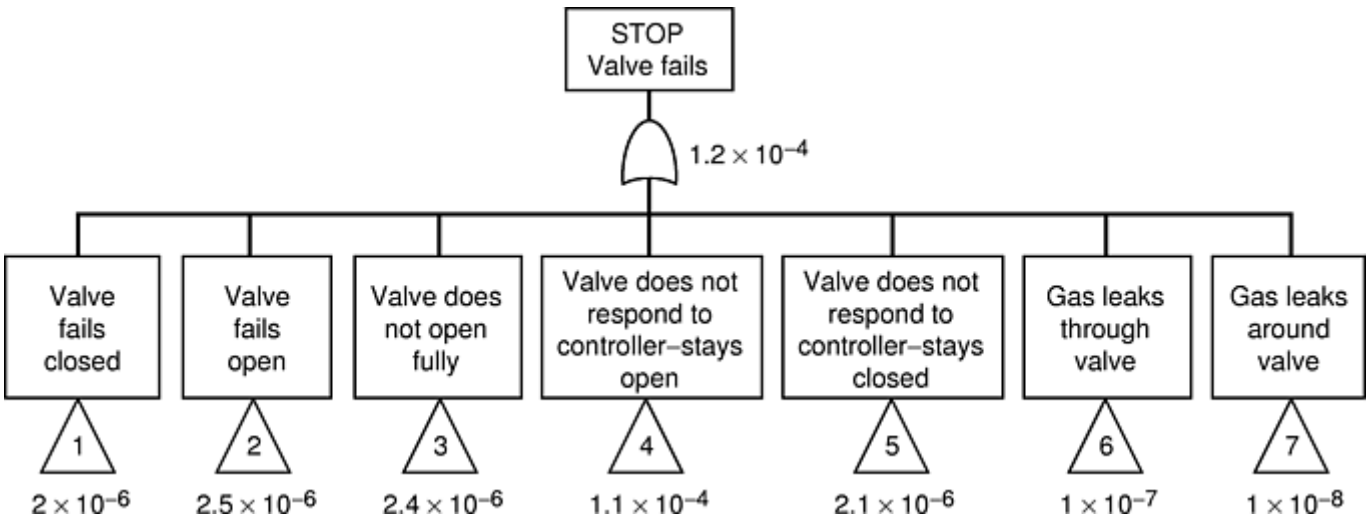


Fig. 8 Fault tree representation of the "stop valve fails" (Ref 27)

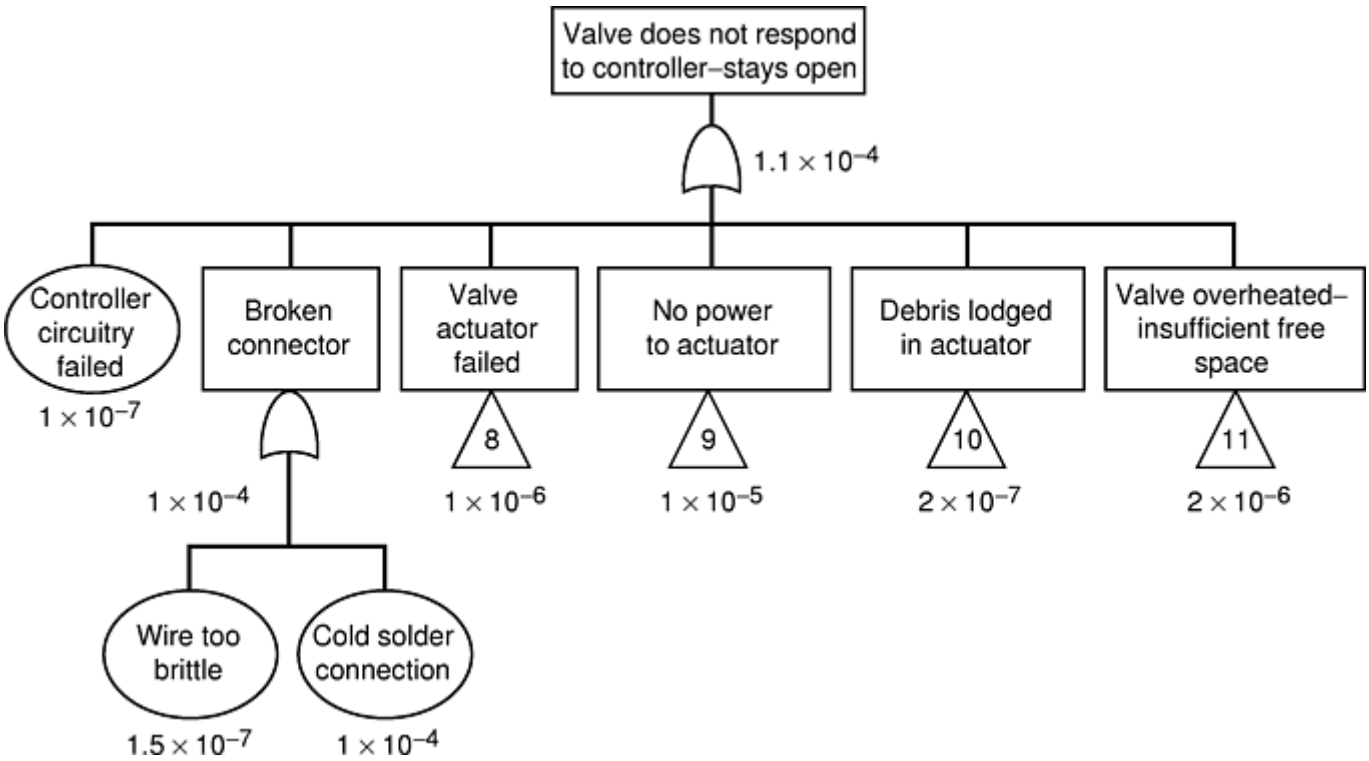


Fig. 9 Fault tree representation of the "valve does not respond to the controller—stays open" failure mode (Ref 27)

The probability of failure for each terminal failure mode can be assessed and then propagated up the fault tree to determine the failure mode probability of occurrence. The failure events in the fault tree are generally assumed to be mutually exclusive and the probabilities of the events at an OR gate are added to obtain the resultant failure probability. If the mutual exclusivity assumption is wrong, the sum must be either reduced by the joint probability or, more commonly, since joint failure probabilities are often difficult to quantify, accepted as an upper bound on the probability. Because the individual event probabilities are usually quite small, this is usually sufficient. If the fault tree includes "repeated events," a more complicated analysis using decomposition is necessary to ascertain the failure mode probability (Ref 28).

The assignment of failure mode probabilities of occurrence, based on the failure cause probabilities, is shown in Fig. 8 and 9. (The numerical values of the probabilities shown are arbitrary and for illustration only.) In Fig. 9 the top event probability of failure is the sum of the terminal event probabilities. From Fig. 9, the most likely cause of the failure mode "Valve does not respond to controller—stays open" is a cold solder connection. Thus, the manufacturing process should be examined, and appropriate controls to prevent this from occurring should be established.

References cited in this section

18. M. Krasich, Use of Fault Tree Analysis for Evaluation of System-Reliability Improvements in Design Phase, *Proc. Annual Reliability and Maintainability Symposium,* 2000, p 1–7

27. J.B. Bowles, Designing for Failure: Managing the Failure Response through Analysis, *Process of Failure Prevention through Education: Getting to the Root Cause,* J. Scutti, Ed., ASM International, 23–25 May 2000, Cleveland, p 62–69

28. D. Kececioglu, *Reliability Engineering Handbook,* Vol 2, Prentice-Hall, 1991

Automation

When done manually, FMEAs tend to be tedious and time-consuming tasks. They are error prone, and often are much disliked by product designers who generally focus on the tasks a system must do and how to design the system to do those tasks. Analyzing what will happen when something fails requires viewing the system in a different way, and propagating a failure mode from the component level to the system level is often difficult.

A number of computer-based tools and customized databases have been developed to aid the analysis process. The simplest, sometimes based on commercial spreadsheet software, do little more than provide a consistent format for the analysis. Others have features that help to improve the quality of the analysis. Some useful features of automated tools are:

- Supplying lists of failure mode and failure effect descriptions from which the analyst can choose. This feature encourages the use of common terminology and standardized descriptions in the FMEA report as well as relieving the tedium of writing such descriptions.
- Providing lists of appropriate failure modes and their apportionments for each type of component. This helps to ensure that all failure modes are analyzed.
- Making completeness and consistency checks such as: (a) checking that all appropriate failure modes for a component have been analyzed and that the sum of all failure mode apportionments is 1.0; (b) verifying that failure effects at each level have been given for each failure mode; and (c) checking terminology, grammar, spelling, and mathematical calculations in the report.
- Tailoring the FMEA report for each user's specific needs by sorting, filtering, and reformatting the data in the reports. Data can be sorted by item, failure mode, failure effect, severity class, probability of occurrence, and so on. Other useful features include providing sub-reports such as critical item lists and reformatting the FMEA so that its data can be automatically input to other databases.

The lists of failure mode and failure effect descriptions and the lists of failure modes and their apportionments for each type of item are often maintained in the analysis libraries discussed earlier, or in a database where they can be readily accessed by the FMEA analyst.

Other types of computer programs have been developed to help the engineer with other aspects of the FMEA analysis. In the later design stages, when detailed specifications are available, numerical simulation programs enable the designer to study the behavior and operational parameters of a system or subsystem. Simulation allows the designer to examine the effects of component tolerances on the system operation and, by changing component parameters to their failed state, to evaluate how various component failure modes will affect the system operation and output.

Researchers are also investigating how expert system technology can be used to help perform a FMEA. They hope that this research will yield tools that the designer can use early in the design process when errors and design weaknesses can be easily and cheaply corrected. Current efforts are focusing on the application of sophisticated database technology; the use of functional, causal, qualitative, and case-based reasoning methodologies; and the development of structural, behavioral, and functional system models. A good overview of this work is given in Ref 22 and 24.

The flame system, developed as a prototype application at the University of Wales at Aberystwyth and now marketed commercially as “AutoSteve,” is perhaps the most automated FMEA tool currently available (Ref 19, 20, 21). It is designed for use with automobile electrical systems. Flame uses both qualitative and quantitative device models for its simulation. This allows it to be used over most of the design cycle. Component models, including their possible faults, are stored in a component library, and new components can be added by the design engineer if they do not already exist. Descriptions of the functions performed by automobile circuits are also stored in a database, and the analyst need only link these descriptions to the particular states of the components in the circuit to build the model. Flame simulates the circuit with each failure, for each component, and generates the effect. The simulation is performed by following a list of events provided by the designer. Each event changes the state of the system; for example, the events “turn wiper on,” and “turn wiper to slow” will take the state of the wiper system from off, to intermittent, to slow. Flame analyzes the system by first executing each of the designated events with all components in their operational states. This provides a baseline history of expected system states. It then applies each failure mode to each of the components for each of the system states and compares the results to the correct working of the system. This comparison yields the effect of the failure mode on the system in terms of the functions that fail to operate, states of the system that are different, and components within the circuit that are in a different state than expected. The effects generator can then filter the effects and report only those of interest to the designer. Faults can be made to appear at different times in the simulation, such as during transient analysis, after establishing the direct current (dc) operating point, at a specified time, and so on. The system is also able to assign severity, detection, and occurrence values to each effect automatically (Ref 19, 21).

References cited in this section

19. T.A. Montgomery, D.R. Pugh, S.T. Leedham, and S.R. Twitchett, FMEA Automation for the Complete Design Process, *Proc. Annual Reliability and Maintainability Symposium,* 1996, p 30–36

20. C.J. Price, Effortless Incremental Design FMEA, *Proc. Annual Reliability and Maintainability Symposium,* 1996, p 30–36

21. C.J. Price, D.R. Pugh, M.S. Wilson, and N. Snooke, The Flame System: Automating Electrical Failure Mode and Effects Analysis (FMEA), *Proc. Annual Reliability and Maintainability Symposium,* 1995, p 90–95

22. D.J. Russomanno, R.D. Bonnell, and J.B. Bowles, Viewing Computer-Aided Failure Modes and Effects Analysis from an Artificial Intelligence Perspective, *Integrated Computer-Aided Engineering,* Vol. 1 (No. 3), Wiley, 1994, p 209–228

24. "The FMECA Process in the Concurrent Engineering (CE) Environment," AIR 4845, Society of Automotive Engineers Aerospace Information Report, 18 June, 1993

Conclusions

Failure Modes and Effects Analysis has evolved from an ad hoc technique, dependent on the designer's "experience," to a formal and accepted analysis technique. Failure modes can be described functionally, and the functional system model can be analyzed early in the design phase. Through fault equivalence, results from the functional and interface fault analyses can be used to reduce the amount of labor required to analyze failure modes in the detailed design phase. Fault trees can be used to develop the causes of failure and their associated probabilities of occurrence. The fault tree is able to combine failures due to design errors, material properties, manufacturing errors, service mistakes, and even user errors to give the probability with which the failure mode occurs. Failure modes can be eliminated by removing their causes or at least have their probabilities of failure reduced to acceptable levels. Researchers are making progress in developing tools to assist the designer in performing the analysis.

References

1. M. Campbell, "History and Background of FMECA," unpublished personal communication

2. W.R. Lomas, "Safety Considerations in the Design of Flight Control Systems for Navy Aircraft," Paper 60-AV-34, American Society of Mechanical Engineers, 1960

3. Mil-F-18372 (Aer.), "General Specification for Design, Installations, and Test of Aircraft Flight Control Systems," Bureau of Naval Weapons, Department of the Navy, Washington, D.C.

4. J.S. Coutinho, Failure-Effect Analysis, *Transactions New York Academy of Sciences,* Vol 26, 1964, p 564–585

5. "Fault/Failure Analysis Procedure," ARP 926, Society of Automotive Engineers Aerospace Recommended Practice, 15 Sept 1967; ARP 926A, 15 Nov 1979

6. "Failure Mode and Effect Analyses," Electronic Industries Association G-41 Committee on Reliability, Reliability Bulletin No. 9, Nov 1971

7. "Analysis Techniques for System Reliability—Procedure for Failure Mode and Effects Analysis (FMEA)," International Electrotechnical Commission, IEC Standard Publication 812, 1985

8. “Fault/Failure Analysis For Digital Systems and Equipment,” ARP 1834, Society of Automotive Engineers Aerospace Recommended Practice, Aug 1986

9. “Procedures for Performing a Failure Mode Effects and Criticality Analysis,” US MIL-STD-1629 (ships), 1 Nov 1974, US MIL-STD-1629A, 24 Nov 1980, US MIL-STD-1629A/Notice 2, 28 Nov 1984

10. C.S. Spangler, Systems Engineering—The Fault Analysis Process for Commercial Avionics Application, *Proc. of the Third Annual International Symposium of the National Council on Systems Engineering*, 1993, p 729–736

11. “Recommended Failure Mode and Effects Analysis (FMEA) Practices for Non-Automobile Applications,” ARP 5580, Society of Automotive Engineers Aerospace Recommended Practice, July 2001

12. J.B. Bowles, The New SAE FMECA Standard, *Proc. Annual Reliability and Maintainability Symposium*, 1998, p 48–53

13. P.L. Goddard, Validating the Safety of Real Time Control Systems Using FMEA, *Proc. Annual Reliability and Maintainability Symposium,* 1993, p 227–230

14. P.L. Goddard, Software FMEA Techniques, *Proc. Annual Reliability and Maintainability Symposium*, 2000, p 118–123

15. “Potential Failure Mode and Effects Analysis in Design (Design FMEA) and for Manufacturing and Assembly Processes (Process FMEA) Instruction Manual,” Ford Motor Company, Sept 1988

16. “Potential Failure Mode and Effects Analysis in Design (Design FMEA) and Potential Failure Mode and Effects Analysis in Manufacturing and Assembly Processes (Process FMEA) Reference Manual,” J1739, Society of Automotive Engineers, Surface Vehicle Recommended Practice, July 1994

17. C.S. Spangler, Equivalence Relations within the Failure Mode and Effects Analysis, *Proc. Annual Reliability and Maintainability Symposium,* 1999, p 352–357

18. M. Krasich, Use of Fault Tree Analysis for Evaluation of System-Reliability Improvements in Design Phase, *Proc. Annual Reliability and Maintainability Symposium*, 2000, p 1–7

19. T.A. Montgomery, D.R. Pugh, S.T. Leedham, and S.R. Twitchett, FMEA Automation for the Complete Design Process, *Proc. Annual Reliability and Maintainability Symposium,* 1996, p 30–36

20. C.J. Price, Effortless Incremental Design FMEA, *Proc. Annual Reliability and Maintainability Symposium*, 1996, p 30–36

21. C.J. Price, D.R. Pugh, M.S. Wilson, and N. Snooke, The Flame System: Automating Electrical Failure Mode and Effects Analysis (FMEA), *Proc. Annual Reliability and Maintainability Symposium*, 1995, p 90–95

22. D.J. Russomanno, R.D. Bonnell, and J.B. Bowles, Viewing Computer-Aided Failure Modes and Effects Analysis from an Artificial Intelligence Perspective, *Integrated Computer-Aided Engineering*, Vol. 1 (No. 3), Wiley, 1994, p 209–228

23. J.B. Bowles and R.D. Bonnell, Failure Mode, Effects, and Criticality Analysis (What It Is and How to Use It), *Topics in Reliability and Maintainability and Statistics, Annual Reliability and Maintainability Symposium tutorial notes,* 1999

24. “The FMECA Process in the Concurrent Engineering (CE) Environment,” AIR 4845, Society of Automotive Engineers Aerospace Information Report, 18 June, 1993

25. H.E. Lambert, Report No. UCID-16238 Lawrence Livermore Laboratory, May 1973

26. “Failure Mode/Mechanism Distributions 1997,” FMD-97, Reliability Analysis Center, IIT Research Institute, 1997

27. J.B. Bowles, Designing for Failure: Managing the Failure Response through Analysis, *Process of Failure Prevention through Education: Getting to the Root Cause*, J. Scutti, Ed., ASM International, 23–25 May 2000, Cleveland, p 62–69

28. D. Kececioglu, *Reliability Engineering Handbook*, Vol 2, Prentice-Hall, 1991

Reliability-Centered Maintenance

Dana Netherton, Athos Corporation

Introduction

RELIABILITY-CENTERED MAINTENANCE (RCM) is a systematic methodology for preventing failures. It is a specific process used to identify the policies that must be implemented to manage the failure modes that could cause the functional failure of any physical asset in a given operating context (Ref 1). In its present form, RCM has been used since the 1980s in dozens of industries at hundreds of sites around the world. Earlier forms were used by commercial aviation in the United States and in Europe starting in the 1960s.

In some ways, the name “RCM” is misleading. Despite the presence of the word “reliability” in its name, RCM is not a field within reliability engineering. Although it can be used with new assets not yet placed in service, RCM is principally intended for use with assets already built, not with assets that are still in the earliest and most fluid stages of the design process.

In addition, despite the presence of the word “maintenance” in its name, RCM is not simply a method for developing maintenance programs; its recommendations reach far beyond the maintenance department. A RCM review may trigger actions for engineering (redesign), for training and technical documentation (addressing human error due to inadequate training or inappropriate procedures), and for operations (addressing errors performed by operators, not just by maintainers).

Reliability-centered maintenance is a process that offers an opportunity for an organization to take a strategic view of its policies for managing the consequences of the failures of its assets. It answers questions such as: Can a facility avoid undesired consequences most effectively by overhauling or replacing everything periodically or by monitoring performance for signs of deterioration? If so, which assets should be monitored, which should be overhauled, and how often? What aspect of performance should be monitored? How much deterioration should trigger repair work?

Most sites have idle assets that are standing by in the event they are needed, such as backup generators or fire safety systems. Being idle, they do not wear out through use and their performance cannot be monitored, so how should they be handled?

Sometimes, no routine cyclic measure can prevent failures with intolerable consequences. What should be done in these cases?

When properly applied, RCM offers rigorous and defensible answers to each of these questions. Its approach to these questions is the result of 40 years of work in the field of physical asset management.

Reference cited in this section

1. “Evaluation Criteria for Reliability-Centered Maintenance (RCM) Processes,” JA1011, Society of Automotive Engineers, 1999

History of RCM (Ref 2)

In the late 1950s, the commercial aviation industry of the United States began to put its first jet airliners, the Boeing 707 and the Douglas DC-8, into passenger service. Industry-wide safety statistics began to be published in 1959, and they were troubling. If the jet airliners were flying at today's operating tempos, they would have an accident roughly every day.* Instead, commercial jets operated by U.S. airlines had seven accidents in the whole of the year 2000, that is, on average, about one accident every two months (Ref 3). The story behind this achievement is the story of RCM.

At that time, conventional wisdom said that reliability depended directly on scheduled maintenance; in other words, the more scheduled maintenance, the more reliable the equipment. Conventional wisdom also said that the best form of scheduled maintenance was a thorough overhaul, which was believed to make the asset like new and thus (it was believed) at its highest level of reliability. In accordance with this conventional wisdom,

the initial scheduled maintenance program for the DC-8, published in 1959, established overhaul intervals for 339 items in the aircraft in addition to periodic overhauls for the engines and for the airplane as a whole.

However, there was some evidence that this conventional wisdom had flaws. The airlines had been experiencing persistent problems with some of their reciprocating engines, which they overhauled regularly under the overall supervision of the Federal Aviation Administration (FAA) of the U.S. government. When the airlines shortened the overhaul intervals of these engines, a few became more reliable. A few more were unchanged, but many engines became less reliable.

This discovery stunned most maintenance managers. The airlines responded with a series of reliability programs in close cooperation with the FAA. These programs studied the relationship between reliability and age, first in reciprocating and jet engines and later in other complex items scheduled for routine overhaul. One lesson learned through these programs was that periodic overhauls were seldom effective with complex items. This accounted for the reciprocating engines whose reliability had been unchanged by more frequent overhauls. However, there was still a need to account for those that had become less reliable with more frequent overhauls.

As overhaul intervals grew longer, some items did fail before their scheduled overhaul. Rather than being repaired with a thorough overhaul, they were repaired by replacing or restoring the specific components affected by the failure and then were put back into service. Beginning in 1962, the airlines compared the postrepair reliability of these items with the reliability of items overhauled at the same age. There was only one difference between them: the items that were overhauled were more subject to premature failure. In other words, the overhauls were reducing their reliability. Overhauls did not simply waste resources; in many cases, they harmed the items being overhauled.

Indeed, when the results were assembled from the studies of one airline, it was clear that the relationship between reliability and age did not follow one or two models, such as simple wearout or the bathtub curve well-known in the field of reliability engineering. It followed six models. Three showed a clear, negative relationship between reliability and age (wearout, bathtub, and a linear relationship); they comprised 11% of the several hundred items studied. Two showed, in effect, a constant relationship between reliability and age; these comprised 21% of the items studied. The model that comprised the majority of the items, 68%, showed a positive relationship between reliability and age; the older the item, the more reliable it proved to be. These items did not wear out, they wore in (Ref 2).

By 1972, these results were so conclusive that the FAA permitted the airlines to abandon routine overhauls of engines and other complex items aboard airliners. Meanwhile, the continuing engine reliability problems were being solved by improving the design of the engines, not by overhauling them more often.

With the collapse of the routine overhaul as the maintenance approach of choice, no one in commercial aviation pretended that scheduled maintenance was no longer required. Instead, the airlines looked for an approach to scheduled maintenance that could be demonstrated to work. In 1964, the FAA authorized United Airlines to begin a program called “Test and Replace as Necessary.” The scheduled tasks consisted only of tests to identify reduced resistance to failure. Only those units that failed the test were removed and sent to a shop for repair. By 1969, United had qualified 209 items on various types of aircraft for this program. The program proved remarkably successful and served as a model for on-condition maintenance, that is, maintenance in which repairs are only performed on the condition that a potential failure had been detected.

By the mid-1960s, it was clear to maintenance experts in commercial aviation that scheduled overhauls were on their way out and that on-condition maintenance was on the way in. Experience with the reliability programs was also beginning to develop criteria for identifying the appropriate approach for a given item when installed in a given application. The time was ripe to consolidate the lessons into a comprehensive process for identifying the best approach for handling the failures of items aboard commercial aircraft.

The first public proposal for a decision diagram was presented at a FAA maintenance symposium in 1965. By 1968, the 747 Maintenance Steering Group, or MSG (responsible for creating the initial maintenance program for the new Boeing 747, which was about to enter service, and composed of representatives from the airlines, the aircraft manufacturer, the engine manufacturer, and the FAA), drafted MSG-1, *Handbook: Maintenance Evaluation and Program Development.* Working groups reporting to the MSG “sorted out the potential maintenance tasks and then evaluated them to determine which must be done for operating safety or essential hidden function protection. The remaining potential tasks were evaluated to determine whether they were economically useful. These procedures provide a systematic review of the aircraft design so that, in the absence of real experience, the best (maintenance) process can be utilized for each component or system” (Ref 5).

This experience led to the improved document, *Airline/Manufacturer Maintenance Program Planning Document:* MSG-2, published in 1970. It was used to develop the initial maintenance programs for the Lockheed TriStar 1011 and the Douglas DC-10. A similar document, prepared in Europe and titled "European Maintenance Systems Guide," was used for the Airbus Industrie A-300 and the Concorde. The impact of the new ideas about maintenance is evident from the number of items scheduled for routine overhaul: eight on the Boeing 747 and seven on the Douglas DC-10, compared to 339 on the initial program of the Douglas DC-8, published in 1959.

In 1972, at least one airline used MSG-2 to rewrite the maintenance program of older aircraft (Ref 2). This additional experience showed further areas for improvement; that is, MSG-2 did not make clear the role of failure consequences in establishing maintenance requirements, nor did it address the role of hidden-function failures in a sequence of multiple failures. During the 1970s, proposals emerged for ways to handle those issues. In 1978, responding to a request from the Department of Defense (DoD) of the U.S. government, two United Airlines executives wrote a 500-page technical report that described the state of the art in commercial aviation preventive maintenance. Stan Nowlan and Howard Heap named their technical report "Reliability-Centered Maintenance," and this title gave the final version of the process the name by which it is known today.

Once the DoD published Nowlan and Heap's report, the U.S. military embarked on developing RCM processes for its own use, one for the U.S. Army, one for the U.S. Air Force, and two for the U.S. Navy, because the shipboard and aviation communities of the Navy insisted that a RCM process that worked for one would not work for the other. Support contractors and equipment vendors learned to use these processes when they sold new equipment to the U.S. military. Developed in test programs during the late 1970s, the processes were formally published in military standards and military specifications in the mid-1980s.

In separate but parallel work in the early 1980s, the Electric Power Research Institute, an industry research group for the U.S. electrical power utilities, carried out two pilot applications of RCM in the U.S. nuclear power industry. Their interest arose from a belief that this industry was achieving adequate levels of safety and reliability but was massively overmaintaining its equipment. As a result, their main thrust was simply to reduce maintenance costs rather than to improve reliability, and they modified the RCM process accordingly. (So much so, in fact, that it bears little resemblance to the original RCM process described by Nowlan and Heap; it should be more correctly described as planned maintenance optimization rather than RCM.) This modified process was adopted on an industry-wide basis by the U.S. nuclear power industry in 1987, and variations of this approach were afterward adopted by various other nuclear utilities, other branches of the electricity generation and distribution industry, and parts of the oil industry.

Also in the late 1980s, other maintenance-strategy specialists became interested in RCM. The most widespread and innovative were John Moubray and his associates. Initially, they worked with RCM in mining and manufacturing industries in southern Africa under the mentorship of Stan Nowlan. Later, they relocated to Britain where they expanded their activities to cover the application of RCM in nearly every industrial sector, spanning more than 40 countries. They built on Nowlan's work while retaining its original focus on equipment safety and reliability. For example, they added environmental issues to the safety issues in the decision-making process, clarified the ways in which equipment functions should be defined, developed more precise rules for selecting maintenance tasks and task intervals, and incorporated quantitative risk criteria directly into the setting of failure-finding task intervals. They call their enhanced version of reliability-centered maintenance "RCM2."

In the 1990s, RCM began to be implemented in U.S. industries outside the military and nuclear power. In response, processes emerged that were called RCM by their proponents but that often bore little or no resemblance to the original meticulously researched, highly structured, and thoroughly proven process described by Nowlan and Heap. As a result, if an organization said that it wanted help in using or learning how to use RCM, it could not be sure what process would be offered.

Responding to this need in 1999, the Society of Automotive Engineers (SAE) published the all-industry standard JA1011, "Evaluation Criteria for Reliability-Centered Maintenance (RCM) Processes." While SAE JA1011 does not present a standard RCM process, it instead presents criteria against which a process may be compared. If the process meets the criteria, it may confidently be called a RCM process; if it does not, it should not.

This section uses SAE JA1011 as its model to describe the key characteristics of a RCM process.

Footnote

* *In commercial aviation statistics, “accident” is defined as “an occurrence associated with the operation of an airplane that takes place between the time any person boards the airplane with the intention of flight and such time as all such persons have disembarked, in which the airplane sustains serious damage, or death or serious injury results from (1) being in or upon the airplane, (2) direct contact with the airplane or anything attached thereto, or (3) direct exposure to jet blast” (Ref 3). In 1959, the accident rate was greater than 30 accidents per million departures (Ref 3). In 2000, U.S. airports saw about 9.6 million departures (Ref 4). At that rate, there would have been at least 288 accidents in 1959, or roughly one per day.

References cited in this section

2. F.S. Nowlan and H.F. Heap, “Reliability-Centered Maintenance,” AD-A066 579, National Technical Information Service, 1978, Appendix B, p 46, 386

3. *Statistical Summary of Commercial Jet Airplane Accidents: Worldwide Operations, 1959–2000,* Boeing Commercial Airplanes, 2001, p 4, 7, 15

4. *Aviation Industry Overview—Fiscal Year 2000,* Federal Aviation Administration, 2001, p 5

5. *Federal Aviation Administration Certification Procedures,* Federal Aviation Administration, May 12, 1972

Overview of the RCM Process

The RCM process involves asking seven questions, always asked in this order (Ref 1):

1. What are the functions and associated desired standards of performance of the asset in its present operating context? The answers to this question will list the functions of the asset.
2. In what ways can it fail to fulfill its functions? The answers to this question will list the functional failures of the asset.
3. What causes each functional failure? The answers to this question will list what RCM calls the “failure modes” of the asset.
4. What happens when each failure occurs? The answers to this question will list the effects of each failure caused by the failure mode.
5. In what way does each failure matter? The answers to this question will list the consequences of each failure caused by the failure mode.
6. What should be done to predict or prevent each failure? The answers to this question will list the proactive tasks that should be performed to address each failure mode, along with the intervals at which these tasks should be performed.
7. What should be done if a suitable proactive task cannot be found? The answers to this question will list the default actions that should be performed to address each failure mode. In the case of cyclic actions, they will also list the intervals at which these actions should be performed.

The following sections expand on each of these questions, offering definitions when necessary. The first four questions comprise a form of failure modes and effects analysis (FMEA) that is tailored specifically to meet the goals of RCM, so the next section describes the approach of RCM to FMEA. (For another perspective on FMEA, see the article “Failure Modes and Effects Analysis” in this Volume.) The following section describes the failure management policies available under RCM. Then, after a section describing the ways that RCM classifies failure effects in terms of consequences, a section describes how RCM uses failure consequences to identify the best failure management policy for each failure mode.

The closing section, “Managing and Resourcing the RCM Process,” discusses some practical issues pertaining to RCM that lie outside the scope of SAE JA1011.

Reference cited in this section

1. “Evaluation Criteria for Reliability-Centered

Failure Modes and Effects Analysis

As described in the section “Overview of the RCM Process,” RCM involves asking seven questions, of which the first four comprise a form of FMEA that is tailored specifically to meet the goals of RCM. Failure modes and effects analysis of RCM differ from the FMEA used elsewhere in engineering in several important respects.

First, the terminology of RCM (Table 1) differs from that of traditional engineering FMEA. Reliability-centered maintenance uses the term “failure modes” to refer to the events called “failure causes” in traditional FMEA. Reliability-centered maintenance uses the term “functional failure” to refer to the state called “failure mode” in traditional FMEA. These uses are entrenched in the RCM community by decades of familiarity, and it is not likely that they will change to match those of traditional FMEA any time soon. (Table 1 provides a glossary of RCM terms.)

Table 1 Glossary of terms used in reliability-centered maintenance

Term	Definition in RCM
Age	A measure of exposure to stress. Computed from the moment an item or component enters service when new or reenters service after a task designed to restore its initial capability. Can be measured in terms of calendar time, running time, distance traveled, duty cycles, or units of output or throughput
Appropriate task	A task that is both technically feasible and worth doing (applicable and effective)
Conditional probability of failure	The probability that a failure will occur in a specific period, provided that the item concerned has survived to the beginning of that period
Desired performance	The level of performance desired by the owner or user of a physical asset or system
Environmental consequences	A failure mode or multiple failure has environmental consequences if it could breach any corporate, municipal, regional, national, or international environmental standard or regulation that applies to the physical asset or system under consideration.
Evident failure	A failure mode whose effects become apparent to the operating crew under normal circumstances if the failure mode occurs on its own
Evident function	A function whose failure on its own becomes apparent to the operating crew under normal circumstances
Failure consequences	The way(s) in which the effects of a failure mode or a multiple failure matter (evidence of failure, impact on safety, the environment, operational capability, direct and indirect repair costs)
Failure effect	What happens when a failure mode occurs
Failure-finding task	A scheduled task used to determine whether a specific hidden failure has occurred
Failure management policy	A generic term that encompasses on-condition tasks, scheduled restoration, scheduled discard, failure finding, run-to-failure, and one-time changes
Failure mode	A single event that causes a functional failure
Function	What the owner or user of a physical asset or system wants it to do
Functional failure	A state in which a physical asset or system is unable to perform a specific function to a desired level of performance
Hidden failure	A failure mode whose effects do not become apparent to the operating crew under normal circumstances if the failure mode occurs on its own
Hidden function	A function whose failure on its own does not become apparent to the operating crew under normal circumstances
Initial capability	The level of performance that a physical asset or system is capable of achieving at the moment it enters service
Multiple failure	An event that occurs if a protected function fails while its protective device or protective system is in a failed state
Net P-F interval	The minimum interval likely to elapse between the discovery of a potential failure and the occurrence of the functional failure
Nonoperational consequences	A category of failure consequences that do not adversely affect safety, the environment, or operations, but only require repair or replacement of any item(s) that may be affected by the failure
On-condition task	A scheduled task used to detect a potential failure

Term	Definition in RCM
One-time change	Any action taken to change the physical configuration of an asset or system (redesign or modification), to change the method used by an operator or maintainer to perform a specific task, to change the operating context of the system, or to change the capability of an operator or maintainer (training)
Operating context	The circumstances in which a physical asset or system is expected to operate
Operational consequences	A category of failure consequences that adversely affect the operational capability of a physical asset or system (output, product quality, customer service, military capability, or operating costs in addition to the cost of repair)
Owner	A person or organization that may either suffer or be held accountable for the consequences of a failure mode by virtue of ownership of the asset or system
P-F interval	The interval between the point at which a potential failure becomes detectable and the point at which it degrades into a functional failure (also known as “failure development period” and “lead time to failure”)
Potential failure	An identifiable condition that indicates that a functional failure is either about to occur or is in the process of occurring
Proactive maintenance	Maintenance undertaken before a failure occurs in order to prevent the item from getting into a failed state (scheduled restoration, scheduled discard, and on-condition maintenance)
Protective device or protective system	A device or system that is intended to avoid, eliminate, or minimize the consequences of failure of some other system
Primary function(s)	The function(s) that constitute the main reason(s) why a physical asset or system is acquired by its owner or user
Run-to-failure	A failure management policy that permits a specific failure mode to occur without any attempt to anticipate or prevent it
Safety consequences	A failure mode or multiple failure has safety consequences if it could injure or kill a human being.
Scheduled	Performed at fixed, predetermined intervals, including continuous monitoring (where the interval is effectively zero)
Scheduled discard	A scheduled task that entails discarding an item at or before a specified age limit, regardless of its condition at the time
Scheduled restoration	A scheduled task that restores the capability of an item at or before a specified interval (age limit), regardless of its condition at the time, to a level that provides a tolerable probability of survival to the end of another specified interval
Secondary functions	Functions that a physical asset or system has to fulfill apart from its primary function(s), such as those needed to fulfill regulatory requirements and those that concern issues such as protection, control, containment, comfort, appearance, energy efficiency, and structural integrity
User	A person or organization that operates an asset or system and may either suffer or be held accountable for the consequences of a failure mode of that system

Second, RCM and traditional FMEA deliver different outputs, because they are aimed at different goals. Traditional FMEA primarily supports the asset in its initial design. At this early stage, the preferred approach to addressing failure causes/failure modes that have intolerable consequences is to change the design of the asset. Once the reliability of the design has reached the highest level achievable within the practical constraints of the design program, any failure causes/failure modes that still remain will be addressed through compensating provisions such as maintenance. Typically, a traditional FMEA will offer a great deal of information to support the design process but will offer relatively little information about the details of such maintenance (such as the type of maintenance task that should be performed or how often it should be performed).

By contrast, RCM primarily supports the asset after it has been designed and once it is to be put in use. At this stage in the life of the asset, design changes tend to be relatively expensive to research and develop. As a result, they also tend to be slow to be installed, absent a disaster that has galvanized decision makers. Therefore, the preferred approach to addressing failure causes/failure modes that would have intolerable consequences is usually to perform some sort of maintenance. For example, the preferred approach to addressing catastrophic failures in automobile engines that are caused by deteriorated lubricants is usually to replace the lubricants periodically, not to redesign the lubricants or the engine.

Therefore, a RCM review will offer a great deal of information to support the maintenance program but will offer relatively little information about the details of any redesign that might be needed (such as details of the features of the new design).

These different goals are also reflected in the failure causes that each process identifies. The failure process is a chain of events beginning far into the distant past and ending in the inability of the asset to perform its function. Which specific event should be called the cause of the failure? Typically, one selects events that can be easily controlled and that will also have a great effect on the outcome of the process.

During the design phase of the lifecycle of an asset, it is easier to control minute details of the failure process, such as whether the casing of the asset fails by brittle fracture or by ductile deformation. Once the asset is in place, that feature can only be changed by replacing its casing. So, during the operating phase of the lifecycle of the asset, casing failures are more likely to be addressed through inspections or through operational restrictions (such as setting a minimum temperature for the area around the asset).

Thus, the similarities between FMEA of RCM and traditional FMEA are superficial, and traditional FMEA seldom serves as a satisfactory replacement for FMEA of RCM.

Functions. According to SAE JA1011, the description of asset functions in a RCM process has four characteristics (Ref 1):

- It defines the operating context. The operating context is a statement of the relevant features of the context in which the asset operates. For example, it is relevant (to the RCM review) to note whether the asset is operated continuously or in cycles. It is relevant (to the RCM review) to note any production targets that the asset supports. It is relevant (to the RCM review) to note features in the surrounding physical environment that may have an effect on the behavior of the asset.
- The list of functions is a complete list of all functions of the asset, that is, all primary and secondary functions, including all protective functions. It is necessary to say this, because some processes purporting to be RCM processes named only the primary function of the asset or added only a handful of secondary functions.
- Each function statement is composed with a verb, an object, and a performance standard (quantified in every case where this can be done). For example, the primary function of a household drill might be “to drill a hole in wood, metal, or masonry, at diameters in increments of $\frac{1}{16}$ in., plus or minus $\frac{1}{64}$ in, up to the length of the drill bit, in less than 30 s.” (Where no performance standard is given, an absolute standard will be inferred. For example, the function statement “to contain fluid,” without a maximum leak rate, implies that no leaks will be tolerated.)
- The performance standards used in these function statements describe the level of performance desired by the owner or user of the asset in its operating context (not the level of performance defined by the design engineer or the purchasing department).

Readers familiar with traditional FMEA may note the presence of performance standards in the function statement of a RCM review. Traditional FMEA does not always include performance standards, partly (at least) because the design engineer who performs the analysis does not always know the level of performance desired by the owner or user of the asset in its operating context.

However, RCM was not developed for the design engineer but for those responsible for ensuring that the asset is available for productive use. If the maintenance function is to know what it needs to do to ensure that the asset is available, its people must be told what level of performance they are supposed to maintain. If the owners or users of the asset cannot tell the maintenance function what they want it to do, they cannot hold the maintenance function responsible when the asset fails to do it.

Looking at these four characteristics, it should be clear that this question is no easy challenge. It is not unusual to find that about one-third of the time spent in a typical RCM review is spent on defining the functions of the asset under review—one-third of the time is spent on one of the seven questions! It was probably that experience that moved some consultants to speed up RCM by halting after identifying only a few functions.

However, an incomplete list of functions is counterproductive and sometimes dangerous. It is not unusual to find that the solution to a particularly stubborn problem is found in this very step, which is likely to reveal when the owner or user of the asset is unwittingly demanding more of the asset than it is capable of providing. (This happens most often when the asset has been in operation for some time and has been subjected to incremental modifications without perfect consultation with all other people in the organization.) In such a case, the solution is typically very simple (either stop using it so hard or get a model with greater capacity) and is discovered very quickly indeed—long before the official end of the RCM review.

It is also not unusual for an asset to have some protective functions that have been overlooked by its owner or user in favor of its more obvious primary and evident secondary functions. When that happens, for example, with a backup generator or a locker of emergency tools, the site will be unprotected when the emergency eventually does arise. A physical asset management program that neglects its protective functions is likely to be complacent and dangerous.

Functional Failures. According to SAE JA1011, the description of functional failures in a RCM process has one characteristic: it identifies all of the failed states associated with each function (Ref 1).

If the performance standards were defined well, this exercise will be simple, yet precise. If the function of a pump is to provide at least 300 L/min of flow, then the failure of that function is to be unable to provide 300 L/min of flow.

Readers who are familiar with traditional FMEA should note again that this term corresponds to “failure mode” in traditional FMEA, albeit failure modes that always have performance standards.

Failure Modes. In RCM, a failure mode is “a single event that causes a functional failure.” According to SAE JA1011, the list of failure modes in a RCM process has five characteristics (Ref 1).

First, it identifies all failure modes reasonably likely to cause each functional failure. Note that this implies the existence of more than one failure mode for each functional failure. Indeed, there may be several events that might cause a pump to fail to deliver the desired flow rate. A RCM process will identify all of them so that it can find a way to address them all.

A RCM process will not identify all imaginable failure modes. Some people can have active imaginations and can dream up failure modes that are frankly not credible. A RCM process will have a method for weeding out the incredible speculations from the reasonably likely assessments of the RCM review.

Second, the method used to decide what constitutes a reasonably likely failure mode is acceptable to the owner or user of the asset. Different owners may need to use different levels of likelihood when screening the failure modes they should include in the review. For example, should the review include “airplane crashes on site” as a reasonably likely failure mode? While most sites will not, the owners of some nuclear power plants in the United States need to do so.

Third, it identifies failure modes at a level of causation that makes it possible to identify an appropriate failure management policy. Level of causation alludes to the fact that functional failure is the last event in a chain of events. Working backward along that chain can be described as looking at deeper levels of causation. Some failure investigators will look at the events that take place immediately before the failure; others will look at the events that take place long before the failure. In the vocabulary of RCM, those investigators are looking at deeper levels of causation.

Various RCM processes may have different methods of identifying the appropriate level of causation, with each method promising to do so more efficiently and more effectively than the others, but SAE JA1011-compliant processes will all be looking for the level of causation at which the most appropriate failure management policy can be identified.

Appropriate failure management policies are those that are technically feasible and worth doing. In RCM, these phrases are technical terms and are discussed later in this article. At this point, it is enough to remember the observation made earlier when discussing the difference between traditional FMEA and the FMEA of RCM, that design engineers and members of a RCM review group are likely to focus on different events that might be said to cause functional failure.

Design engineers have relatively more control over events that take place early in the design life of an asset, such as the selection of materials or of a specific geometry for controlling stresses in a structure; they have relatively little control over the way that the asset is operated and maintained. A RCM review group has relatively little control over the design of the asset and relatively more control over its operation and maintenance. Although the details of technical feasibility are discussed later, it should go without saying that it is not feasible to expect people to solve a problem by focusing on events over which they have little or no control.

Fourth, the list of failure modes in a RCM process includes those failure modes that have happened before, those that are currently being prevented by existing maintenance programs, and those that have not yet happened “but that are thought to be reasonably likely (credible) in the operating context.”

This is the first of two lists of types of failure modes that are often excluded from reviews. This list sorts failure modes by the type of information used to identify them, such as failure history or the current maintenance program. Some review processes focus on the most recent failure, because they are only intended to solve the problem presented by that failure—after the failure has already occurred. Some review processes examine only the existing maintenance program, because they are intended only to update or optimize that program.

However, the purpose of RCM is “to identify the policies that must be implemented to manage the failure modes that could cause the functional failure of any physical asset in a given operating context.” This purpose is much broader than simply cleaning up after a failure or optimizing a maintenance program. A process that supports this purpose may require changes in work procedures, training, or the design of the asset—and may require them before the failure has actually taken place. In this respect, RCM is truly a process that prevents failures from happening—those failures whose consequences make the effort worth doing, as described subsequently in this article.

Fifth, the list should also include “any event or process that is likely to cause a functional failure, including deterioration, design defects, and human error whether caused by operators or maintainers (unless human error is being actively addressed by analytical processes apart from RCM).”

This, the second list, sorts failure modes by type of failure mechanism. It is human for investigators to limit their examinations to the kinds of failure mechanisms they understand. A metallurgist may prefer to solve materials-related failure modes and may prefer to give human error to someone else to solve. A maintenance worker may prefer to solve the failure of workmanship and may prefer to give design defects to someone else to solve. However, RCM is intended to find solutions to problems that lead to significant failures, wherever the solutions may be found.

Although SAE JA1011 does not address the resources required to perform RCM, this aspect of RCM does suggest that the best setting for conducting a rigorous RCM review is a multi-disciplinary setting, probably in the context of a small

group of technical experts. Indeed, as is discussed later, such a setting has become the norm in most commercial applications of RCM.

Once this step has been completed, the RCM review will have identified all failure modes that are reasonably likely to cause the asset to experience functional failures in its operating context.

Failure Effects. One of the most distinctive features of RCM is its explicit assessment of whether a proposed action is worth doing, that is, whether the benefit gained from the action justifies the effort required to take the action. In order to make this assessment, it is necessary to know what would happen if the failure took place without taking any action. Reliability-centered maintenance collects that information in the form of a description of the effects of the failure.

According to SAE JA1011, the description of failure effects in a RCM process has two characteristics.

First, the description specifies what would happen if no specific task is done to anticipate, prevent, or detect the failure.

There is a temptation to describe what would normally happen if the failure mode occurred, including the normal response of maintenance and casualty personnel to signs that the failure is about to occur—a response that is intended to prevent or mitigate the consequences of the failure.

However, this part of the RCM review is intended to help the reviewers assess whether any such response is needed. After all, not all failures need to be prevented. For example, most homeowners permit their light bulbs to burn out, and they replace them after failure. So, the description of failure effects must not assume the very responses that are to be questioned.

Second, the description includes all information needed to support the evaluation of the consequences of the failure.

Such information includes any evidence that the failure has occurred. Because alarms or other indications sometimes appear after the failure itself occurs (such as low-level alarms in tanks, which may appear some time after the feed pump of the tank has failed), this information can indicate how long the effects of the failure are likely to have been developing without an opportunity for intervention.

It also includes anything the failure does that would kill or injure someone, harm the environment, hinder production or operations, or cause physical damage to other assets. These categories relate directly to consequence categories, which are discussed later.

It also includes any actions that must be taken to restore the function of the system after the failure. Broadly speaking, if a great deal of work is needed after a failure, it is more likely to be worth doing something to prevent having to do it.

Once this list is compiled, the RCM review will have a complete FMEA tailored to meet the goals of RCM.

Reference cited in this section

1. “Evaluation Criteria for Reliability-Centered Maintenance (RCM) Processes,” JA1011, Society of Automotive Engineers, 1999

Failure Management Policies and “Technical Feasibility”

Unlike some other processes, RCM considers the full range of options for managing the consequences of the failure of physical assets. This section describes the failure management policies available under RCM and explains the criteria of RCM for deciding when a specific failure management policy is technically feasible.

Determining whether a given policy is technically feasible involves comparing the technical features of the failure mode with the criteria of RCM for technical feasibility for that policy. The criteria of RCM for technical feasibility vary from one type of failure management policy to another, because they are based on the technical characteristics of each policy.

According to SAE JA1011, a RCM process offers two kinds of failure management policies: scheduled tasks and unscheduled policies. In the context of RCM, scheduled tasks consist of tasks that are performed at fixed, predetermined intervals. (Tasks that are performed at irregular intervals but that are planned in advance, such as repairs performed in response to equipment condition, are not called scheduled tasks in the RCM context.)

Scheduled Tasks. According to SAE JA1011, RCM offers four kinds of scheduled tasks: on-condition tasks, scheduled restoration tasks, scheduled discard tasks, and failure-finding tasks (Ref 1).

On-condition tasks look for a potential failure—for an identifiable condition or warning sign that indicates that a functional failure is either about to occur or is in the process of occurring. For example, checking air pressure in tires is an on-condition task. The owner or user of the car only adds air to a tire on the condition that the check revealed low air pressure. It is possible to look for potential failure by using specialized condition monitoring equipment, existing process instrumentation (such as flow rates, pressures, and temperatures), existing quality management techniques (such as statistical process control), or the human senses.

It should be clear that an on-condition task is only technically feasible if there is a potential failure that can be discovered. If failure occurs suddenly and without practical warning (often the case with brittle fractures), then an on-condition task is not technically feasible. According to SAE JA1011, a technically feasible on-condition task has five characteristics (Ref 1):

- A clearly defined potential failure does exist. If the potential failure cannot be readily recognized, it will do little good to look for it.
- There is an identifiable interval between potential failure and functional failure, known as the P-F interval (or failure development period). This and the rest of the characteristics depend on the point that it will do little good to look for the warning sign if it does not give enough time to do something useful.
- The task interval is less than the shortest likely P-F interval. Note that the definition of a "task" in a RCM review includes the definition of the interval at which the task should be performed.
- It is physically possible to do the task at intervals less than the P-F interval.
- The shortest time between the discovery of a potential failure and the occurrence of the functional failure is long enough to take predetermined action to avoid, eliminate, or minimize the consequences of the failure mode. The time between the discovery of a potential failure and the occurrence of the failure mode is shorter than the P-F interval; that is, it is equal to the P-F interval of the asset minus the interval of the task, because a potential failure might emerge soon after the task has been accomplished (with no potential failure in sight), giving the asset nearly the entire time of the task interval to develop toward functional failure. Once the next task does detect the potential failure, there might be little time left until the onset of functional failure.

Human senses often detect the most obvious potential failures and have the shortest P-F intervals and the lowest costs. Specialized inspection technology, which may detect more subtle potential failures, has the longest P-F intervals and the highest costs. Installed instrumentation can also be useful and can lie between these two extremes.

Every inspection technique has some likelihood that a single inspection might fail to detect the potential failure. This likelihood may be increased if the technique is expected to detect extremely subtle potential failures. If a user, exercising such care and skill as it is reasonable to expect in the relevant operating context, is reasonably likely to miss a potential failure with a specific technique, then the potential failure is not clearly defined (see the first characteristic of a technically feasible on-condition task, mentioned previously) and an on-condition task using that technique to focus on that potential failure is not technically feasible.

Depending on local circumstances, people using RCM might decide, instead, to use the same technology to detect more-obvious failure modes that it can detect reliably (such as looking for small cracks rather than extremely tiny cracks), accepting the shorter P-F interval that would result. They might decide to use different technology that is more reliable (if available). They might decide to abandon on-condition tasks altogether and try something else. (In this last case, the technology may still be useful when looking for other degraded areas after a potential failure has been detected, in order to define the scope of the repair work, or when performing postrepair inspections.)

Not all failure modes show potential failures, so on-condition tasks are not always technically feasible. (In fact, experience indicates that tasks using equipment to discover potential failures are technically feasible for no more than 20% of the failure modes encountered in RCM reviews [Ref 6].) Reliability-centered maintenance, therefore, considers other kinds of scheduled tasks as well.

Scheduled discard tasks discard an item at or below a specified age limit regardless of its condition at the time. For example, automobile manufacturers recommend that car owners discard their engine oil at specified intervals of miles driven.

According to SAE JA1011, a technically feasible scheduled discard task has two characteristics (Ref 1).

First, there is a clearly defined (preferably a demonstrable) age at which there is an increase in the conditional probability of the failure mode under consideration.

Conditional probability of failure is the probability that a failure will occur in a specific period on the condition that the item has survived to that period. It is possible for this probability to change over the lifetime of an item. Indeed, people perform scheduled discard tasks because they believe that the conditional probability of failure is higher after the age associated with the task than before that age. If this is true, then we can speak of a wearout age before which the item might be discarded.

However, if the conditional probability of failure remains the same—if it is constant versus age—then a scheduled discard task will offer no improvement. If the conditional probability of failure falls with age, as it did with roughly two-thirds of the items examined by the commercial airlines in their reliability programs of the 1960s and early 1970s, a scheduled discard task could harm the asset by replacing a reliable surviving part with an unreliable new part.
However, the existence of a wearout age is not enough to make a scheduled discard task technically feasible.
Second, a sufficiently large proportion of the occurrences of this failure mode occur after the wearout age to reduce the probability of premature failure to a level that is tolerable to the owner or user of the asset.
This could be called the "barn door" characteristic. An item might have an age at which its conditional probability of failure skyrockets; however, if it is highly unreliable before that age, then waiting until the wearout age to take action solves nothing. By the time such an asset reaches its wearout age, it will probably have already failed—closing the barn door after the horse has already left.
The next type of scheduled task, the scheduled restoration task, is similar to the scheduled discard task in many respects. It restores the capability of an item at or before a specified interval (an age limit), regardless of its condition at the time, to a level that provides a tolerable conditional probability of survival to the end of another specified interval.
According to SAE JA1011, a technically feasible scheduled restoration task has three characteristics (Ref 1). The first two are identical to those of a technically feasible scheduled discard task; the third characteristic is unique to the scheduled restoration task:

- There is a clearly defined (preferably a demonstrable) age at which there is an increase in the conditional probability of the failure mode under consideration.
- A sufficiently large proportion of the occurrences of this failure mode occur after the wearout age to reduce the probability of premature failure to a level that is tolerable to the owner or user of the asset.
- The task restores the resistance to failure (condition) of the component to a level that is tolerable to the owner or user of the asset. Clearly, if the scheduled restoration task is unable to restore the item to a tolerable level, it is not technically feasible.

The last type of scheduled task offered by RCM is the failure-finding task. This task determines whether a specific hidden failure has occurred.
A hidden failure is one whose effects do not become apparent to the operating crew under normal circumstances if the failure mode occurs on its own. It is most frequently associated with protective devices or functions, such as alarms or automatic shutdowns, and with standby assets, such as backup equipment or safety equipment. Such equipment can quietly rust in place or be used as a source of spare parts and then be unavailable for use if needed. The failure-finding task is not intended to prevent such failures; it is intended to detect them before the asset is needed, so that corrective action can be taken before an emergency arises.
Hidden failures are different from other failures, because they have no consequences if they occur on their own. They only have consequences in the event of a multiple failure, that is, only if the protected asset or function fails while the hidden function is in a failed state. Therefore, the evaluation of a failure-finding task takes into account some of the features of a multiple failure.
According to SAE JA1011, a technically feasible failure-finding task has four characteristics (Ref 1):

- The basis on which its task interval is selected takes into account the need to reduce the probability of the multiple failure of the associated protected system to a level that is tolerable to the owner or user of the asset. It should be clear from the previous discussion that the purpose of the failure-finding task is to reduce the probability that the multiple failure may occur and to reduce it to a level that is tolerable to the owner or user of the asset. If the task cannot do this, it is not technically feasible. Note, once again, the close association that RCM makes between the task and its interval.
- The task confirms that all components covered by the failure mode description are functional. Because the description of a hidden failure mode tends to be broad ("pump fails"), a technically feasible failure-finding task tends to be broad in scope ("test-run the pump"). This can be a challenge, because test features do not always cover all components. For example, the test button on a household smoke detector only tests that the battery and the horn are working; it does not test the smoke sensor.

- The task and its interval take into account any probability that the task itself might leave the hidden function in a failed state.
- It is physically possible to do the task at the specified intervals. Sometimes, if the owner or user is to be sufficiently confident that an unreliable protective device will not be in a failed state when the protected device fails, the protective device must be tested at intervals of seconds or minutes. Clearly, such a task is not technically feasible.

Unscheduled Policies. In the event that no scheduled task is an appropriate way to handle the consequences of a failure mode, RCM offers two alternatives: a one-time change and allowing the failure mode to progress to failure.

One-time changes represent a change to the operating context in which the asset works. This change may be applied to the physical system, such as a design change. It may be applied to the methods used to operate or maintain the asset through procedural changes (and any required refresher training). It may be applied directly to the people who operate or maintain the asset through training that increases their understanding of the asset.

Such changes usually require investment in redesign, procedural development, and training development. Investments generally require additional levels of approval before work can begin and thus are generally harder to implement than a scheduled task in a maintenance program. This is especially true for systems already in operation whose funding for initial design of hardware, procedures, and training has already been expended.

Thus, RCM tries to achieve desired levels of performance in the system as it is currently configured and operated by applying appropriate scheduled tasks (Ref 1). A one-time change is only technically feasible if such a task is not available.

In the event that no other failure management policy is appropriate, the only technically feasible alternative is to allow the asset to run to failure. Run-to-failure is an acceptable failure management policy in the RCM context, if it is also worth doing according to the criteria of RCM.

The criteria of RCM for deciding whether a failure management policy is worth doing depend on the consequences of the failure mode, so the next section addresses failure consequences.

References cited in this section

1. "Evaluation Criteria for Reliability-Centered Maintenance (RCM) Processes," JA1011, Society of Automotive Engineers, 1999

6. J.M. Moubray, *Reliability-Centered Maintenance,* Industrial Press, 1999, p 155, 290

Failure Consequences and "Worth Doing"

Technical feasibility is not the only factor to be used when considering whether a given failure management policy is appropriate. The other factor is whether the policy is worth doing.

Reliability-centered maintenance bases its examination of this factor on the consequences of the failure mode, first by identifying the consequences of the failure mode and then by applying the appropriate worth-doing criteria to the failure mode and the proposed failure management policy. This section describes the categories of consequences used in RCM and how RCM uses them.

Reliability-centered maintenance uses the term "consequences" to refer to the reason why a failure mode matters to the owner or user of the asset. According to SAE JA1011, RCM has five kinds of consequences: safety, environmental, operational, nonoperational, and hidden. Failure modes that are not hidden are called evident. Operational and nonoperational consequences are sometimes called economic consequences for reasons that will become apparent soon (Ref 1).

Safety. A failure mode with safety consequences matters, because it could injure or kill a human being. Failure modes that could damage equipment (including secondary damage from the failure mode) are addressed in one of the economic categories (operational, nonoperational).

This use of the term "safety" sometimes surprises newcomers to RCM who are accustomed to thinking of a failure with extensive secondary damage as a safety hazard. However, RCM handles the consequences of secondary damage and the consequences of injury or death in very different ways, as discussed subsequently.

Environmental. A failure mode with environmental consequences matters, because it might breach "any corporate, municipal, regional, national, or international environmental standard or regulation which applies to the physical asset or system under consideration" (Ref 1).

Reliability-centered maintenance defines environmental consequences separately from safety consequences, but, as described subsequently, it uses them in the same way as it uses safety consequences.

Operational. A failure mode with operational consequences matters, because it reduces the operational capability of a physical asset or system, that is, its output, product quality, customer service, military capability, or operating costs in addition to the cost of repair (Ref 1).

Reliability-centered maintenance examines the economic impact of this reduced operational capability in order to determine whether a task that addresses such a failure mode is less expensive than the failure mode itself over a comparable period of time. In some cases, the economic impact can be measured directly (for example, if it reduces production of a product that would certainly have been sold when produced). In other cases, the economic impact must be measured indirectly.

For example, the economic impact of lost military capability cannot usually be measured directly. However, when military services plan how many units to procure, they generally take into account the capability they expect to lose in assets that are unavailable due to scheduled maintenance or to unscheduled repairs.

For example, American fighter squadrons flying on missions to Europe in World War II were routinely accompanied by several spare aircraft that would slip into the formation when the failure of equipment in the high-performance (but mechanically unreliable) fighters would force some pilots to turn back before they reached enemy airspace. In some cases, mechanical breakdowns could have a great effect on military capability. When P-51 Mustangs escorted their first daylight bombing raid all the way to Berlin in March 1944, one of the three 16-plane squadrons reached Berlin with only two fighters. The rest had turned back because of equipment problems (Ref 7).

In a modern example, the United States Navy operates and maintains 12 nuclear aircraft carriers in order to have four carriers on-station around the world. With the rate of equipment failures aboard these aircraft carriers and the amount of scheduled maintenance performed on them, the U.S. Navy has found that if n number of carriers are on-station, then that same n number of carriers will be receiving shore-based repairs at any given time, on average, and that n number (again) will be travelling between their station and the shore-based repair depot. When $n = 4$, the total inventory of carriers becomes 12.

If World War II fighters had been more reliable, the military would have been able to buy fewer of them in order to have the same number of aircraft over the target. If nuclear aircraft carriers were more reliable, the military would be able to buy (and maintain) fewer of them in order to have the same number of carriers on-station. The higher procurement numbers—and the higher costs—reflect the indirect economic consequences of the failure modes that reduce the operational capability of these military assets.

Nonoperational. A failure mode with nonoperational consequences matters only because it requires the repair or replacement of one or more items. The magnitude of the consequences is measured simply in economic terms, that is, the cost of the repair, considering both parts and labor (Ref 1).

Every failure mode without noticeable safety, environmental, or operational consequences will be categorized as having nonoperational consequences.

Hidden failure modes are assigned to the category of hidden consequences. Failure-finding tasks are only considered if the failure mode is hidden, because there is no need to find failures when the failure mode is evident.

Some hidden failure modes also have safety or environmental consequences, because many protective devices or functions are installed to protect lives or the environment. The remaining hidden failure modes apply to protective devices or functions that are installed to protect equipment. These hidden failure modes are coupled with economic (operational or nonoperational) consequences.

Worth Doing. It is common for engineers to judge a technical approach on the grounds of its technical feasibility. Reliability-centered maintenance adds another assessment, that is, a comparison of the value added by the approach to the drawbacks of the approach. If the value outweighs the drawbacks, the approach is worth doing.

According to SAE JA1011, a RCM process declares a scheduled task worth doing if it satisfies the criteria in Table 2 (Ref 1). A one-time change is worth doing if the change meets the appropriate criteria in Table 3 (Ref 1).

Table 2 Worth-doing criteria for scheduled tasks

	Evident failure mode	**Hidden failure mode**
Safety or environmental consequences	The task reduces the probability of the failure mode to a level tolerable to the owner or user of the asset.	The task reduces the probability of the multiple failure to a level tolerable to the owner or user of the asset.
No safety or environmental consequences	The direct and indirect costs of doing the task are less than the direct and indirect costs of the failure mode, when measured over comparable periods of time.	The direct and indirect costs of doing the task are less than the direct and indirect costs of the failure mode, plus the cost of repairing the hidden failure mode, when measured over comparable periods of time.

Table 3 Worth-doing criteria for one-time changes

	Hidden failure mode	Evident failure mode
Safety or environmental consequences	The one-time change must reduce the probability of the multiple failure to a level tolerable to the owner or user of the asset.	The one-time change must reduce the probability of the failure mode to a level tolerable to the owner or user of the asset.
No safety or environmental consequences	The one-time change must be cost-effective, in the opinion of the owner or user of the asset.	The one-time change must be cost-effective, in the opinion of the owner or user of the asset.

Once it has been decided that a failure mode has safety or environmental consequences, it will be handled in terms of probability that it will occur. Then, that probability will be compared to the probability that is tolerable to the owner or user of the asset. It can sometimes be a challenge for owners or users to decide what probability of a dangerous failure they are prepared to tolerate. Still, if they do not know what probability they are prepared to tolerate—until they know how much protection they need—they cannot know whether the protection they will get from the task they are evaluating will be adequate.

When examining a failure mode with economic consequences, it is important to compare the costs of the failure mode against the costs of the task over comparable periods of time. It is misleadingly easy to compare the costs of one failure against the costs of performing the task one time. However, the costs of a task done many times will add up over time, and it is those accumulated costs that must be compared against the cost of an expensive but infrequent failure.

Failure modes and effects analysis, failure management policies and technically feasible, and failure consequences and worth doing are the building blocks of RCM. In the next section, they are put together to create a failure management program.

References cited in this section

1. "Evaluation Criteria for Reliability-Centered Maintenance (RCM) Processes," JA1011, Society of Automotive Engineers, 1999

7. G.C. Hall, Jr., *Death Squadron*, Kensington Publishing Corp., 1946, p 178, 183–84

Failure Management Policy Selection

With the list of functions, functional failures, failure modes, and failure effects as well as an understanding of failure consequences and the characteristics of the failure management policies available, all the pieces are in place to support the selection of the appropriate failure management policies for the asset under review.

According to SAE JA1011, all scheduled tasks selected in a RCM process must be technically feasible and also worth doing (Ref 1). The other options, one-time changes and a policy of run-to-failure, must meet their own specific criteria. If more than one policy meets these criteria, then the more cost-effective policy will be selected (Ref 1).

Also, policies are selected as if no specific task is currently being done to anticipate, prevent, or detect the failure (Ref 1).

The criteria used by RCM at this stage are complex, and it can be hard to keep them (and their results) straight. Most RCM processes guide their users through the process with a decision logic diagram. Indeed, the use of decision logic diagrams is so widespread in RCM that some people mistakenly believe that any process with a decision logic diagram is a RCM process. However, no decision logic diagram is the sole RCM decision logic diagram.

The most complex version—also the most highly optimized version—considers all failure management policies for each failure mode, identifies all those that are technically feasible and worth doing, and selects the one that is least expensive.

Figure 1 shows one such decision logic diagram, in this case developed by the U.S. Naval Air Systems Command (Ref 8). It begins the failure management policy selection process by determining whether the failure mode is hidden or evident. This determination makes failure-finding tasks available as an option later if the failure mode is hidden.

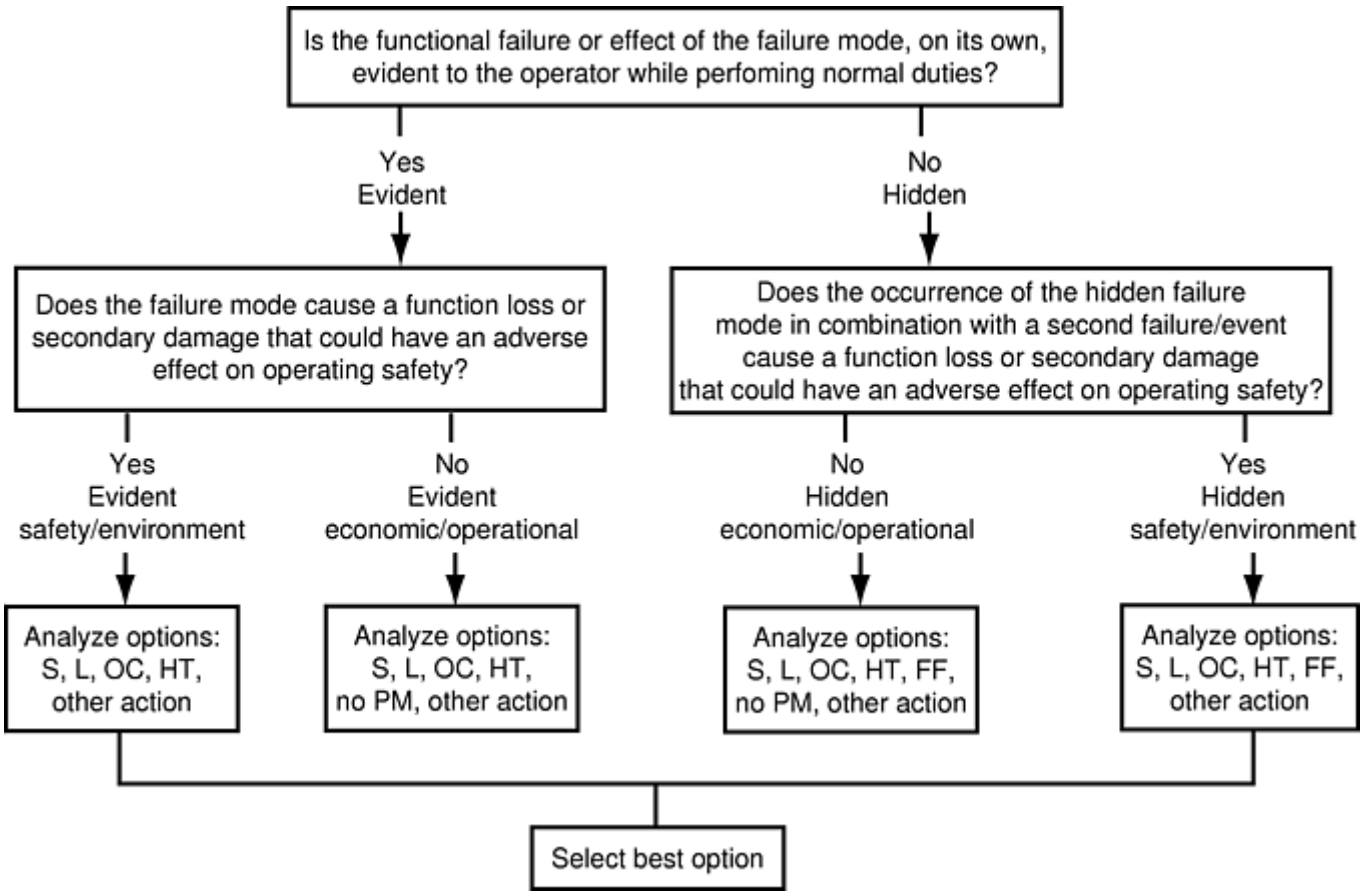


Fig. 1 Typical RCM decision logic diagram. S, servicing; L, lubrication; OC, on-condition; HT, hard-time (comprises scheduled restoration and scheduled discard); FF, failure-finding; PM, preventive maintenance. Note: S, L, and HT are aviation-unique terms and are not standard RCM terms.

It then examines the consequences of the failure mode, determining whether the consequences are safety, the environment, operational, or nonoperational (which U.S. Naval Aviation calls “economic”). This determination makes the specific worth-doing criteria available when considering each failure management policy.

In the next step, the process examines each failure management policy in turn, to determine whether the policy is technically feasible (considering the technical characteristics of the failure mode and the technical characteristics of the policy) and whether it is worth doing (considering the consequences of the failure mode and the consequences of executing the policy). (The abbreviations used to name the failure management policies in Fig. 1 reflect the slightly different vocabulary used in the RCM program of U.S. Naval Aviation.)

In the final step, the process that uses Fig. 1 selects the most cost-effective policy that is both technically feasible and worth doing.

While SAE JA1011 supports such a process, it does not require it. Indeed, most practical RCM processes stop considering failure management policies for a failure mode when the review identifies the first policy that is both technically feasible and worth doing. This is reflected in their use of decision logic diagrams in which types of failure management policies are considered in turn until a suitable policy is identified.

The decision logic diagrams used by different RCM processes show this policy examination in different ways. Figure 1 simply has a box labeled “Analyze options.” Diagrams developed for some other RCM processes spell out the steps of the analysis, that is, which policies to examine first and whether to continue looking after a policy has been found to be both technically feasible and worth doing. Some diagrams even provide a summary of the specific criteria for technically feasible and worth doing for each kind of consequence and each kind of failure management policy. The choice of a specific decision logic diagram is usually part of the choice of a specific RCM process and rests on assumptions about the level of RCM expertise held by those who are expected to use the process, because highly expert people probably need fewer hints.

When this step is concluded, the RCM review will have answered all seven questions listed earlier in “Overview of the RCM Process.” It will have a list of the functions of the asset, its functional failures, its failure modes, and the effects of those failures. It will also have a list of the failure management policies that are technically feasible for those failure modes and that are worth doing, considering the consequences of those failure modes. In many cases, the review also

gathers a wealth of supporting technical information relating to the behavior of the asset—information that is often very useful for other purposes, such as training or troubleshooting.

This concludes the discussion of the RCM process itself. However, the success of a RCM program does not depend only on technical proficiency in using the RCM process. It depends on other matters as well that are outside the scope of SAE JA1011. They are the subject of the next section.

References cited in this section

1. "Evaluation Criteria for Reliability-Centered Maintenance (RCM) Processes," JA1011, Society of Automotive Engineers, 1999

8. "Guidelines for the Naval Aviation Reliability-Centered Maintenance Process," Management Manual NAVAIR 00-25-403, Commander, Naval Air Systems Command, 2001

Managing and Resourcing the RCM Process

In this article, repeated reference has been made to SAE JA1011 and to RCM processes. In practice, several existing processes conform to SAE JA1011 (and are RCM processes) without being identical processes. By and large, they differ from each other in the way that they address issues outside the scope of the SAE standard, which focuses only on the technical issues relating to RCM. These additional issues pertain to the challenge of managing and resourcing the RCM process. Chief among them are:

- Prioritizing assets and establishing objectives
- Planning
- Level of analysis and asset boundaries
- Technical documentation
- Organization
- Training
- Role of computer software
- Data gathering
- Implementation

Each process offers an approach to these issues that is intended to enable an organization to perform the technical process of RCM in the most efficient way. Many organizations find the benefits of RCM so valuable that they wish to apply it to many of their assets, and as the scope of RCM expands, the efficiency of the RCM process becomes increasingly important. If an organization cannot use the process efficiently in its sites, it will find that RCM itself is not worth doing; that is, its benefits do not justify the effort needed to produce them. In practice, this has sometimes led to the collapse of the RCM program in the organization.

Therefore, when setting up a RCM program, it is important to keep in mind the following considerations.

Prioritizing Assets. If RCM is to be applied to a large number of assets at a site, the assets will have to be reviewed in an order or sequence. Some RCM processes incorporate additional steps intended to help assign different priorities to different assets. The usefulness of such additional steps is the subject of debate within the RCM community. Its proponents argue that they help produce clear thinking and clear decisions. Its opponents argue that the additional steps divert time and attention from the most important thing, the RCM reviews themselves. The most important thing, they argue, is to start doing RCM reviews on the assets that are clearly the most important.

Establishing Objectives. There is no universal objective for a RCM review. The objective of a review will depend on the deficiencies currently perceived in the operation and maintenance of the asset. The deficiencies may be related to the performance of the asset itself (safety hazards, interrupted operations, expensive repairs), or they may be related to the performance of the work done to keep the asset running (problems with training, with scheduling work, and so forth).

Because the objectives differ with each RCM review, it is prudent to identify the objectives of each RCM review on a case-by-case basis.

Planning. The RCM process and its supporting processes (discussed in this section) are complex and not likely to be accomplished efficiently without an explicit and focused plan that takes into account all of the processes and resources involved.

Level of Analysis and Asset Boundaries. Failing to find the right level of detail for analysis of failure modes can have a serious effect on the efficiency and effectiveness of a RCM review. Too high a level of detail can prevent the review from identifying all reasonably likely failure modes. Too low a level of detail can force the review to consider the same failure mode many times, thus wasting time and effort. It also makes it more difficult to identify functions and the consequences of failures.

Different RCM processes have different methods for identifying the best level of detail. Some always go straight to the level of the lowest replaceable component. Others examine everything at the system or major equipment level. Still others try to be yet more flexible. Before selecting a process, the owner or user of the asset should examine and approve the method used by that process for identifying the best level of detail.

Sometimes the best level of analysis on a large asset would produce more failure modes than a RCM review can handle in a reasonable amount of time. In such an event, it is appropriate to divide the asset into subsystems and analyze each subsystem separately. When dividing such an asset, it is important to ensure that the boundaries of each review are identified clearly and that no components are overlooked.

Technical Documentation. Existing documentation about the asset under review should certainly be gathered beforehand. Some RCM processes develop specific documentation, if not already available, such as functional block diagrams. This practice is another subject of debate within the RCM community. Its proponents argue that specific documentation always improves the clarity of discussion about the asset. Its opponents argue that it is only needed occasionally and should only be developed when needed.

Organization. An essential part of the planning required to support RCM is the organization of people who will support it. This organization must:

- Ensure that the RCM process to be used complies with SAE JA1011 and plan each RCM review
- Ensure the RCM reviews are executed as planned
- Apply the process during each review
- Provide information and assist in decision-making (by operators, maintainers, management representatives, and (on a case-by-case basis) designer/vendor representatives)
- Provide physical facilities, such as offices, meeting rooms, and computer hardware and software

One fundamental choice of organization is the option of using individual analysts or using a review team with a facilitator. Most U.S. military RCM programs were organized in the late 1970s and early 1980s, and many still rely on individual analysts who are expected to interview technical experts as-needed. Most commercial RCM programs were organized after the shift to teamwork in the 1990s, and most rely on review teams with a facilitator.

The proponents of individual analysts argue that that approach requires less time from technical experts. The proponents of teams argue that team meetings can be a more efficient way to assemble technical information, if the meetings are managed well, and that they also increase the buy-in of the people who are likely to be the leaders of those who will actually perform the recommended actions. Of course, meetings are only efficient if they are well organized and well managed, so the skill of the meeting facilitator is crucial to their success.

Training. It should be clear that RCM is different from other disciplines bearing the word “reliability” in their names. It is different enough to require specialized training. Because RCM is not routinely offered in the curriculum of undergraduate engineering programs or technical training programs, this training usually must be provided as part of any initiative to begin or support a specific RCM program at a specific site.

Some people in a RCM program require more training than others. A short orientation presentation can be enough for senior managers. Experience has shown that those who need to understand the details of the technique need at least three days of formal classroom training in order to absorb the concepts. Those who will take the lead in applying the process, such as meeting facilitators or individual analysts, require still more training, including on-site mentoring, before they will possess the necessary skills.

Role of Computer Software. Computer software can serve a useful role in recording the information and decisions that accumulate during the RCM review. If a large number of assets are to be analyzed, a computer may be the only practical way to store and manipulate these findings for later use.

Computer software can also perform some of the mathematical and statistical calculations in RCM, such as the cost comparisons that support the determination whether a task is worth doing for a failure mode with economic consequences, or the statistical calculations that identify the best task interval for a failure-finding task. However, computer software also brings two risks to a new RCM program.

The first risk is that people's attention might be diverted to the software and the tasks of populating its database and away from the real needs of the asset being reviewed. This approach can reduce RCM to a mechanical process that adds little real value, because the source of the value of RCM is its effectiveness in eliciting information about the asset and organizing people's thoughts about that information. As RCM expert John Moubray puts it, “RCM is thoughtware, not software” (Ref 6).

The second risk is that the ability of the computer to quickly perform calculations will tempt the software designer to incorporate algorithms based on processes other than RCM. Algorithms that have tempted RCM software designers include:

- Best-replacement-interval algorithms, which balance the cost of repairing failures (based on a given failure rate) against the cost of replacing an item periodically
- Condition-monitoring algorithms that combine condition trend data with age data, triggering repairs based either on evidence of a potential failure or on age

The best-replacement-interval algorithm works only if scheduled restoration or scheduled discard is the most appropriate failure management policy. It adds value (compared to simply performing the task shortly before the wearout age) only if there is a dependable historical record of failures (and their ages). Routine reliance on this algorithm rests on the assumption that both conditions are usually met.

Practical RCM programs do not accept this assumption. One of the key insights of the early efforts in the development of RCM was the discovery that scheduled restoration and scheduled discard tasks are almost never appropriate. In addition, a historical record of failures is almost never available, in practice, especially for the most important failures (which tend to be the most infrequent failures), making mathematical predictions of the cost of replacements versus age too imprecise to make such an algorithm practical.

The combination condition-monitoring algorithms are based on a common practice in vibration analysis: the use of a chart with warning and danger levels that vary with the age of the asset. As the asset ages, the levels that trigger warning and danger alerts on these charts get lower. If the asset survives long enough, these charts will force a restoration or discard task even if the condition of the asset is unchanged or improving.

These charts are only appropriate in cases where an on-condition task must be combined with a scheduled restoration or scheduled discard task in order to create an appropriate task. In practice, this happens about as often as cases where a scheduled restoration or discard task is appropriate on its own—again, almost never.

The standard SAE JA1011 has two specific requirements aimed at such risks. First, it says that the failure management policy selection process in a RCM process takes account of the fact that the conditional probability of some failure modes increases with age (or exposure to stress), that the conditional probability of failure of others does not change with age, and that the conditional probability of failure of yet others decreases with age (Ref 1).

Thus, a process that assumes that all or nearly all assets wear out (and that scheduled restoration or scheduled discard is always or nearly always technically feasible) is not a RCM process.

Second, SAE JA1011 states that the formulae used to support a RCM process are “logically robust,” which means that they are consistent with the expectations of RCM about the behavior and deterioration of physical assets (Ref 1). Clearly, algorithms with formulae that expect most physical assets to wear out over time do not comply with this requirement.

Data Gathering. A RCM review gathers and reviews information about the asset and its operating context in order to find the most appropriate ways to manage the consequences of the failure of the asset. The various steps in a RCM review require historical data about failures, the performance of the asset (including how the asset degrades over time), associated maintenance and operating costs, and the performance of scheduled maintenance. It also needs information about existing scheduled maintenance tasks and about the consequences of failure.

Sometimes, adequate information about the failure modes that may occur and how often they might occur is simply not available (especially with assets containing large amounts of new technology). When the consequences of this uncertainty are intolerable, it can be tempting to ignore the technical feasibility of tasks

and “just do something.” A better choice is to take action that changes the consequences so that the uncertainty no longer matters. For example, if the design of the asset makes it difficult to be sure of detecting warning signs that the failure mode is taking place, a different design might eliminate that problem. If the reliability of new technology is not known precisely enough to be able to assess whether a scheduled task is worth doing, the installation of a backup unit (which will change the nature of the consequences to hidden) may make it possible to address failures through a failure-finding task.

Once this information has been gathered for the review, it should be kept current so that the recommendations of the review can be quickly updated later. The owners of the asset should avoid the common mistake of recording only the repair work performed after a failure and should place equal emphasis on recording the causes of functional failure and the associated consequences (such as equipment downtime).

Implementation. After the RCM review has been completed (or updated), its results must be implemented. Organizations that fail to implement results fail to reap any benefits from their review. Successful implementation involves three key steps:

- Obtain management support for the results. This entails auditing the results to ensure that they are technically valid and that they meet the goals of the management. It also entails presenting the results to management and inviting their support for implementation.
- Plan the implementation of the results. This entails writing descriptions of the scheduled tasks and one-time changes in enough detail to ensure that the work will be done correctly by the people who will do it.
- Execute the implementation. Make the one-time changes and plan and execute the scheduled tasks.

These management and planning activities are outside the scope of the RCM process itself, so those who wish to learn more about them should look for references about maintenance management and maintenance planning.

References cited in this section

1. “Evaluation Criteria for Reliability-Centered Maintenance (RCM) Processes,” JA1011, Society of Automotive Engineers, 1999

6. J.M. Moubray, *Reliability-Centered Maintenance,* Industrial Press, 1999, p 155, 290

Conclusions

From its origins in the quandaries of airline maintenance managers in the late 1950s and early 1960s to its worldwide use in nearly every industry on the face of the earth today, RCM has grown into a formidably powerful tool for identifying the causes of asset failures and the most appropriate ways to manage their consequences before the failures occur, thus truly preventing the intolerable consequences of failures outright.

In its systematic examination of failures, failure causes (or failure modes), their consequences, and the failure management policies available in a given operating context, RCM keeps in mind these key points:

- Most assets do not wear out at a consistent age; whereas, a considerable number of assets give warning signs before they fail. As a result, RCM always examines the suitability of on-condition maintenance (checking for signs of potential failure), whether or not it examines the suitability of scheduled restoration or scheduled discard tasks.
- Not every failure needs to be prevented. The consequences of preventing the failure may be worse than the consequences of accepting the failure, especially if the asset tends to be less reliable right after it is worked on (a very common experience). As a result, RCM always checks to see whether a scheduled task is worth doing, even if the task has already been shown to be technically feasible.
- The key to solving reliability problems lies in finding the cause of the failure, not merely in finding the mode or manner of failure. A functional failure is the last event in a chain of events. The goal of failure prevention (and of RCM) is to identify the right point at which to intervene and break the chain.

- Further, the key to finding the cause of the failure is understanding the level of detail required in the current context. In the design context, "component A3 breaks" may be a satisfactory "cause of the failure of asset A," and a satisfactory solution may be to specify a more reliable version of component A3. However, in an in-service context, where equipment is out of warranty and fully paid for, a design change may be difficult to justify in any but the most urgent cases. In such a context, a deeper level of detail may be required in order to identify an appropriate scheduled maintenance task that will be just as technically feasible as a design change but more cost-effective (and thus more likely to be approved for implementation). Reliability-centered maintenance always examines the suitability of scheduled tasks, whether or not it examines the suitability of design changes.
- If the consequences of the failure are intolerable and scheduled tasks cannot make them tolerable, the effort must not stop there. One-time changes must be examined until the consequences can be made tolerable. (In rare cases, the only appropriate one-time change may be to cease operating the dangerous asset entirely.) Reliability-centered maintenance is aimed at the reliability of the asset, not simply the maintenance program of the asset.
- Hidden failures matter, too. Backup and safety equipment are often ignored until they are needed, at which point it may be too late to fix problems. As a result, RCM explicitly asks whether a failure is hidden. If it is, RCM adds failure-finding tasks to the list of options available.

Through these questions, along with the supporting criteria that apply these points, RCM enabled U.S. and Canadian airlines to reduce the crash rate of their jet airliners from more than 30 per million take-offs to less than one—a reduction of more than 3,000%. It has enabled international industry to improve the reliability and availability of their industrial assets on sites around the world, in nearly every field that depends on having its assets being reliable and available for use. It remains the most thorough and flexible tool in the world for identifying and addressing failures before they happen—for truly preventing the failures that matter.

References

1. "Evaluation Criteria for Reliability-Centered Maintenance (RCM) Processes," JA1011, Society of Automotive Engineers, 1999

2. F.S. Nowlan and H.F. Heap, "Reliability-Centered Maintenance," AD-A066 579, National Technical Information Service, 1978, Appendix B, p 46, 386

3. *Statistical Summary of Commercial Jet Airplane Accidents: Worldwide Operations, 1959–2000,* Boeing Commercial Airplanes, 2001, p 4, 7, 15

4. *Aviation Industry Overview—Fiscal Year 2000,* Federal Aviation Administration, 2001, p 5

5. *Federal Aviation Administration Certification Procedures,* Federal Aviation Administration, May 12, 1972

6. J.M. Moubray, *Reliability-Centered Maintenance,* Industrial Press, 1999, p 155, 290

7. G.C. Hall, Jr., *Death Squadron,* Kensington Publishing Corp., 1946, p 178, 183–84

8. "Guidelines for the Naval Aviation Reliability-Centered Maintenance Process," Management Manual NAVAIR 00-25-403, Commander, Naval Air Systems Command, 2001

Products Liability and Design

Charles O. Smith, Engineering Consultant

Introduction

PRODUCTS LIABILITY is a legal term for the action whereby an injured party (plaintiff) seeks to recover damages for personal injury or property loss from a producer and/or seller (defendant) when the plaintiff alleges that a defective product caused the injury or loss.

If a products liability suit is entered against a company, the plaintiff's attorney and technical experts attempt to convince a jury that the manufacturer did not exercise reasonable care in one or more features of design and/or manufacture, and that because the company did not exercise reasonable care, an innocent party was injured. The defendant's team attempts to convince a jury that the manufacturer was not responsible for the injury.

Products liability is not new. The first law code known to be in writing was established by Hammurabi, King of Babylon, about 4000 years ago, and it contained clauses that clearly relate to products liability (Ref 1).

Who may be a plaintiff? Essentially any consumer, user, or bystander may seek to recover for injury or damages caused by a defective and unreasonably dangerous product. Who may be a defendant? Any corporation, business organization, or individual who has some degree of responsibility in the “chain of commerce” for a given product, from its inception as an idea or concept to its purchase and use.

The situation is schematically summarized in Fig. 1. The editors believe it is important for some general background on the legal aspects of failure analysis to be included in this Volume. This article is not intended to provide legal conclusions, and the opinions are those of the author. Because every situation is different, this article should not be applied to any specific product without additional analysis. It is written from an engineering perspective, and the author is not a lawyer. Designers may wish to consult with counsel to better understand what the law requires of them.

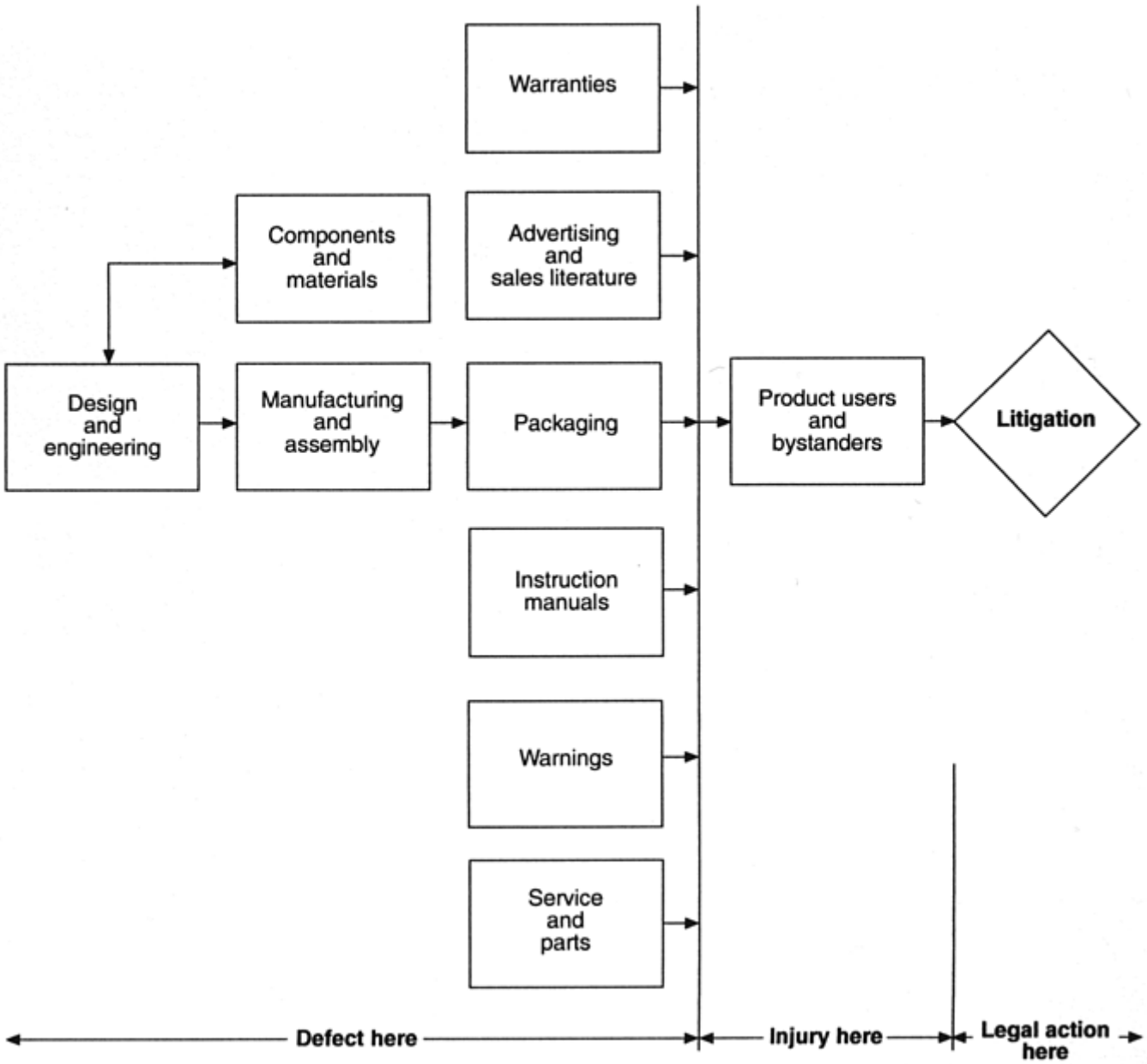


Fig. 1 The essence of products liability

Reference cited in this section

1. *The Code of Hammurabi,* University of Chicago Press, 1904

Legal Bases for Products Liability

The three legal theories on which a products liability lawsuit can be based are negligence, breach of warranty, or strict liability. All three are predicated on the fault system (i.e., a person whose conduct causes injury to another is required to fully and fairly compensate the injured party).

The basic method of imposing liability on a defendant requires the plaintiff to prove that the defendant acted in a negligent manner. Under the negligence theory, the plaintiff must essentially establish proof of specific negligence (i.e., prove that the defendant was almost intentionally negligent). Proof of specific negligence is a difficult task.

A user of a product may, as a result of express oral or written statements, or implication, reasonably rely on the manufacturer's express or implied assurance (including advertising material) as to the quality, condition, and merchantability of goods, and as to the safety of using them for their intended purpose and use. If the user relies on these assurances and is injured, suit can be entered on the basis of breach of warranty.

Both negligence and breach of warranty require proof of some fault on the part of the defendant (i.e., the focus is on the action of an individual). Strict liability, however, focuses on the product itself (Ref 2):

1. One who sells any product in a defective condition unreasonably dangerous to the user or consumer or to his property is subject to liability for physical harm thereby caused to the ultimate user or consumer, or to his property, if (a) the seller is engaged in the business of selling such a product and (b) it is expected to and does reach the user or consumer without substantial change in the condition in which it is sold.
2. The rule stated in subsection (1) applies although (a) the seller has exercised all possible care in the preparation and sale of his product, and (b) the user or consumer has not bought the product from or entered into any contractual relation with the seller.

Although these three bases apply in general, it should be recognized that there are variations. Interpretation and application may vary greatly, depending on the jurisdiction.

Reference cited in this section

2. *Restatement of the Law, Second, Torts,* 2d, Vol 2, American Law Institute Publishers, 1965

Hazard, Risk, and Danger

There is substantial confusion about the meaning of words such as hazard, risk, and danger. Webster (Ref 3) defines danger as: "liability to injury, pain, damage or loss; hazard; peril; risk." Webster makes some distinction by further saying: "Hazard arises from something fortuitous or beyond our control. Risk is doubtful or uncertain danger, often incurred voluntarily."

One can also consider a hazard as any aspect of technology or activity that produces risk; as the potential for harm or damage to people, property, or the environment; and as including the characteristics of things and the actions (or inactions) of individuals. One can also consider risk as a measure of the probability and severity of adverse effects.

With all the products liability litigation in the United States, there has developed a clear distinction among these three words for legal purposes. In this context, a hazard is a condition or changing set of circumstances that present an injury potential (e.g., a railroad crossing at grade, a toxic chemical, a sharp knife, the jaws of a power press). Risk is the probability of injury and is affected by proximity, exposure, noise, light, experience, attention arresters, intelligence of an involved individual, and so on. Risk (probability of exposure) is obviously much higher with a consumer product than with an industrial product used by trained workers in a shop environment. Danger is the unreasonable or unacceptable combination of hazard and risk. The U.S. courts generally hold that any risk that can be eliminated by reasonable accident-prevention methods is unreasonable and unacceptable. A high risk of injury could be considered reasonable and acceptable if the injury is minimal and the risk is recognized by the individual concerned.

As might be expected, there is extensive and ongoing debate over the meaning of "reasonable" and "unreasonable." The American Law Institute (Ref 2) says unreasonably dangerous means that: "The article sold must be dangerous to an extent beyond that which would be contemplated by the ordinary consumer who purchases it, with the ordinary knowledge common to the community as to its characteristics. Good whiskey is not unreasonably dangerous merely because it will make some people drunk, and is especially dangerous to alcoholics; but bad whiskey, containing a dangerous amount of fusel oil, is unreasonably dangerous." The American Law Institute (Ref 2) further says: "There are some products which, in the present state of human knowledge, are quite incapable of being made safe for their intended and ordinary use. … Such a product, properly prepared, and accompanied by proper directions and warnings, is not defective, nor is it unreasonably dangerous."

References cited in this section

2. *Restatement of the Law, Second, Torts,* 2d, Vol 2, American Law Institute Publishers, 1965

3. *Webster's New Twentieth Century Dictionary, Unabridged,* 2nd ed., Simon & Schuster, 1979

Definitions of Defects

The American Law Institute (Ref 2) says that a product is in a defective condition if "it leaves the seller's hands, in a condition not contemplated by the ultimate user, which will be unreasonably dangerous to him." Peters (Ref 4) indicates that a California Supreme Court decision, Barker v Lull (Ref 5), established a good assessment of "defective condition." This provides three definitions (or criteria) for manufacturing defects and two for design defects:

Manufacturing defects

- Nonconformance with specifications
- Nonsatisfaction of user requirements
- Deviation from the norm

Design defects

- Less safe than expected by ordinary consumer
- Excessive preventable danger

Manufacturing Defects

A failure to conform with stated specifications is an obvious manufacturing defect and not a new criterion. The aspect of user satisfaction may not be well known, but in the legal context it has long been recognized that a manufacturing defect exists when there is such a departure from some quality characteristic that the product or service does not satisfy user requirements. Under the third criterion (deviation from the norm), added by Barker, a manufacturing defect occurs when a product leaves the assembly line in a substandard condition, differs from the manufacturer's intended result, or differs from other, ostensibly identical units of the same product line.

Design Defects

A product may be considered to have a design defect if it fails to perform as safely as an ordinary consumer would expect. This failure to perform safely is interpreted in the context of intended use (or uses) in a reasonably foreseeable manner, where "foreseeable" has the same meaning as "predicted" in failure modes and effects, fault-tree, or hazard analyses. It appears that many "ordinary" consumers would have no concept of how safe a product should, or could, be without the expectations created by statements in sales material, inferences from mass media, general assumptions regarding modern technology, and faith in corporate enterprise.

A design defect also exists if there is excessive preventable risk. The real question is whether the risk outweighs the benefits. A risk-benefit analysis should include at least five factors:

- Gravity of the danger posed by the design (i.e., severity of the consequences in the event of injury or failure)
- Probability (including frequency and exposure of the failure mode) that such a danger will occur
- Technical feasibility of a safer alternative design, including possible remedies or corrective action
- Economic feasibility of these possible alternatives
- Possible adverse consequences to the product and consumer that would result from alternative designs

Additional relevant factors may be included, but design adequacy is evaluated in terms of a balance between benefits from the product and the probability of danger. Quantification of the risk-benefit analysis is not required but may be desirable.

Proving design adequacy places the burden of proof on the defendant. Once the plaintiff proves that the product is a proximate cause of injury, the defendant must prove that the benefits outweighed the risk. Discussion of manufacturing and design defects of various products is given in Ref 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24.

Note: No paper, book, or handbook relative to products liability can be truly current. In addition, there is substantial variation among jurisdictions (federal, state, and local). All cited publications, however, do have something that is currently pertinent.

Other Defects

The engineer must be alert for other possibilities. Smith and Talbot (Ref 25) point out that a marketing defect exists when there is a failure to provide any warning of hazard and risk involved with use of a product, provide adequate warning of hazard and risk involved with use of a product, or provide appropriate, adequate directions for safe use of a product. In other words, a marketing defect exists when a product, free of design and manufacturing defects, is unreasonably dangerous due to absence of warnings and directions. The designer/manufacturer has control over the directions and warnings provided. The designer/manufacturer is the most knowledgeable about the product and thus presumably the most able to determine the necessary directions and warnings.

Suits against manufacturers often allege a defective product. Careful investigation, however, sometimes shows that the problem is due to improper maintenance (e.g., Ref 26). The designer/manufacturer obviously has no control over the maintenance actually conducted but can try to minimize the possibility of improper practices by providing proper and adequate instructions with the product. In any event, the designer should not overlook the possibility of misuse and improper maintenance on the part of the user.

Example 1: Failure of a High-Speed Steel Twistdrill. A 1.905 cm (0.75 in.) stud broke in the vertical wall of a metalworking machine known as an ***upsetter.*** A parallel vertical wall left a limited amount of space in which mechanics could work. A pilot hole was drilled with a 0.476 cm (0.19 in.) drill. The drill was held in a Jacobs chuck in a portable drill press that, in turn, was held to the workpiece by an electromagnet. After the pilot hole was finished, the drill press was removed, the pilot drill was replaced by a 1.6 cm (0.63 in.) high-speed steel drill, and the press was repositioned.

One man was doing the drilling while another man was squirting oil into the hole. When the drill was about 1.27 to 1.90 cm (0.50 to 0.75 in.) into the pilot hole, there was a “bang.” The drill shattered, causing a chip to lodge in the right eye of the oiler, ultimately resulting in loss of vision in that eye. Suit was entered against the drill manufacturer alleging a defective drill.

The plaintiff's attorney retained a metallurgist who examined the fragments. An unetched longitudinal section showed a large nonmetallic inclusion parallel to the axis near the center of the drill. After etching, this section showed carbide bands in a martensitic matrix. Hardness measurements indicated 65 to 66 HRC at the edge of the flute with a bulk hardness of 62 to 64 HRC. The drill tip is shown in Fig. 2.

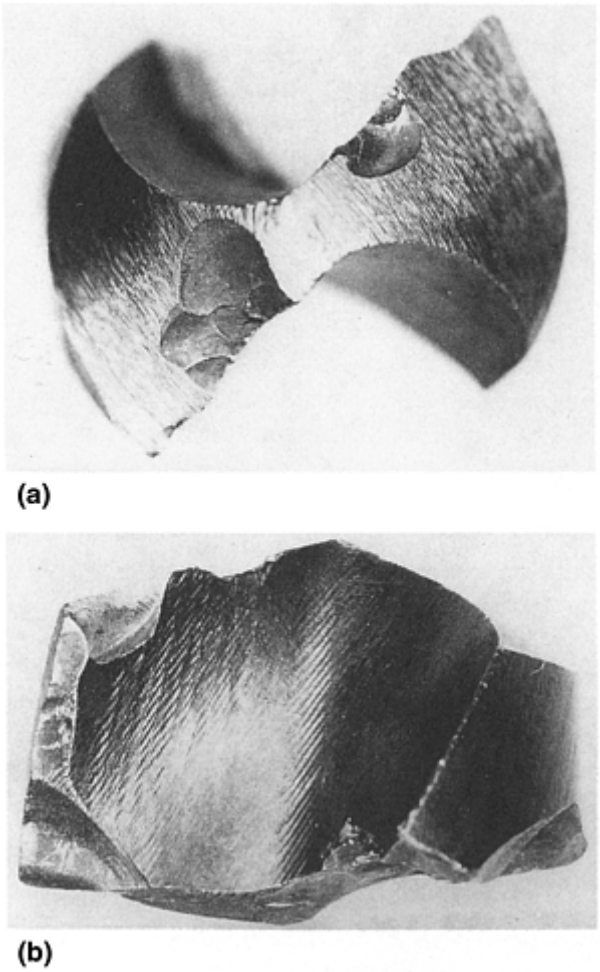


Fig. 2 The tip of the broken twistdrill. (a) End view. (b) Top view. A second drill from the same lot was also examined. There was carbide banding but to a significantly lesser

degree. Hardness was measured as 63 to 64 HRC at both the flute edge and in the bulk of the drill.

Plaintiff's expert concluded that the failed drill was defective while the other drill was satisfactory. He claimed that failure was a cumulative result of the following defective conditions: the steel contained nonmetallic inclusions that were detrimental to the properties of the drill; the carbide segregation was excessive, causing the drill to be brittle; and the cutting edge of the flutes was excessively hard. In his opinion, this high hardness made the edge brittle, so that the edge would chip during drilling. The chips caused the drill to bind and then shatter in a brittle manner because of excessive carbide segregation.

Plaintiff alleged defective design, defective manufacture, unsuitable or defective material, lack of sufficient quality control, and failure to foresee.

The defendant manufacturer believed the twist drill met all specifications for M1 high-speed steel. Both the supplier and manufacturer inspected for carbide segregation, with the poorest rating being “slight to medium.” A “medium” rating was permitted. Heat treatment and nitriding practices were consistent with those published by ASM International. After heat treatment, the drills were within the specified range of 64 to 66 HRC.

Some twenty other inspections for dimensional accuracy, shape, and finish were made after heat treatment. Fourteen drills were given a severe drilling performance test (manufacturer's routine) with no breaking or chipping.

It could be argued that there was nonsatisfaction of user requirements. The counter argument was that there was too high a demand on the drill.

Conclusions by the plaintiff's experts can be viewed as indicating a deviation from the norm. (Of the 5360 drills made from this one lot of steel, the manufacturer received only this one complaint.) The observed nonmetallic inclusion had a maximum width less than 0.0127 mm (0.0005 in.) and a maximum length less than 0.84 mm (0.033 in.). It was located in the shank, more than 6.35 cm (2.5 in.) from the drill tip and along the central axis, which is subjected to essentially none of the bending and twisting loading. While the inclusion is relatively large, it is not likely that it could contribute to the failure.

The largest carbide band was about 8 mm (0.32 in.) long and located about one-fourth of the distance from the central axis to the outer edge. It was also some distance from the drill tip. This location implies relatively light loading.

The manufacturer made hardness measurements on the two drills examined by plaintiff's expert. The results are given in Table 1. These indicate no significant difference between the two drills. The higher hardness at the cutting edge is expected and reasonable for a nitrided M1 steel. The hardness of both drills is within normal specification ranges.

Table 1 Hardness examination of drills

See Example 1 in text.

Measurement	Average hardness, HRC[(a)]	
	Bulk	Cutting edge
Plaintiff's measurements		
Broken drill	62–64	65–65
Unbroken drill	63–64	63–64
Defendant's measurements		
Broken drill	64.9, 64.8	65.1[(b)], 66.5[(c)]
Unbroken drill	65.0, 65.1	65.6[(b)], 66.5[(c)]

(a) Tukon readings (100 g) converted to HRC.

(b) At 0.1270 mm (0.005 in.) from the surface.

(c) At 0.254 mm (0.010 in.) from the surface

The manufacturer examined a number of other M1 drills that had satisfactorily met (corporate) standard drilling tests. One of these had a nonmetallic inclusion 1.5 times longer than in the failed drill. Two had edge hardnesses in the 66 to 68 HRC range with carbide banding more pronounced than in the failed drill.

From the viewpoint of design defects, was the drill less safe than expected by the ordinary consumer? Maybe. Presumably the workers did not expect drill failure. It is well known, however, that twistdrills do fail, no matter how well designed and manufactured. Using a drill to remove material after drilling a pilot hole is a common practice and clearly foreseeable. It is clearly more hazardous than drilling without a pilot hole. The drill may have been less safe than expected, but it seems more credible that too much was expected.

Existence of a design defect related to “excessive preventable danger” seems doubtful. The drill design was highly similar to that used by other manufacturers. All dimensions, tolerances, clearances, and so on were consistent with those used by other manufacturers and were based on years of drill use by a great variety of users. There is no question of potentially severe damage and relatively high probability of exposure. But there are no apparent alternatives that are technically and/or economically feasible.

What is your judgment on the validity of the allegations? How should this litigation have been resolved?
Examination of Fig. 2 indicates that one cutting lip is about 0.725 cm (0.286 in.) long while the other is about 0.802 cm (0.316 in.) long, so that the chisel edge is about 2.7 mm (0.106 in.) off center. The shorter lip will contact the work before the longer lip and thus bears all of the initial drilling stresses. The larger of the two chipped areas along the cutting edges in Fig. 2 is on the shorter lip. The broken point also had improper clearance angles (one was close to a negative angle). It was clear that the point of the broken drill was not the original point put on at manufacture but came from regrinding (presumably from “eyeballing” rather than using a jig). The work conditions, including a small pilot hole, a portable drill press, relocation of the press between the two drilling operations, and a questionable supply of coolant, placed abnormal stress on the drill.
The case was eventually settled out of court, with the plaintiff receiving a sum of less than $10,000 at a time when similar injury cases were receiving judgments of $50,000 to $150,000. This was clearly a so-called “nuisance settlement” to get rid of the suit. Much greater detail, both technical and legal, can be obtained by reference to Ref 27, 28, 29. A side aspect of this case relates to the expert witness. It developed that the plaintiff's expert was not sufficiently knowledgeable about high-speed steels, although he was a competent metallurgist.
Finally, it should be noted that this accident could have been avoided if the victim had been wearing safety glasses. The possibility of eye injury when working with tools is well known. Manufacturers usually add a warning to the product itself or its packaging. Employers require both workers and visitors to wear eye protection and provide training on its importance. Eye protection, as well as other personal protective equipment, is also required by both federal and state laws and regulations governing workplace safety.

Human Factors

Two additional examples arising from litigation are given subsequently. Designers certainly must consider the hazards in the design when it is used or operated in the intended manner. A hazard is any aspect of technology or activity that produces risk or danger of injury. The designer must also recognize that the product may be used in unintended but foreseeable ways. Protection must be provided against hazards in all uses that can be foreseen by the designer. Unfortunately, even the most diligent search for foreseeable uses may still leave a mode of use undiscovered. In litigation, a key issue often revolves around the question of whether the specific use was foreseeable by a reasonably diligent designer.
Example 2: A Snowmobile Collision. A female teenager and her boyfriend rented a snowmobile for a few hours to ride around on a lake. He was driving; she was a passenger. They collided with another snowmobile rented from the same rental agency. She was thrown on the ice, injuring her left knee. Surgery was performed but left her with permanent injury.
The owner of the agency admitted that the brakes had been disconnected by the previous owner of the agency (while he was an employee) and that he had left them disconnected. The reason given was that renters became confused about the operation of the throttle and the brakes, with the result that they “tore up the equipment.” Without brakes, the only mechanism for slowing the snowmobile was to release the throttle and let the engine provide braking. The owner claimed that the throttle had been adjusted to limit the speed to between 20 and 25 mph.
The owner's manual read: “The squeeze-type throttle lever, conveniently located on the right side of the steering handle, enables the operator to control the sled and the engine rpm at the same time.” The manual further read: “The hand-operated brake lever is located on the left side of the steering handle. By mounting the brake and throttle levers on the steering handle, the operator is able to maintain complete control of his snowmobile using his hands only.”
The throttle was opened (to increase speed) by closing (i.e., squeezing) the right-hand lever. The brakes were applied (to decrease speed) by closing (i.e., squeezing) the left-hand lever. When the throttle was released, the engine slowed. At some lower (undetermined) speed, the clutch disengaged and the tracks stopped moving, thereby generating significant drag.
As an exercise, this design is analyzed from the human factors viewpoint. If changes would improve the design, what changes would be made?
Most people immediately see this design as very poor from the viewpoint of human factors. Most operators never see the owner's manual, but in this context, that is minor. The natural (automatic) thing for people is that both the left and right hands will do the same thing simultaneously (i.e., both will open or both will squeeze), unless deliberate thought is given to doing otherwise. When a collision is imminent, there is no time for conscious thought, only conditioned reflex. Only well-experienced operators might be expected to have the left hand squeeze while the right hand opens.
It is not necessary, moreover, to have the design that was used. Reversing the action of the brake lever would be a better design (i.e., both hands squeeze to move forward; both hands release to stop). This “deadman” arrangement is common in railway locomotives, power movers, snowblowers, and so on. This arrangement is technically and economically feasible. The argument that keeping the left hand closed is undesirable is weak, because the right hand must be kept closed while the snowmobile is operating.

Another, perhaps superior, alternative is found on many motorcycles and would be familiar to many people. The throttle is operated by rotating a grip on the steering handle. The brake is operated by a squeeze lever on the same side of the steering handle. The two cannot be operated simultaneously. This alternative is technically and economically feasible, with no foreseeable adverse effects.

Example 3: A Ladder Label. Figure 3 shows a full-scale copy of a label “permanently attached” (by some sort of adhesive) to the inside of a rail on a fiberglass ladder. How good is this label? How effective is it? Assuming that users do indeed see the label, how many will read it? Of those who read it, how many will really comprehend what the manufacturer is trying to say? It appears to the author that the label was not well thought out, either in content or in phrasing, which is unclear or ambiguous in many places. The label does not provide clear instructions on use or explicitly warn of the dangers. Thus, in the author's opinion, the label is clearly inferior and essentially ineffective. One might infer that the manufacturer was trying to cover all possibilities to provide “protection” against products liability suits. It may be ineffective in that respect as well. Perhaps the ladder manufacturer is responding to jury awards in which a manufacturer was found negligent for not including a specific warning against some misuse.

FIBERGLASS
SINGLE & EXTENSION LADDER
FOR SAFETY, READ CAREFULLY

INSPECTION

1. Inspect upon receipt and before use.
2. Never climb a damaged ladder. Return for repair or discard.
3. Check all working parts, rivets, bolts, rope and cable for good working order.
4. Never use ladder with missing parts.
5. Discard if exposed to fire or chemicals.

SELECTION

1. Use 300 lb., and 200 lb. Duty-Rated Ladder for maintenance and heavy-duty work. Never use ladder jacks on 200 lb. or 225 lb. Duty-Rated Ladders.
2. Use ladder with correct duty rating to support combined weight of the user and material. Ladders are available with duty ratings of 200, 225, 250, 300 lb.

SET-UP AND USE

1. Set up ladder at 75½° by placing bottom ¼ of length being used out from vertical resting point.
2. Set ladder on firm level ground. Never lean sideways and never use on ice or snow.
3. Use proper size ladder. Never use temporary supports to increase length or to adjust for uneven surfaces.
4. Keep rungs free from wet paint, mud, snow, grease, or other slippery material.
5. Extend only from ground. Never extend from top or by bouncing.
6. Never walk or jog ladder while on it.
7. Securely engage ladder locks before climbing.
8. Erect ladder with fly (upper) section above and resting on base (lower) section.
9. Each section of a multi-section ladder shall overlap the adjacent section by 3 ft. up to and including 36 ft.; by 4 ft. over 36 ft., up to and including 48 ft.; by 5 ft. over 48 ft., up to and including 60 ft.
10. Always have the four ends of the ladder rails firmly supported.
11. Always tie top and base to building.
12. Project ladder minimum of 3 feet above roof edge.
13. Tie down ladder before stepping onto roof.
14. Never over-reach. Move ladder instead. Keep belt buckle inside ladder side rails.
15. Never use in high winds.
16. Never overload. Ladder designed to support one person when properly used.
17. Never use as a horizontal platform, plank or material hoist.
18. Never use on a scaffold.
19. Never fasten different ladders together to increase length.
20. Never apply a side load to ladder to push or pull anything while on ladder.
21. Never drop or apply impact load to ladder.
22. Never sit on end of ladder rails.
23. When reassembling, properly engage all guide brackets and lock prior to use.
24. Never use in front of unlocked doors.
25. Fly section must have safety shoes if used as a single ladder.
26. Hooks may be attached at or near top for added security.
27. To support the top of a ladder at a window opening, a stabilizer should be attached to span the window.
28. Never use ladder when you are in poor health.
29. Never use if taking drugs or alcoholic beverages.
30. Recommend never using if over 65 years of age.

CLIMBING INSTRUCTIONS

1. Never climb onto ladder from the side or from one ladder to another.
2. Face ladder when ascending or descending. Maintain a firm grip and stand on middle of rung.
3. Never stand above 3rd rung from top.
4. Never climb above support point.

STORAGE

1. Support ladder on racks when stored.
2. Never store material on ladder.
3. Properly support ladder in transit.

FOR ADDITIONAL INSTRUCTIONS, SEE ANSI A14.5

Fig. 3 A black-and-white reproduction of a decalcomania label to be placed on the inside of the side rail of a fiberglass ladder. The heading was yellow lettering on black. The text lettering was black on yellow. Reproduction is 100% of original size.

At least two issues deserve comment. One is the content of the text on the label. There are 44 items under five different headings, and many of these items are somewhat ambiguous. For example, in the first three items under “Inspection,” it is not clear just what one looks for (i.e., there is no definition of terms). Under “Set-up and Use,” the first item is poorly stated. Engineers have no difficulty in understanding what is intended, but many people besides engineers use ladders. In addition, does any user ever actually measure 75° or even “of length being used”? A measurement is not required—a simple estimate of angle or length would be sufficient. Item 10 seems to conflict with item 12. Item 11 says: “Always tie top and base to building.” Are ladders used only on buildings? No, but the user might think twice before using it on, for example, a tree. Does anyone ever tie a ladder to a building? If so, how? Item 30 says: “Recommend never using if over 65 years of age.” This age limitation presumably refers to the age of the user (not the ladder), but why 65 rather than some other age? One could go on at length about other items. At the bottom, the label says: “For additional instructions, see ANSI A14.5.” Will the average user have any concept of what this means or where to find a copy? Will any user, even the intelligent engineer, obtain and read American National Standards Institute (ANSI) A14.5? In any case, the manufacturer is providing a reference for those who feel they might need the additional information that a standard provides. Reference to the standard also tells the user that the manufacturer is conscious of standards and is trying to follow them. This is important to the user even if he or she does not personally obtain and read the standard. It is unclear just what “additional instructions” should be consulted in the standard. Warning of all possible foreseeable hazards and misuses in a very limited space can be difficult, but some further design of the label seems desirable.

The second problem is the type size. It does conform to ANSI Z535.4. The text of the original label had black lettering on a yellow background in keeping with ANSI Z535.1. The three-line heading was yellow on black. The text print size was 6-point type. (“Point” is a printer's measure equivalent to 0.01384 in. or essentially 72 points per in.) ANSI Z535.4 specifies type (in the text) of 1.5 mm high (minimum), or 5 points. However, do you have any difficulty in reading it? Bailey (Ref 30) notes that “type size in books and magazines usually ranges from 7 to 14 points with the majority being about 10 to 11 points. Probably the optimum size is from 9 to 11 points—sizes smaller or larger can slow reading speed.”

Anticipating Errors. When humans are involved in the use of a product or system, there will be errors. Some errors are extremely difficult, if not impossible, to anticipate. Also, in many situations, people abuse equipment. This is commonly a result of poor operating practices or lack of maintenance. In other situations, there may be deliberate action by the user (e.g., trying to fit two components together in a manner that is not intended, such as installing thread adapters on pressurized gas containers). There is no question that the designer cannot anticipate all these possibilities and provide protection. Nevertheless, the designer is not relieved of a substantial effort to anticipate such actions and to try to thwart them.

How does one proceed? The designer must be well informed on anthropometrics (physical characteristics), how people tend to behave or perform, and how to combine such data to achieve a suitable, effective, and safe design. A wealth of literature is available.

Hunter (Ref 31) includes enough anthropometric data to give insight into the kind of data to expect. He also provides many examples of sources of information. He comments on Department of Defense documents that provide substantial and significant information. The objectives of these various documents can be applied with equal validity to both civilian and military products.

Human behavior is largely a question of psychology, a topic about which most engineers know little. There seems to be little information readily available that is focused for use by engineers. Possible sources are Ref 32, 33, 34, 35, 36, 37.

Many publications provide varying degrees of insight and help in applying human factors information to design. Some that may be particularly useful are Ref 38, 39, 40.

One of the many objectives of the designer is to minimize the probability of “human error,” where human error is any personnel action inconsistent with established behavioral patterns considered to be normal or that differs from prescribed procedures. Predictable errors are those that experience shows will occur repeatedly under similar circumstances. The designer must minimize the possibility of such errors.

People have a strong tendency to follow procedures that require minimum physical and mental effort, discomfort, and/or time. Any task that conflicts with this tendency is highly likely to be modified or ignored by the person who is expected to execute the task.

One of many important considerations in design is to follow common stereotypical expectations as much as possible. Consider a few examples:

- Clockwise rotation of a rotary control (knob) is expected to increase the output.
- Moving a lever forward, upward, or to the right is expected to increase the output.
- On a vertically numbered scale, the higher numbers are expected to be at the top.

- In vehicles, depressing the accelerator is expected to increase speed, and depressing the brake is expected to decrease speed. One expects the right foot to be used to apply force to the accelerator, then moved to the brake pedal. This is true whether one drives on the right or left side of the road.

Smith (Ref 41) tells of a forklift truck that violated the fourth item. The left foot was used to depress a pedal that increased speed, and a brake was applied when the foot was lifted.

References cited in this section

2. *Restatement of the Law, Second, Torts,* 2d, Vol 2, American Law Institute Publishers, 1965

4. G.A. Peters, New Product Safety Legal Requirements, *Hazard Prevention,* Sept/Oct 1978, p 21–23

5. *Barker v Lull Engineering Co.,* 20 C. 3d 413

6. C.O. Smith, Manufacturing/Design Defects, Paper 86-WA/DE-14, American Society of Mechanical Engineers

7. C.O. Smith, Mobile Ladder Stand, Paper 87-DE-5, American Society of Mechanical Engineers

8. C.O. Smith, Design of a Saw Table, Paper 87-WA/DE-9, American Society of Mechanical Engineers

9. C.O. Smith, Coffee Grinder: Safe or Not?, Paper 88-WA/DE-6, American Society of Mechanical Engineers

10. C.O. Smith, Collapse of an Office Chair, Paper 89-WA/DE-18, American Society of Mechanical Engineers

11. C.O. Smith, Some Subtle (or Not So Subtle?) Product Defects, Paper 90-WA/DE-23, American Society of Mechanical Engineers

12. C.O. Smith, A Fatal Helicopter Crash, Paper 91-WA/DE-8, American Society of Mechanical Engineers

13. P.D. Beard and T.F. Talbot, What Determines if a Design is Safe, Paper 90-WA/DE-20, American Society of Mechanical Engineers

14. T.F. Talbot and C.S. Hartley, Failure of Fastening Devices in Pump Packing Gland Flange, Paper 89-WA/DE-12, American Society of Mechanical Engineers

15. T.F. Talbot and M. Crawford, Wire Rope Failures and How to Minimize Their Occurrence, Paper 87-DE-7, American Society of Mechanical Engineers

16. T.F. Talbot and J.H. Appleton, Dump Truck Stability, Paper 87-DE-3, American Society of Mechanical Engineers

17. T.F. Talbot, Safety for Special Purpose Machines, Paper 87-WA/DE-8, American Society of Mechanical Engineers

18. T.F. Talbot, Chain Saw Safety Features, Paper 86-WA/DE-16, American Society of Mechanical Engineers

19. T.F. Talbot, Hazards of the Airless Spray Gun, Paper 85-WA/DE-13, American Society of Mechanical Engineers

20. T.F. Talbot, Man-Lift Cable Drum Shaft Failure, Paper 87-WA/DE-19, American Society of Mechanical Engineers

21. T.F. Talbot, Bolt Failure in an Overhead Hoist, Paper 83-WA/DE-20, American Society of Mechanical Engineers

22. W.G. Ovens, Failures in Two Tubular Steel Chairs, Paper 91-WA/DE-9, American Society of Mechanical Engineers

23. J.A. Wilson, Log Loader Collapse: Failure Analysis of the Main Support Stem, Paper 89-WA/DE-13, American Society of Mechanical Engineers

24. T.A. Hunter, Design Errors and Their Consequences, Paper 89-WA/DE-14, American Society of Mechanical Engineers

25. C.O. Smith and T.F. Talbot, Product Design and Warnings, Paper 91-WA/DE-7, American Society of Mechanical Engineers

26. C.O. Smith and J.F. Radavich, Failures from Maintenance Miscues, Paper 84-DE-2, American Society of Mechanical Engineers

27. C.O. Smith, Failure of a Twistdrill, *J. Eng. Mater. Technol.,* Vol 96 (No. 2), April 1974, p 88–90

28. C.O. Smith, Legal Aspects of a Twistdrill Failure, *J. Prod. Liabil.,* Vol 3, 1979, p 247–258

29. C.O. Smith, "ECL 170, Tortured Twist Drill," Center for Case Studies in Engineering, Rose-Hulman Institute of Technology, Terre Haute, IN

30. R.W. Bailey, *Human Performance Engineering: A Guide for System Designers,* Prentice-Hall, 1982

31. T.A. Hunter, *Engineering Design for Safety,* McGraw-Hill, 1992

32. E. Grandjean, *Fitting the Task to the Man,* 4th ed., Taylor and Francis, Washington, D.C. 1988

33. B.A. Sayers, *Human Factors and Decision Making: Their Influence on Safety and Reliability,* Elsevier Science Publishers, 1988

34. K.S. Park, *Human Reliability,* Elsevier Science Publishers, 1987

35. L.S. Mark, J.S. Warren, and R.L. Huston, Ed., *Ergonomics and Human Factors,* Springer-Verlag, 1987

36. B.S. Dhillon, *Human Reliability,* Pergamon Press, 1986

37. C.D. Wickens, *Engineering Psychology and Human Performance,* 2nd ed., Harper-Collins, 1992

38. W.E. Woodson, *Human Factors Design Handbook,* McGraw-Hill, 1981

39. G. Salvendy, Ed., *Handbook of Human Factors,* Wiley, 1987

40. W.G. Ireson, Chapter 12, *Reliability Handbook,* McGraw-Hill, 1966

41. C.O. Smith, Two Industrial Products—Defective Design?, Paper 93-WA/DE-11, American Society of Mechanical Engineers

Preventive Measures

What are the implications of the previously mentioned example for the design engineer? It is necessary to look carefully at the completed design to be sure that it is indeed appropriate and that it does not incorporate problems for which proper technological solutions have existed for some time. (For example, an independent assessment by a design review board, whose members have no parental pride in the design, is highly appropriate.) In addition, there must be recognition that many, perhaps most, consumers have no objective basis to judge how safe a product should be. An engineer making a judgment about safety must understand this lack of appreciation of an appropriate safety level.

Acting as a prudent manufacturer is not enough. The focus should be on the product itself, not the reasonableness of a manufacturer's conduct. Obviously, there will be no viable lawsuits if there are no injuries

or if there are no violations of the law. Undoubtedly the best practice is to sell a well-designed, well-manufactured product. The manufacturer needs to make certain that all reasonable preventive measures have been used in the design and manufacturing process. Much evidence, however, suggests that one of Casey Stengel's comments applies in the area of preventive measures: “In many areas we have too strong a weakness.” While many preventive measures are well known to most design engineers, some comments may be appropriate, even if only in the sense of a checklist of items to be considered.

Design review is an effort, through group examination and discussion, to ensure that a product (and its components) will meet all requirements. In a design of any complexity, there is necessity for a minimum of three reviews: conceptual, interim, and final. Conceptual design reviews have a major impact on the design, while interim and final reviews have relatively less effect as the design becomes more fixed and less time is available for major design changes. It is much easier and much less expensive to include safety in the initial design than to include it retroactively.

A more sophisticated product may require several reviews during the design process. These might be: conceptual, definition, preliminary (review of initial design details), critical (or interim review, perhaps several reviews in sequence—review details of progress, safety analyses, progress in hazard elimination, etc.), prototype (review of design before building a prototype), prototype function review, and preproduction review (final review—last complete review before release of the design to production).

These periodic design reviews should review progress of the design, monitor design and development, ensure that all requirements are met, and provide feedback of information to all concerned.

A design review is conducted by an ad hoc design review board composed of materials engineers, mechanical designers, electrical designers, reliability engineers, safety engineers, packaging engineers, various other design engineers as appropriate, a management representative, a sales representative, an insurance consultant, an attorney in products liability, outside “experts” (be sure they are truly expert!), and so on. Members of the design review board should not be direct participants in day-to-day design and development of the product under review, but the engineers should have technical capability at least equal to that of the actual design team. Vendor participation is highly desirable, especially in conceptual and final design reviews.

Design review checklists should be prepared well in advance of actual board meetings. These checklists should cover all aspects of the design and expected performance, plus all phases of production and distribution. A new checklist should be developed for each new product.

It is good practice for a designer/manufacturer to have some sort of permanent review process in addition to the ad hoc board for each individual product. This permanent group should evaluate all new products, reevaluate old products, and keep current with trends, standards, and safety devices.

If properly conducted, a design review can contribute substantially to avoiding serious problems by getting the job done right the first time. Formal design review processes are effective barriers to “quick and dirty” designs based on intuition (or educated guesses) without adequate analyses.

Some Common Procedures. Many engineers and designers are familiar with such techniques and procedures as hazard analysis; failure modes and effects analysis (FMEA); failure modes, effects, and criticality analysis (FMECA); fault-tree analysis (FTA); fault hazard analysis (FHA); operating hazard analysis (OHA); use of codes, standards, and various regulatory acts; and the Occupational Safety and Health Act (OSHA). These are discussed in the article “Safety in Design” in *Materials Selection and Design*, Volume 20 of *ASM Handbook*. Some other aspects of products liability are perhaps less well known and require some comment.

Prediction methods are necessary in applying FMEA, FTA, and so on. From statistics it is possible to predict performance of a large group of similar products, but it is not possible to predict performance of any one individual item of that group. Various statistical and probabilistic techniques can be used to make predictions, but these are predicated on having good databases.

State of the Art. The meaning of the term “state of the art” should be defined for each specific product. This might be done by comparing the product to those produced by competitors, but this comparison may not be enough. A jury is not bound by negligent practices of a negligent industry, and unfortunately, in some areas, industry practices and standards are low-level consensus practices and standards. Being in step with the state of the art may not be enough—one should strive to be ahead of the state of the art if possible (i.e., better than the competitors). It is not enough to explain what was done, because the plaintiff's expert witnesses may point out what could have been done. Purely economic reasons may not be a compelling defense argument in the courtroom and should be avoided if possible.

Quality Assurance and Testing. A primary function of quality control is to feed back inspection, testing, and other data, showing designers what is happening and revealing any need for design improvement. Manufacturers should test products in various stages of development, including field service, especially if critical components or subassemblies are involved. Final tests are necessary on each individual product or on representative samples of plant output. Care must be taken that quality control is not relaxed, intentionally or unintentionally, for production expediency.

Foreseeability is a factor that requires special attention. It is necessary to determine not only how the product is intended to be used, but also every reasonably conceivable way that it can be used and misused. (Who has never used a flat-tang screwdriver for some other purpose?) All reasonable conditions of use, or misuse, that might lead to an accident should be detailed. The designer must conclusively demonstrate that the product cannot be made safer, even to prevent accidents, during use or misuse. The problem of foreseeability is one that seems especially difficult for engineers to accept.

Consumer Complaints. Data on product failures from test facilities, test laboratories, and service personnel are valuable. Each complaint should be quickly, carefully, and thoroughly investigated. An efficient reporting system can result in product corrections before large numbers of the product reach users, or a product recall before there has been a major exposure of the public to an unsafe product.

Warranties and Disclaimers. Warranties and disclaimers detail the limitations of the manufacturer's liability. When used, they must be written in clear, simple, and easily understood language. Both should be reviewed by highly competent legal counsel knowledgeable in both the industry and products liability. A copy of the warranty and/or disclaimer must be packaged with the product. All practical means must be used to make the buyer aware of the contents. It must be recognized, however, that warranties and disclaimers, no matter how well written, may not be the strongest defense.

Warnings and Directions. Directions are intended to ensure effective use of a product. Warnings are intended to ensure safe use. Both should be written to help the user understand and appreciate the nature of the product and its dangers. If directions and warnings are inadequate, there is potential liability, because it cannot be said that the user had contributory negligence in failing to appreciate and avoid danger.

The burden of full and effective disclosure is on the manufacturer. Directions and warnings, although essential, do not relieve the manufacturer of the duty to design a safe product. The law will not permit a manufacturer, who knowingly markets a product with a danger that could have been eliminated, to evade liability simply because a warning is placed on the product. One must design against misuse.

This topic is discussed in greater detail in the article “Safety in Design” in *Materials Selection and Design,* Volume 20 of *ASM Handbook.* A label is discussed in some detail in the article “Human Factors in Design” in *Materials Selection and Design,* Volume 20 of *ASM Handbook.*

Written Material. All advertising, promotional material, and sales literature must be carefully screened. Warranties can be implied or inferred by the wording on labels, instructions, pamphlets, sales literature, advertising (written and electronic broadcast), and so on, even though no warranty is intended. There must be no exaggeration in such material. The manufacturer must be able to show that the product is properly rated and that the product can safely do what the advertisement says it will do. Additional information on the appropriate level of language is given in the article “Safety in Design” in *Materials Selection and Design,* Volume 20 of *ASM Handbook.*

Human Factors. Many products and systems require operation by a human who thereby becomes an integral part of the system. As such, the human can have a very significant effect on system performance. One must recognize that the human being is the greatest, and least controllable, variable in the system. Some attorneys believe that most products liability suits result because someone (usually the designer) did not thoroughly think through how the product interfaced with society.

Products Recall Planning. It is a fact of life that mistakes are sometimes made even by highly experienced professionals exercising utmost care. When such errors occur, a products recall may be necessary. Unless the specific troublesome part can be readily and uniquely identified as to source, production procedure, time of manufacture, and so on, there will be great difficulty in pinpointing the problem within the producing organization. Placing one advertisement for recall purposes in newspaper and magazines (not including TV) throughout the country is very expensive. An obvious economic need, as well as a regulatory requirement, exists for manufacturers (and importers) to have systems in place for expeditious recall of a faulty product.

Records. Once involved in litigation, one of the most powerful defenses that manufacturers and engineers can have is an effective, extensive, and detailed record. Records should document how the design came about, with

notes of meetings, assembly drawings (including safety features), checklists, the state of the art at the time, and so on. These records, while no barrier to products liability lawsuits, will go a long way toward convincing a jury that prudent and reasonable care has been taken to produce a safe product.

Paramount Questions

No matter how carefully and thoroughly one executes all possible preventive measures, it is necessary to ask:

- What is the probability of injury?
- Who determines the probability of injury?
- What is an acceptable probability of injury?
- Who determines the acceptable probability of injury?

As Lowrance (Ref 42) suggests, determining the probability of injury is an empirical, scientific activity. It follows that engineers are better qualified by education and experience than most people to determine this probability. Presumably, designers will use organized approaches to cope with the complexity. One obvious place for assessing this probability is the design review process. While design review is a most valuable aid for the designer, it is not a substitute for adequate design and engineering.

As Lowrance (Ref 42) further suggests, judging the acceptable probability of injury is a normative, political activity. Obviously, assessing the probability of injury is not a simple matter. Assessing the acceptable probability of injury is far more complex and difficult. Use of the word “acceptable” emphasizes that safety decisions are relativistic and judgmental. It implies three questions:

Acceptable in whose view? Acceptable in what terms? Acceptable for whom? This use of “acceptable probability of injury” avoids any implication or inference that safety is an intrinsic, absolute, measurable property.

In assessing acceptable danger, one major task is determining the distribution of danger, benefits, and costs. This determination is both an empirical matter and a political issue. It involves questions such as:

Who will be paying the costs? Will those who benefit be those who pay? Will those endangered be those who benefit? Answers to these questions may be based on quantifiable data but often must be based on estimates or surveys. A related major task is to determine the equity of distribution of danger, benefits, and costs. This asks a question of fairness and social justice for which answers are a matter of personal and societal value judgment.

Who determines the acceptable level of probability of injury? In terms of ability to judge acceptability, designers/engineers are no better qualified than any other group of people and, in general, are less qualified than many others. It is often alleged that engineers (because of their inherent characteristics, education, and experience) are less sensitive to societal influences of their work and products than others. As for most stereotypes, there is some truth in this view. Clemenceau reportedly said: “War is much too serious a matter to be entrusted to the military.” Perhaps product design is much too serious a matter to be entrusted solely to designers and (especially) business managers.

Jaeger (Ref 43) has summarized the situation thus:

“Nowadays it seems to me that the risk problem in technology has turned out to become one of the most pressing questions concerning the whole of industrial development. This problem is of fundamental as well as of highly practical importance. The answer to the question “How safe is safe enough?” requires a combination of reflective and mathematical thinking as well as the integration of technological, economic, sociological, psychological and ecological knowledge from a superior point of view.”

If the designer cannot adequately make the determination, then who can? Various ideas have been proposed (e.g., Ref 44), but no suggestion yet made is fully satisfactory. The designer/producer must resolve this for each product. References 42 and 45 can be helpful in developing sensitivity to assessing an acceptable probability of injury.

References cited in this section

42. W.W. Lowrance, *Of Acceptable Risk,* William Kaufman, Inc., 1976

43. T.A. Jaeger, Das Risikoproblem in der Technik, *Schweizer Archiv fur Angewandte Wissenshafter und Technik,* Vol 36, 1970, p 201–207

44. C.O. Smith, "How Much Danger? Who Decides?" paper presented at American Society of Mechanical Engineers Conference "The Worker in Transition: Technological Change," Bethesda, MD, 5–7 April 1989

45. R.A. Schwing and W.A. Albers, *Societal Risk Assessment,* Plenum Press, 1980

Acceptable Level of Risk

Because "danger" has been previously defined as the unreasonable or unacceptable combination of hazard and risk, it is a contradiction to speak of an acceptable level of danger. However, an acceptable level of risk or hazard might be considered, following from the previous section, to be acceptable probability of injury. It should also be mentioned in this section, as it has been in previous sections, that the acceptable level is related to the benefits gained by taking the risk and the risk of alternatives.

It has been suggested that an acceptable level of danger might be 1 in 4000 per year. Currently (2002), 1 in 7000 per year may be a better number; that is, about 40,000 die per year in the United States out of a population of about 275 million. Statistics indicate that this is about the danger of dying from an automobile accident in the United States. One might infer that U.S. citizens consider this an acceptable level in view of the fact that little apparent effort is expended in trying to decrease the accident rate. The National Highway Traffic Safety Administration indicates that about 50% of fatal traffic accidents in the United States are alcohol related. If there were severe penalties for driving under the influence of alcohol (as there are in some other countries), this danger would presumably decrease to about 1 in 8000 per year. Either level of danger may be rational for the public as a whole (obviously debatable), but it probably is not perceived as such by a bereaved family. Such a rate hardly seems acceptable for consumer products. It certainly is unacceptable for nuclear applications. While the majority of manufactured products have a much lower level of danger than this, many of these products are considered to have a level of danger too high to be acceptable. Juries regularly make this decision in products liability actions.

One aspect of a potentially acceptable level of danger is the manner in which it is stated. Engineers might prefer to state the level in terms of probability. The general public, however, might well prefer it otherwise, or even unstated. The general public must be aware of fatalities from automotive accidents. It is possible that if automobile manufacturers were to point out that there is an annual chance of about 1 in 4000 that an individual will be killed, and a much greater chance of being injured (even seriously, such as spinal injuries, which not only incapacitate the victim but require constant attention by others), the attitude of the public might be different.

It must be recognized that while it is possible to reduce the level of danger to a very small number, danger cannot be completely eliminated, no matter how much effort is expended. We do not think there is any one level of acceptable danger. Each situation must be judged independently. The question is not what level of danger the engineer/designer thinks is acceptable for the public but what level the public perceives to be acceptable.

Acknowledgment

This article was adapted from C.O. Smith, Products Liability and Design, *Materials Selection and Design,* Volume 20, *ASM Handbook,* p 146–151.

References

1. *The Code of Hammurabi,* University of Chicago Press, 1904

2. *Restatement of the Law, Second, Torts,* 2d, Vol 2, American Law Institute Publishers, 1965

3. *Webster's New Twentieth Century Dictionary, Unabridged,* 2nd ed., Simon & Schuster, 1979

4. G.A. Peters, New Product Safety Legal Requirements, *Hazard Prevention,* Sept/Oct 1978, p 21–23

5. *Barker v Lull Engineering Co.,* 20 C. 3d 413

6. C.O. Smith, Manufacturing/Design Defects, Paper 86-WA/DE-14, American Society of Mechanical Engineers

7. C.O. Smith, Mobile Ladder Stand, Paper 87-DE-5, American Society of Mechanical Engineers

8. C.O. Smith, Design of a Saw Table, Paper 87-WA/DE-9, American Society of Mechanical Engineers

9. C.O. Smith, Coffee Grinder: Safe or Not?, Paper 88-WA/DE-6, American Society of Mechanical Engineers

10. C.O. Smith, Collapse of an Office Chair, Paper 89-WA/DE-18, American Society of Mechanical Engineers

11. C.O. Smith, Some Subtle (or Not So Subtle?) Product Defects, Paper 90-WA/DE-23, American Society of Mechanical Engineers

12. C.O. Smith, A Fatal Helicopter Crash, Paper 91-WA/DE-8, American Society of Mechanical Engineers

13. P.D. Beard and T.F. Talbot, What Determines if a Design is Safe, Paper 90-WA/DE-20, American Society of Mechanical Engineers

14. T.F. Talbot and C.S. Hartley, Failure of Fastening Devices in Pump Packing Gland Flange, Paper 89-WA/DE-12, American Society of Mechanical Engineers

15. T.F. Talbot and M. Crawford, Wire Rope Failures and How to Minimize Their Occurrence, Paper 87-DE-7, American Society of Mechanical Engineers

16. T.F. Talbot and J.H. Appleton, Dump Truck Stability, Paper 87-DE-3, American Society of Mechanical Engineers

17. T.F. Talbot, Safety for Special Purpose Machines, Paper 87-WA/DE-8, American Society of Mechanical Engineers

18. T.F. Talbot, Chain Saw Safety Features, Paper 86-WA/DE-16, American Society of Mechanical Engineers

19. T.F. Talbot, Hazards of the Airless Spray Gun, Paper 85-WA/DE-13, American Society of Mechanical Engineers

20. T.F. Talbot, Man-Lift Cable Drum Shaft Failure, Paper 87-WA/DE-19, American Society of Mechanical Engineers

21. T.F. Talbot, Bolt Failure in an Overhead Hoist, Paper 83-WA/DE-20, American Society of Mechanical Engineers

22. W.G. Ovens, Failures in Two Tubular Steel Chairs, Paper 91-WA/DE-9, American Society of Mechanical Engineers

23. J.A. Wilson, Log Loader Collapse: Failure Analysis of the Main Support Stem, Paper 89-WA/DE-13, American Society of Mechanical Engineers

24. T.A. Hunter, Design Errors and Their Consequences, Paper 89-WA/DE-14, American Society of Mechanical Engineers

25. C.O. Smith and T.F. Talbot, Product Design and Warnings, Paper 91-WA/DE-7, American Society of Mechanical Engineers

26. C.O. Smith and J.F. Radavich, Failures from Maintenance Miscues, Paper 84-DE-2, American Society of Mechanical Engineers

27. C.O. Smith, Failure of a Twistdrill, *J. Eng. Mater. Technol.*, Vol 96 (No. 2), April 1974, p 88–90

28. C.O. Smith, Legal Aspects of a Twistdrill Failure, *J. Prod. Liabil.*, Vol 3, 1979, p 247–258

29. C.O. Smith, "ECL 170, Tortured Twist Drill," Center for Case Studies in Engineering, Rose-Hulman Institute of Technology, Terre Haute, IN

30. R.W. Bailey, *Human Performance Engineering: A Guide for System Designers*, Prentice-Hall, 1982

31. T.A. Hunter, *Engineering Design for Safety*, McGraw-Hill, 1992

32. E. Grandjean, *Fitting the Task to the Man*, 4th ed., Taylor and Francis, Washington, D.C. 1988

33. B.A. Sayers, *Human Factors and Decision Making: Their Influence on Safety and Reliability*, Elsevier Science Publishers, 1988

34. K.S. Park, *Human Reliability*, Elsevier Science Publishers, 1987

35. L.S. Mark, J.S. Warren, and R.L. Huston, Ed., *Ergonomics and Human Factors*, Springer-Verlag, 1987

36. B.S. Dhillon, *Human Reliability*, Pergamon Press, 1986

37. C.D. Wickens, *Engineering Psychology and Human Performance*, 2nd ed., Harper-Collins, 1992

38. W.E. Woodson, *Human Factors Design Handbook*, McGraw-Hill, 1981

39. G. Salvendy, Ed., *Handbook of Human Factors*, Wiley, 1987

40. W.G. Ireson, Chapter 12, *Reliability Handbook*, McGraw-Hill, 1966

41. C.O. Smith, Two Industrial Products—Defective Design?, Paper 93-WA/DE-11, American Society of Mechanical Engineers

42. W.W. Lowrance, *Of Acceptable Risk*, William Kaufman, Inc., 1976

43. T.A. Jaeger, Das Risikoproblem in der Technik, *Schweizer Archiv fur Angewandte Wissenshafter und Technik*, Vol 36, 1970, p 201–207

44. C.O. Smith, "How Much Danger? Who Decides?" paper presented at American Society of Mechanical Engineers Conference "The Worker in Transition: Technological Change," Bethesda, MD, 5–7 April 1989

45. R.A. Schwing and W.A. Albers, *Societal Risk Assessment*, Plenum Press, 1980

Selected References

- S. Brown, I. LeMay, J. Sweet, and A. Weinstein, Ed., *Product Liability Handbook: Prevention, Risk, Consequence, and Forensics of Product Failure,* Van Nostrand Reinhold, 1990
- V.J. Colangelo and P.A. Thornton, *Engineering Aspects of Product Liability,* American Society for Metals, 1981
- R.A. Epstein, *Modern Products Liability Law: A Legal Revolution,* Quorum Books, Westport, CT, 1980
- R.L. Goodden, *Product Liability Prevention: A Strategic Guide,* ASQ Quality Press, Milwaukee, WI, 2000
- P.W. Huber and R.E. Litan, Ed., *The Liability Maze: The Impact of Liability Law on Safety and Innovation,* The Brookings Institute, 1991
- W. Kimble and R.O. Lesher, *Products Liability,* West Publishing Co., 1979
- J. Kolb and S.S. Ross, *Product Safety and Liability,* McGraw-Hill, 1980
- M.S. Madden, *Products Liability,* Vol 1 and 2, West Publishing Co., 1988
- C.O. Smith, *Products Liability: Are You Vulnerable?,* Prentice-Hall, 1981
- J.F. Thorpe and W.H. Middendorf, *What Every Engineer Should Know about Products Liability,* Dekker, 1979