NASA/TM—2019-220179 AIAA–2018–5413

# Prognostic Launch Vehicle Probability of Failure Assessment Methodology for Conceptual Systems Predicated on Human Causal Factors

*Craig H. Williams*
*Glenn Research Center, Cleveland, Ohio*

*Lawrence J. Ross and J. Joseph Nieberding*
*Aerospace Engineering Associates LLC, Bay Village, Ohio*

December 2019

# NASA STI Program . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA Scientific and Technical Information (STI) Program plays a key part in helping NASA maintain this important role.

The NASA STI Program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI Program provides access to the NASA Technical Report Server—Registered (NTRS Reg) and NASA Technical Report Server—Public (NTRS) thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counter-part of peer-reviewed formal professional papers, but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., "quick-release" reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at http://www.sti.nasa.gov

- E-mail your question to help@sti.nasa.gov

- Fax your question to the NASA STI Information Desk at 757-864-6500

- Telephone the NASA STI Information Desk at 757-864-9658

- Write to:
  NASA STI Program
  Mail Stop 148
  NASA Langley Research Center
  Hampton, VA 23681-2199

NASA/TM—2019-220179 AIAA–2018–5413

# Prognostic Launch Vehicle Probability of Failure Assessment Methodology for Conceptual Systems Predicated on Human Causal Factors

*Craig H. Williams*
*Glenn Research Center, Cleveland, Ohio*

*Lawrence J. Ross and J. Joseph Nieberding*
*Aerospace Engineering Associates LLC, Bay Village, Ohio*

Prepared for the
2018 AIAA Space and Astronautics Forum and Exposition
sponsored by the American Institute of Aeronautics and Astronautics
Orlando, Florida, September 17–19, 2018

National Aeronautics and
Space Administration

Glenn Research Center
Cleveland, Ohio 44135

December 2019

## Acknowledgments

The authors would like to thank all the original sources of NASA, industry, and foreign accident investigation board reports. Appreciation is extended to Mr. Ken O'Connor, Mr. Steve Lilley, and Mr. Keith Knudson of the NASA Headquarters Safety Center for their guidance and assessment of the probabilistic methodology. We appreciate the thoughtful advice on potential limitations of this approach provided by Ms. Christine Kilmer of the Reliability and System Safety Engineering Branch at the NASA Glenn Research Center. We are most thankful for the inspiration, support, and guidance provided by the Defense Advanced Research Projects Agency (DARPA) Experimental Spaceplane Program: program managers Mr. Jess Sponable and Mr. Scott Wierzbanowski, Chief Engineer Mr. Vijay Ramasubramanian, and Government team leaders Mr. Joseph Padavano (ARES Corporation) and Mr. Jay Penn (The Aerospace Corporation).

Trade names and trademarks are used in this report for identification only. Their usage does not constitute an official endorsement, either expressed or implied, by the National Aeronautics and Space Administration.

*Level of Review*: This material has been technically reviewed by technical management.

Available from

NASA STI Program
Mail Stop 148
NASA Langley Research Center
Hampton, VA 23681-2199

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
703-605-6000

This report is available in electronic form at http://www.sti.nasa.gov/ and http://ntrs.nasa.gov/

# Prognostic Launch Vehicle Probability of Failure Assessment Methodology for Conceptual Systems Predicated on Human Causal Factors

Craig H. Williams
National Aeronautics and Space Administration
Glenn Research Center
Cleveland, Ohio 44135

Lawrence J. Ross and J. Joseph Nieberding
Aerospace Engineering Associates LLC
Bay Village, Ohio 44140

## Summary

Lessons learned from past failures of launch vehicle developments and operations are used to create a new method to predict the probability of failure of conceptual systems. Existing methods such as Probabilistic Risk Assessments and Human Risk Assessments are considered but found to be too cumbersome for this type of system-wide application for yet-to-be-flown vehicles. The basis for this methodology is historic databases of past failures, where it was determined that various faulty human interactions were the predominant root causes of failure rather than deficient component reliabilities that were evaluated through statistical analysis. This methodology contains an expert scoring part, which can be used in either a qualitative or a quantitative mode. The method produces two products: (1) a numerical score of the probability of failure and/or (2) guidance to program management on critical areas in need of increased focus to improve the probability of success. In order to evaluate the effectiveness of this new method, data from a concluded vehicle program (U.S. Air Force's Titan IV with the Centaur G-Prime upper stage) was used as a test case. Although the theoretical versus actual probabilities of failure were found to be in reasonable agreement (4.46 vs. 6.67 percent, respectively) the underlying subroot cause scoring had significant disparities attributable to significant organizational changes and acquisitions. Recommendations are made for future applications of this method to ongoing launch vehicle development programs.

## Nomenclature

$a$ lower limiting score of root causes
$b$ upper limiting score of root causes
$E$ event
$F$ cumulative distribution function
$P$ probability of failure
$X$ random variable of interest (the score of root causes for any case)
$\Omega$ sample space
$\omega$ possible cases

# 1.0 Introduction

Analytic methods to evaluate a launch vehicle's probability of failure are frequently hardware-centric. Analysis tends to rely on component failure rates used in statistical analyses to predict the chance of failure of an integrated vehicle. The methods used in such approaches are sound and produce a defendable numerical results. However, assessments of historic launch vehicle failures repeatedly show that the underlying causes of failures generally originate from humans, rather than hardware component failure or other manifestations of poor quality control.

This incongruity between presumed cause and actual cause is problematic when attempting to quantify a credible probability of failure of a conceptual launch system based on historic real examples. Although probabilistic risk assessments (PRAs) are intended to do just that, they tend to be resource intensive to perform. Even then, their accuracy is subject to scrutiny. Figure 1 shows PRAs performed a dozen years after the first shuttle launch (and also after the Challenger Space Transportation System (STS) 51–L failure) produced failure probabilities that were many times more optimistic than the actual risk progression was (calculated after shuttle retirement) (Ref. 1). As will be discussed, much of the optimistic assessments can be attributed to the lack of addressing human causal factors.

A concise, easily implemented new method, which accurately takes into account human causal factors, is developed here to quantify launch-system reliability for future concepts.

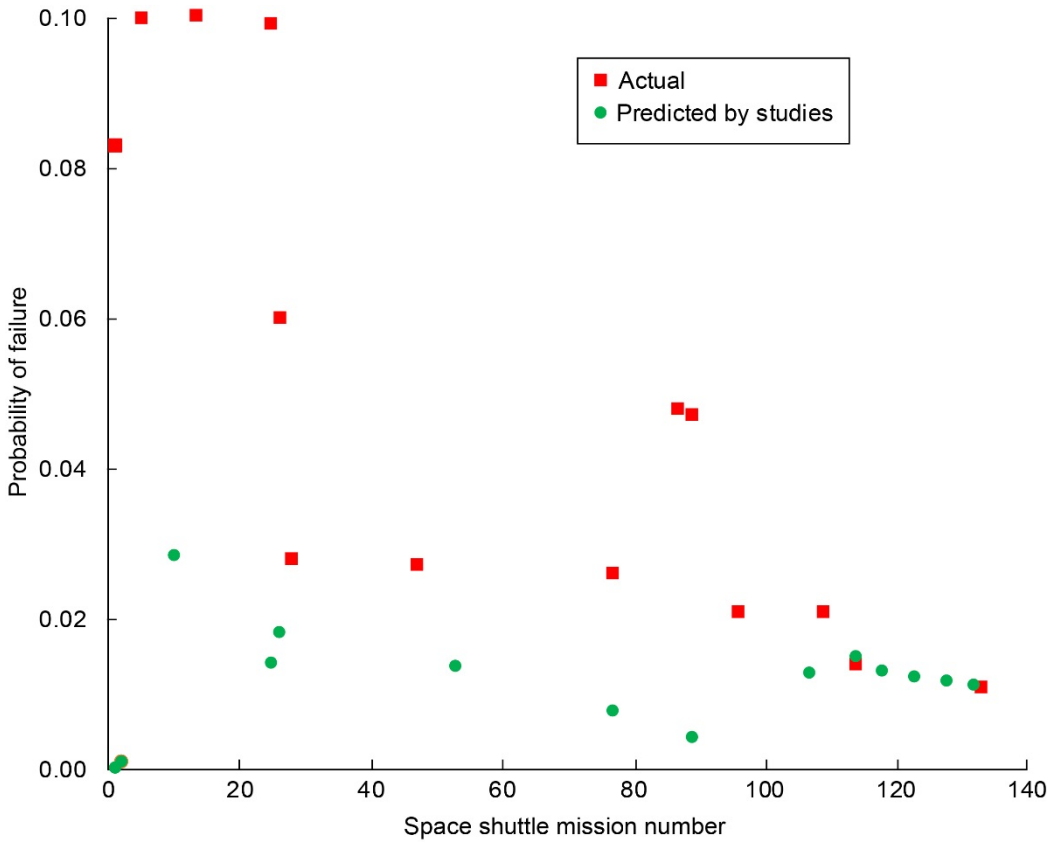


Figure 1.—Actual versus predicted probability of failure for space shuttle system.

## 2.0 Proximate Versus Root Causes of Failure

The distinction between "proximate" and "root" causes is given in NASA's Procedural Requirements (Ref. 2):

(1) "**Proximate Cause**: The event(s) that occurred, including any condition(s) that existed immediately before the undesired outcome, directly resulted in its occurrence and, if eliminated or modified, would have prevented the undesired outcome. Also known as the direct cause(s)."
(2) "**Root Cause**: An event or condition that is an organizational factor that existed before the intermediate cause and directly resulted in its occurrence (thus indirectly it caused or contributed to the proximate cause and subsequent undesired outcome) and; if eliminated or modified, would have prevented the intermediate cause from occurring, and the undesired outcome. Typically, multiple root causes contribute to an undesired outcome."

As will be illustrated, root causes of most launch vehicle failures (despite differing proximate causes) share a lot of similarities. For example, in the case of the U.S. Air Force (USAF) Titan IVB/Centaur launch of a Milstar spacecraft failure in 1999, the vehicle tumbled during the Centaur upper stage phase of flight, which left the payload in a useless orbit. The proximate cause of the failure found by the accident investigation board was a loss of Centaur upper stage roll control due to a software error. Specifically, a value of an exponent within the flight software was entered as a "zero" instead of a "one." The root causes, however, were human in nature, where "the software development process that allowed a human error to go undetected" (Ref. 3):

(1) Human entered erroneous flight constants.
(2) Human software checks failed to detect the error due to lack of understanding by staff.
(3) Software testing lacked formality and was performed with default values (not the entered flight values).
(4) Cape personnel did not diligently follow up when they noticed something atypical.

## 3.0 Existing Methods to Assess Probability of Failure

It is reasonable to assume that such a mature field would have created methods to assess probability of failure for entire aerospace systems that included human-centric root causes. Discussions were held with the NASA Headquarters Safety Center and the NASA Glenn Research Center Safety, Reliability and Quality Assurance Branch, and literature searches were performed on the subject. Two comprehensive documents were identified and reviewed.

### 3.1 NASA Headquarters Study of Human Reliability Analysis Methods

There have been approximately 50 different methods to assess and predict complex system probability of failure developed over the past half-century (Ref. 4). Most of these methods were created to assist the nuclear power industry and are largely hardware centric. Out of these 50 methods, 14 were selected by NASA Headquarters (HQ) for further study based on their applicability for launch vehicle failure assessments (Ref. 4). This subset was predicated on methods that contained human reliability analysis (HRA), which enabled incorporating effects and probabilities of human errors for a more effective use of PRAs. Outside HRA experts were brought into the HQ study team from academia, other Federal laboratories, and the private sector. (Note that the existing NASA PRA guidance provides a method similar to those practiced by industry (Ref. 5).) These combined HRA and PRA techniques were

compared comprehensively in order to determine which were best suited to help guide the development of future NASA space systems. However, the HRA process (problem definition; task decomposition; and the identification, representation, and quantification of human error) was most readily applied to "bottoms-up" initial design, analysis of individual tasks, and operating specific components or systems. The two initial HRA steps can become quite complex if not applied to clearly defined problems that are limited in scope. Even though there are commercially available software tools designed to facilitate this work, the process can easily become unwieldy if applied on an entire launch vehicle system.

One of the methods studied by NASA HQ was human factors process failure mode and effects analysis (HF PFMEA) (Ref. 4). Originally created by NASA HQ, this method was designed to identify human errors and their consequences. However, HF PFMEA was designed to focus on specific subsystems that have a limited number of operation steps. HF PFMEA methodology then defines all possible combinations of acts a person could make in order to correct undesirable sequences of events. This produces a considerable number of possible scenarios and actions, making it unwieldy for systemwide application on a conceptual design. In addition, HF PFMEA does not calculate human error probabilities (a primary reason for it not being further considered in the HQ study).

The HQ's study chose 4 of the 14 methods for further assessment, finding them superior for space-system development, each with varying strengths and weaknesses. Upon closer examination, each seemed unwieldy for assessing an entire space system (launch vehicle and its ground systems) from the perspective of known past (human) root causes of failures. Further, event modeling in the HRA/PRA process became even more tedious and complex for problems beyond a finite subsystem when using any of these methods. Since none of these methods appeared capable of assessing a launch vehicle system without requiring considerable effort, investigation continued for other methods.

## 3.2 STAMP and STPA

In a comprehensive assessment external to NASA, the Systems-Theoretic Accident Model Process (STAMP) was found to be a viable prospect using an all-encompassing accident model based on systems theory (Ref. 6). STAMP both analyzes accidents after they occurred and creates approaches to prevent them from occurring in developing systems. This method is not focused on failure prevention per se, but on reducing hazards by influencing human behavior through the use of constraints, hierarchical control structures, and process models to improve system safety. (Aside: a reliable system may not be safe, and vice versa.) In addition, its author takes exception to making a distinction between proximate and root causes, maintaining that the distinction is at least artificial and at most an obstacle to discovering the true causes of unsafe operations. STAMP's top-down approach guides the user to produce "safety constraints" and ensure that they are enforced (rather than generating a time-sequenced "series of events"). This dynamic treatment of the launch system was proposed as a superior method to create an accident model (Figure 2). STAMP is very comprehensive, analyzing not just immediate causes of the failure but also the societal, organizational, cultural, and governmental environments surrounding it.

The predictive part of the problem (a hazard analysis) was built on STAMP, using it as a preceding analysis. It was called the System-Theoretic Process Analysis (STPA). The primary reason for creating STPA was to include all causal factors identified in STAMP: "…design errors, software flaws, component interaction accidents, cognitively complex human decision-making errors, and social organizational and management factors contributing to accidents," precisely the human sources of failure lacking in many existing methods (Ref. 6). A primary output of this method is "functional control diagrams" rather than component (i.e., hardware) diagrams. The two overall objectives of STPA are to (Ref. 6) (1) identify every safety constraint that (if violated) could result in an unsafe or failed condition and (2) ascertain how each of these constraint violations could occur, including deterioration over time.

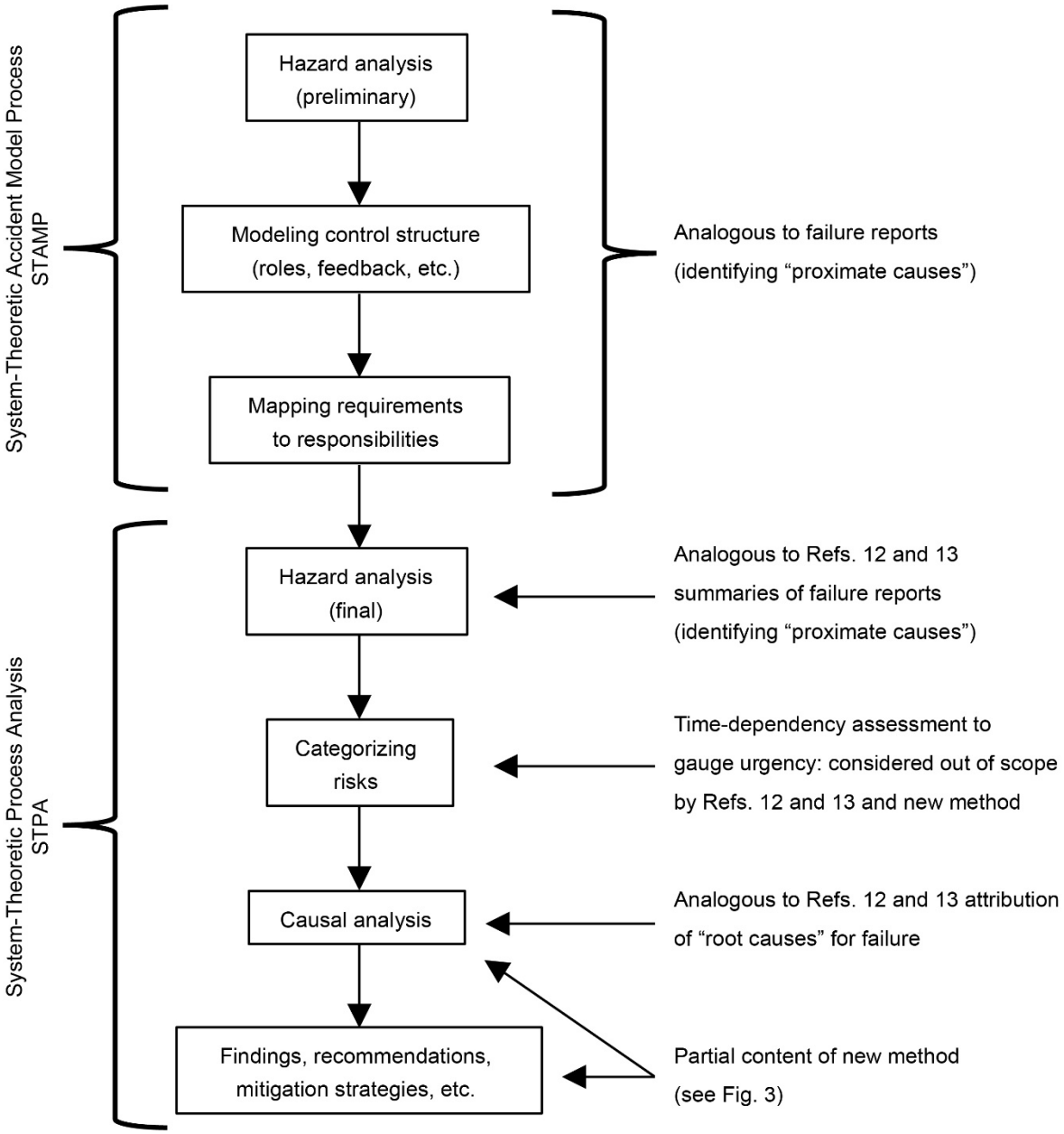


Figure 2.—Relationship of STAMP/STPA to new method and its underlying basis.

STAMP/STPA guides the design process rather than requires the design to exist beforehand, making it attractive for conceptual design applications. Thus, the composite STAMP/STPA method, based largely on human factors, might be used to more accurately predict a systemwide probability of failure. However, as with the four methods assessed in the NASA HQ study, the STAMP/STPA's exhaustively detailed nature could drive the analysis to become unwieldy if it is not narrowly tailored to a specific subsystem. For example, Reference 6 applied STAMP to the failed Titan IVB/Centaur-32 launch, narrowly focusing only on the proximate cause of the failure: faulty flight software development and insufficient testing. Yet the resulting STAMP processes limited to the initial assessment of this specific cause of failure alone required 30 pages to summarize (Ref. 6). If this approach had been used to assess the entire launch vehicle system, it would have been overwhelming. Other examples were provided, also in Reference 6, that were similarly focused on the specific cause of failure, not a systemwide analysis. Although STAMP/STPA can be used for organizational and managerial issues, "Less experimentation has been done on applying it at these levels and, once again, more needs to be done" (Ref. 6). Thus, for similar reasons as the NASA HQ study, it did not appear that STAMP/STPA could be concisely applied to perform major systemwide assessment of probability of failure of a conceptual design. Reference 6 went further, suggesting that attempting to quantify human actions impacting future system reliability may not even be possible because of the unpredictability of human interaction with the surrounding conditions (questioning this paper's premise). That assertion was rejected and a new method was created.

### 3.3 Other Methods

More traditional methods were also examined. These methods included models based on subsystem characteristics (both descriptive and functional) where all conceivable failure modes were attempted to be analytically described (Ref. 7). Here, assumed subsystem reliabilities by the authors were limited to technical parameters (no human factors) such as component life and vehicle configuration (number of engines, length time of operation, etc.). A similar method assigned subsystem reliabilities and then combined block diagrams and prediction modules to address functionality, operability, and other interdependencies of subsystems (Ref. 8). These methods appeared to be limited by lack of human factors and assumptions on hardware reliability statistics.

Other techniques relied on past reliability improvements in the aggregate of various launch vehicles, which were then curve fitted and adjusted for various approaches to modeling reliability growth. These methods took into account the entire system (rather than components) and assumed that whatever past vehicle improvements took place would similarly occur in future vehicles. Each model had a different shortcoming in forecasting failure rates of future systems (Ref. 9). More sophisticated prognostication methods of this approach also exist (Ref. 10).

Various NASA program standards and guidelines now recommend that some type of human factors assessments be performed, with no preferred practices. Either a Human Error Risk Assessment or Human Factor Task Analysis of some type is required to be performed, where the latter must "Analyze and list all the things people will do in a system, procedure, or operation" (Ref. 11). The overly broad "…all the things…" could easily result in significant effort.

The above methods represent a considerable body of work. Nevertheless, the authors failed to find a method that reasonably prognosticates launch vehicle probability of failure for conceptual systems that was predicated on human causes.

# 4.0 Proposed New Method

Since an existing, straightforward technique could not be found that relied on historic human causality data to assess the likelihood of failure of a conceptual launch vehicle on an entire-system-wide basis, the development of a new method was pursued. This new method was intended to guide conceptual vehicle design, development, and testing to increase the probability of success during operation. We propose this new method based on a rational probabilistic approach using historic data from accident investigation board reports. Figure 3 illustrates the steps to this approach:

(1) Establish new method's basis
   a. Review of past proximate causes of launch vehicle failures
   b. Establish root causes of past launch vehicle failures based on expert judgment
   c. Categorize, then consolidate similar root causes into finite categories
   d. Establish baseline model using root causes of past launch vehicle failures
      i. Selection of cases to be used
      ii. Scoring of root and subroot causes
      iii. Plotting resultant data
   e. Derive function for probability of failure of launch system
(2) Apply new method: NASA/USAF Shuttle/Centaur G-Prime upper stage (as flown on Titan IV)

It is important to emphasize that the first part of this process to "establish new method's basis" is a one-time-only effort, reliant on the coauthors' (from AEA, see Sec. 4.1) experience and judgement. The second part of this process to "apply the new method to conceptual designs" is the application of this method by the aerospace community on conceptual designs and development programs.

How this new method compares to existing methods such as STAMP/STPA can be seen in Figure 2. Note that most of STAMP/STPA pertains to identifying proximate causes and root causes of specific past failures. Only the bottom-most part of Figure 2 (findings, recommendations, mitigation strategies, etc.) corresponds with (part of) this new methodology.

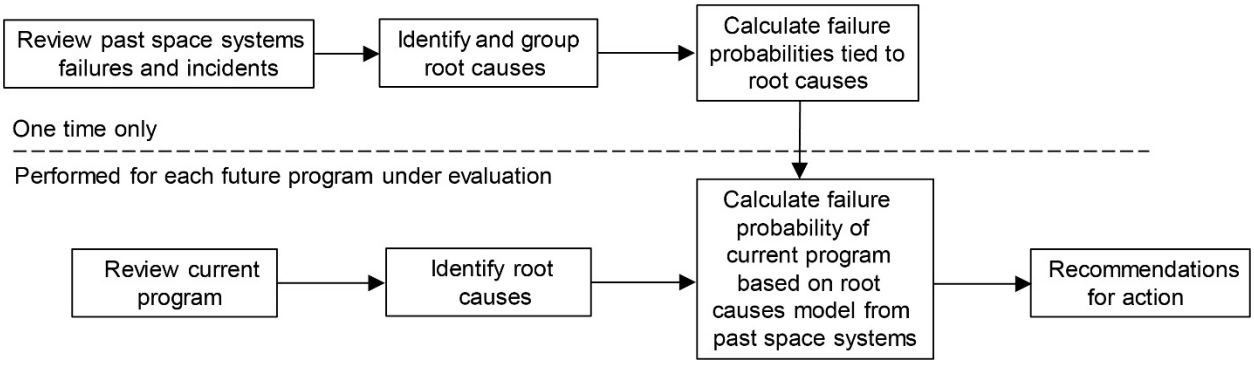


Figure 3.—Approach of new method.

## 4.1 Review of Past Proximate Causes of Launch Vehicle Failures

A comprehensive source of aerospace failure case studies was produced by two former NASA Glenn Research Center executives, now leaders of Aerospace Engineering Associates LLC (AEA) (Refs. 12 and 13). They are coauthors of this report. Over the course of their 30+-year careers, they successfully led launch vehicle development programs and actively served in leadership roles on more than 60 launch teams. It is this comprehensive experience that was fundamental to establishing the credibility of this new method. At AEA, they reviewed and assessed over 50 NASA and international case studies of launch vehicle and spacecraft failures as well as other major system incidents, which became the database for this new methodology. The proximate causal data were obtained from accident investigation board reports, interviews with those directly involved, and subject matter experts. The failure case studies consisted of 26 launch vehicles, 16 spacecraft, and 12 other aerospace or major systems (ground systems, aircraft, and major test facilities, etc.). This was not intended to be an all-inclusive database of past launch vehicle failures; only the cases evaluated by AEA were used in the formulation of this new methodology.

## 4.2 Establish Root Causes of Past Launch Vehicle Failures Based on Expert Judgment

After analyzing the failures and their proximate causes, the coauthors from AEA developed specific actions to remedy the mistakes. The absence of these specific actions can be considered the root causes of the failures. They found that the nature of the root causes did not depend on the type of aerospace system (launch vehicle, spacecraft, major ground test site, etc.). Also, root cause types did not change with time. What did matter was that root causes attributed to humans dominated over those attributed to hardware failure. Indeed, they state, “An examination of space mission and other mishaps finds human error to be a dominant factor (Ref. 13).” Further, it was found that most failures had more than one root cause. These findings substantiate the major problem with aerospace systems probability of failure analyses stated earlier: whereas methods to assess probability of failure tend to be focused on hardware, the root causes tend to be human-centric. Although a human-factors-based method may be difficult to repeat consistently, lack statistical rigor, or be somewhat deficient in system engineering, it nevertheless would focus on the overwhelming majority of the true (i.e., root) causes of failure. Therefore, as long as the methodology is reasonably sound, a human-factors-based probability of failure assessment methodology should be more predictive and useful than methods currently used.

## 4.3 Categorize, Consolidating Similar Root Causes Into Finite Categories

There have been efforts in the past to categorize and consolidate similar root causes. The report by the Mishap Investigation Board of the Astro-E2 mission in 2005 had a graphic that illustrated “Recurring Project Themes from Previous Mishaps,” which documented 28 distinct possible root causes (Ref. 14). The coauthors of the current study have published an earlier presentation with (only) four distinct causation categories (one of which was subdivided into six subcategories) (Ref. 13). Upon reflection of the results in Section 4.2 above, it was felt that a dozen distinct categories were needed to adequately capture the various types of root causes without becoming unwieldy. Some categories were noticeably absent, such as “legacy hardware,” a frequent area of concern and topic of discussion. Yet it is the actions people fail to take with legacy hardware that mattered: insufficient testing, reliance on prior similar design requirements, erroneously assuming that implicit limits did not apply, and so forth. Testing was separated into two categories: system and subsystem or component. This was because system testing is designed to pick up integration and ambient environment issues, whereas subsystem or component testing is largely focused on individual self-functionality. Hardware and software failure root causes (the types that receive

a disproportionate amount of attention in other probability of failure assessments) were found to be relatively minor root causes of failure. Complete explanations of the subgroups within each category are as follows:

(1) Insufficient or lack of prudent integrated testing is a major root cause of failures in launch vehicles. Not pursuing a "test as you fly; fly as you test" philosophy is a related characteristic. Without sufficient understanding of interactions within the entire system (which implies careful review and comprehension of data from an otherwise well-executed test campaign), the risk of system-to-system problems increases significantly. Test data of an operating system while in relevant environment (thermal, vacuum, vibration, etc.) are particularly essential for success.

(2) Engineering errors can be in the form of faulty hardware design and/or fabrication. Incorrect analytical modeling (where the actual operation or the environment is not correctly represented) or computational errors (where engineers make mistakes), if left uncaught, can result in launch failures.

(3) Unsound systems engineering (SE) practices have been a major impediment to mission success. Inadequate SE (correct design requirements, robust margins, etc.) by individuals lacking sufficient depth of experience, judgment, or critical understanding of the relevant technical field is captured within this area. Directly related are insufficient meaningful reviews (where major problems are identified, data presented and discussed, and decisions made), which are displaced by pro forma reviews with delayed critical decisions. SE experts are also expected to challenge analyses, heritage, and other assumptions in order to gage their soundness to substantiate their decisions. Analytic models not correlated with actual flight-derived data, scaled from other source, or of questionable validity are also expected to be rooted out by sound SE.

(4) Insufficient or lack of prudent component or subsystem testing is also a major root cause of failures. Prudent testing prior to integration permits discovery while each subsystem or component is isolated from others. Relying on verification by analysis or comparison with requirements without first obtaining test data can give the program a false sense of security. Heritage hardware or software may appear to save money and effort, but not validating either for new application, range of operation, or a new environment can risk significant cost and schedule downstream. Lastly, forgoing lower level testing can result in overlooking the opportunity to establish instrumentation needs, which are typically first brought to light during subsystem-level testing.

(5) Failure to follow established processes (or errors in processes) span fabrication, test, integration, and launch operations. Nonstandard events, loosely controlled changes, and workarounds not formally incorporated into the standard process (or not included in the program documentation) have caused serious mishaps.

(6) Failures of hardware are categorized here. These root causes include random part failure, poor quality, and/or statistically out-of-tolerance components ($-3\sigma$). Also included here are multiple unforeseen changes in program, the environment, and secondary effects on hardware, where a low probability chain of events unfortunately appear to conspire to doom a mission.

(7) "Better-Faster-Cheaper" is an expression originally coined by a NASA Administrator in the 1990s and used as a basis for policy for creating and managing major programs with deliberately compressed schedules, highly constrained cost, and highly visible to the public. It is used here more generically to describe a root cause of failure that can be attributed to imprudently low funding and overworked staff due to an insufficient schedule imposed to carry

out policy initiatives. These conditions sometimes drove staff to take (or not take) actions against their better judgment, believing that resistance was futile.

(8) Poor program management has been a highly visible root cause of failures. Inattentiveness to (or ineffectiveness in) managing problems even when they are program-threatening is chief among the characteristics. The "Normalization of Deviance" is something associated with the Challenger and Columbia Space Shuttle disasters: an unexpected deviation in system performance accompanied by revised expectations continue until a catastrophic occurrence results. Regrettably, also part of this category is lack of leadership integrity—such as provable knowledge that a program cannot succeed technically, yet senior management continues to spend money and consume resources until termination.

(9) Failures of software are categorized here. Differences between functional specifications and true requirements can lead to software failures. An all too common aspect is insufficient (or no) independent verification and validation (IV&V), which invites broken software to remain undetected until too late.

(10) Effective communication between organizations, management, and other members of the program's broader team is essential. When it fails, the consequences can be devastating. Sometimes there are subtle, but fundamentally important, differences in how organization-to-organization relationships function. Insufficient formality between working groups have led to unresolved action items that later proved lethal to the program.

(11) Independent reviews are intended to surface problems that are complex, cross many department or system lines of authority, are too subtle for all but the most experienced staff to identify, and/or have escaped all customary checks and balances. Sometimes reviews are treated as pro forma, where true problems are either ignored or rationalized. An absence of independent assessment sometimes occurs in programs, where a conflict of interest gets in the way of the duty to hold the review. There have been occasions where the independent review has functioned well, yet the program for whatever reason fails to heed or fully implement the recommendations. Despite experienced and diligent program managers, sometimes bad things just happen.

(12) There are other root causes of failure, sometimes unique to a specific program but just as devastating: for example, the urgency to compete with a foreign adversary may push a program's leadership to act (or not act) in a way he would otherwise not; or an extremely talented, well-respected leader might have such an inspiring effect on his staff that his untimely departure may cause everyone to lose faith in the project.

## 4.4 Assessment of Root Causes

The root causes of each failure case were grouped into the 12 categories previously described as similarities became apparent. The groupings were then tallied and are shown in Figure 4. There is a fairly even distribution among the human-failure-based root causes with no clearly dominant category. The leading root causes are lack of sufficient testing (both integrated system and component), lack of appropriate systems engineering, and engineering and process errors—altogether totaling 63 percent. All human-factor root causes amount to 87 percent, whereas hardware failures, by contrast, contribute less than 9 percent. In fact, there is only a single case where random hardware component failure was the sole root cause (Ref. 13). These results indicate that focusing on only one or two root causes to assess the probability of failure would be inadequate and that emphasizing statistical hardware failure would be misplaced focus. This assessment confirms that a multifaceted approach focusing on a variety of human causal factors is needed.

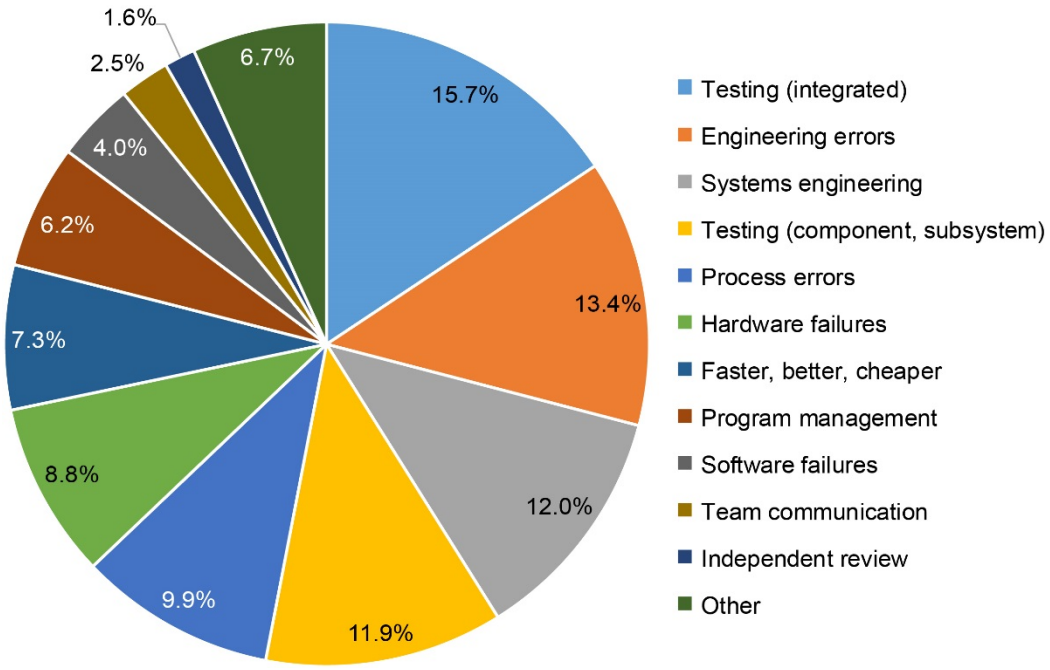


Figure 4.—Distribution of root causes in launch vehicle system failures.

## 4.5 Downselection of Cases To Be Used for Basis of Method

Not all cases that were assessed in References 12 and 13 were used in the analysis employed by the new method. Desiring the largest reasonable sample space initially inspired the inclusion of the spacecraft failures as well as those of the launch vehicles. Both types of vehicles had similar characteristics from a general engineering perspective, and indeed, the failure mechanisms were similar (if not the same). However, a practical problem became obvious when the statistical part of this methodology was exercised (Section 4.8): how to account for the numerical total of spacecraft in the sample space? Although the total number of launch vehicles in the sample space can be reasonably quantified given their near similarity, it became problematic when addressing spacecraft. For example, should all Intelsats be grouped together or just within series? How should "one-of-a-kind" interplanetary spacecraft such as Galileo be treated? Even though the qualitative (color-coded) part of this methodology could be useful for spacecraft, the quantitative part of calculating failures per total sample space was problematic. Since there was a small, but adequate, number of launch vehicle cases, a practical decision was made to exclude spacecraft in the analysis. Another concern over which launch vehicle cases should be included was raised with respect to using only "operational" vehicles and avoid "test or research and development (R&D) infant mortality." However, that would have reduced the total sample set to a mere 14. Further, the characteristics of the R&D failures were very similar to those of the operational vehicle failures. So it was decided to include all launch vehicle failures contained in References 12 and 13, while excluding the spacecraft and other systems. Thus, of the 54 cases in the total database, a subset of 21 case studies of launch vehicles only (both "development" and "operational") were selected as the basis for this methodology (Table I).

TABLE I.—SELECTED FAILURE CASE STUDIES OF LAUNCH VEHICLE SYSTEMS

| Mission | | Problem | Result | Number in series | Description of series |
|---|---|---|---|---|---|
| Research and development | | | | | |
| 1 | Atlas/Centaur F-1 | Premature shield separation | Loss of mission | 8 | Test flights: 7 LeRC[a] led + F-1 |
| 2 | Atlas/Centaur AC-5 | Premature booster engine shutdown | Loss of mission, pad | | See AC F-1 |
| 3 | [b]N-1, no. 1 | Stage 1 failure | Loss of mission | 4 | Four N-1s in series |
| 4 | [b]N-1, no. 2 | T–0 explosion | Loss of mission, pad | | See N-1, no. 1 |
| 5 | [b]N-1, no. 3 | Uncontrolled roll | Loss of mission | | See N-1, no. 1 |
| 6 | [b]N-1, no. 4 | Pogo[c] | Program termination | | See N-1, no. 1 |
| 7 | Titan IIIE/Centaur TC-1 | Centaur engine start failure | Loss of mission | 1 | Test flight only |
| 8 | X-43A | Loss of control | Loss of mission | 3 | Three (expendable) vehicles; one failure |
| Operational | | | | | |
| 1 | Apollo 13 Service Module | LOX[d] tank explosion | Loss of mission | 20 | Total service module flights |
| 2 | Apollo 13 Stage II | Pogo[b] | Potential loss of mission | 13 | Total Saturn V flights |
| 3 | Ariane 5 (501) | Loss of control | Loss of mission | 92 | Total up through May 2017 |
| 4 | Atlas/Centaur AC-21 | Fairing separation failure | Loss of mission | 61 | Total nontest flight AC up to 1990 (AC-69) |
| 5 | Atlas/Centaur AC-24 | Avionics hardware failure | Loss of mission | | See AC-21 |
| 6 | Atlas/Centaur AC-33 | Loss of control | Loss of mission | | See AC-21 |
| 7 | Atlas/Centaur AC-43 | Booster engine failure | Loss of mission | | See AC-21 |
| 8 | Atlas/Centaur AC-62 | Loss of control during coast | Compromised mission | | See AC-21 |
| 9 | Atlas/Centaur AC-67 | Lightning strike | Loss of mission | | See AC-21 |
| 10 | Space Shuttle Challenger | SRM[e] failure | Loss of mission | 135 | Total space shuttle flights |
| 11 | Space Shuttle Columbia | Launch-induced wing damage | Loss of mission | | See Space Shuttle Challenger |
| 12 | Titan IIIE/Centaur TC-6 | Stage 2 LOX[d] tank problem | Potential loss of mission | 6 | Post TC-1 |
| 13 | Titan IVB/Centaur-32 | Loss of control | Loss of mission | 16 | Total Titan IV/Centaur flights |
| Total | | | | 359 | |

[a]NASA Lewis (now Glenn) Research Center.
[b]Russian rocket.
[c]Engine oscillation (caused by combustion instability) in resonance with vehicle's natural frequencies.
[d]Liquid oxygen.
[e]Solid rocket motor.

## 4.6 Scoring of Root and Subroot Causes With Requisite Expertise

The scoring for the first part of the methodology was done based on judgment of the root causes identified in References 12 and 13. “Expert judgment” credibility of both AEA coauthors of this report was established and demonstrated by their successful engineering and managerial leadership of several launch vehicle developments, more than 60 launches spanning over three decades, accident investigation boards, and several major conceptual launch vehicle design studies. Such experience and demonstrated accomplishments were essential in order to correctly identify and judge roots causes of past failures. Credibility to score the second part of the methodology—applying the root cause basis to a conceptual design—required much less experience: at least one launch vehicle development program (doing actual engineering and project management) that resulted in a successful launch. Further, active launch team member experience was essential (i.e., on console with lead responsibilities during at least one successful countdown.)

The definitions of the root causes (though generally similar) varied somewhat in how they were characterized and discussed across the 21 cases (because the proximate causes were unique). It was therefore necessary to identify common “subroot causes” (at least two and up to four for each root cause) to ensure all aspects of each root cause was captured and properly categorized. Each failure summary was assessed on a qualitative basis (i.e., color coded) with respect to each subroot cause. Figure 5 explains the scoring scale. Initially, only a qualitative soring was pursued, as the main intent of this methodology was to alert the development program manager to those areas most in need of attention. A “green” score was assigned if there were minimal (or no) meaningful problems in that particular subroot cause area. A “yellow” score was assigned if problems appeared within a range of “correctable within existing program definition and resources” up to “prominent problems requiring prompt resolution” and possibly necessitating additional funding, staff, and/or schedule relief. A “red” score indicated even more serious concerns culminating in “serious problems threatening program viability” (either in technical feasibility or resource allocation).

As the methodology of this analysis evolved, a lead representative of the program (which was the impetus for developing this method) expressed a strong desire for a quantified measure of probability of failure. The method employed currently by the program (predicated on the mean time between failures of components used as input to a statistical analysis) produces a numerical result. Therefore, the representative desired a quantified (numerical) result, not just a qualified one. It was for that reason numerical values were introduced for each of the dozen root causes in conjunction with the color-coded scoring. Here the evaluator was free to specify any decimal value between 0 and 1, with the color-coded subroot causes used as a guide: ($0.0 \leq$ green $< 0.3$), ($0.3 \leq$ yellow $< 0.7$), and ($0.7 \leq$ red $\leq 1.0$) (Figure 5). An example of the scoring done on one of the 21 failure cases is in Table II: an assessment of the Titan IVB/Centaur-32 failure. The duality of this scale (color and numeric) allowed for either subjective or objective scoring. Although both means of scoring are subjective to varying degrees, it should be understood that what is being attempted to measure are human errors—which are, by definition, subjective. Although each root cause was distinct, it was recognized that they were not necessarily

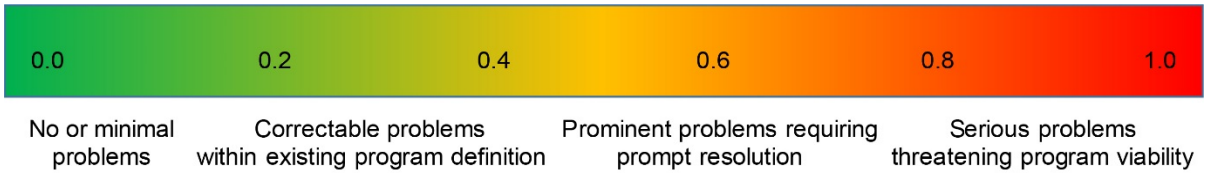


Figure 5.—Scoring scale for root causes of past launch vehicle failures (both qualitative and quantitative).

TABLE II.—SCORING OF ROOT CAUSES OF TITAN IVB/CENTAUR-32 FAILURE

| Root cause | Score | |
|---|---|---|
| | Subroot cause[a] (qualitative) | Root cause (quantitative) |
| Insufficient testing (integrated system) | | 0.70 |
| Lack of prudent integrated system testing | | |
| Not pursuing "test as you fly; fly as you test" | | |
| Insufficient understanding of interactions within entire system | | |
| Lack of test data of functioning system while in relevant environment | | |
| Engineering errors | | 0.60 |
| Faulty hardware design, fabrication | | |
| Incorrect analytical modeling, or computational errors | | |
| Ineffective systems engineering (SE) | | 0.00 |
| Inadequate SE, engineer judgment, understanding, resolution of critical problems | | |
| Insufficient meaningful reviews | | |
| Failure to challenge analyses, heritage, assumptions | | |
| Analytic models that are uncorrelated with actuals, ill- scaled, or of questionable validity | | |
| Insufficient testing (components, subsystems) | | 0.00 |
| Lack of prudent component and subsystem testing | | |
| Verification by analysis or comparison with requirements only | | |
| Heritage hardware and/or software: not validated for new application | | |
| Not establishing instrumentation needs | | |
| Process errors | | 0.80 |
| Fabrication, test, integration, or launch process not followed | | |
| Nonstandard events or work-arounds that have not incorporated into process | | |
| Hardware failure (flight or ground) | | 0.00 |
| Poor quality or statistically out-of-tolerance component | | |
| Multiple unforeseen program changes, environment changes, or secondary effects | | |
| Faster, Better, Cheaper | | 0.00 |
| Overworked staff due to imprudently short schedule | | |
| Imprudently low funding | | |
| Poor program management | | 0.00 |
| Lack of leadership integrity | | |
| Inattentiveness to (or ineffectiveness in) managing problems | | |
| Normalization of deviance (unexpected deviation, revised expectation) | | |
| Software failure (flight or ground) | | 0.80 |
| Differences between functional specifications and true requirements | | |
| Insufficient (or no) independent verification and validation (IV&V) | | |
| Poor team communication | | 0.65 |
| Organization-to-organization differences | | |
| Insufficient formality between working groups | | |
| Insufficient use of independent review team guidance | | 0.00 |
| Absence of independent assessment | | |
| Failure to heed or fully implement recommendations | | |
| Others | | 0.00 |
| International pressures | | |
| Loss of key leader without comparable replacement | | |
| Others | | |
| **Total** | | **3.55** |

[a]Green indicates there were no or minimal problems; yellow, moderate problems impacting program viability; and red, serious problems threatening program existence (see Figure 5).

independent of each other. But the complexity of quantifying the interdependencies was thought to result in too many hypotheticals and assumptions, so the root cause scores were merely summed to produce a resultant total root score. Further, since the distribution of root causes was fairly even (Figure 4), merely summing the individual root cause scores appeared to be reasonable.

## 4.7 Plotting of Resultant Root Cause Scores From Historical Launch Vehicle Data

Each of the failure cases listed in Table I was scored according to the method described in Section 4.6. The resultant total root scores were plotted in the order of increasing total score of root causes (Figure 6). Scores ranged from 0.10 (for Atlas/Centaur-24) to 6.25 (for Russian N-1 no. 4) where the maximum possible score was 12.0. Conveniently, a somewhat uniform distribution of scores resulted from the assessment even though no deliberate attempt was made to arrive at such a result. Even though no generalizations could be made of the results, by observation there did appear to be a rough grouping of the lowest scores by the unmanned Atlas and Titan vehicles, followed by the manned space shuttle and Apollo/Saturn vehicles, with the greatest scores for the Russian N-1 vehicles.

## 4.8 Derivation of the Cumulative Distribution Function to Calculate Probability of Failure

Because every nonzero score represented a case of a failed launch, and increasing nonzero scores represented increasing severity and/or diversity of human causal factors, the probabilistic approach to be applied needed to take into account both of these characteristics. A cumulative distribution function was chosen to calculate the probability of failure of conceptual vehicle concepts. Concepts would be scored similarly as with the historic cases in Section 4.6, and then the probability of failure calculated by finding the corresponding cumulative number of failures of historic cases with that score or lower. Note that if the cumulative scoring curve in Figure 6 were to be expanded to include all of the launches that were successful, then the first part of the curve (as well as the corresponding bar chart) would be identically zero for all these cases.

Since the cumulative distribution function can be set up as a probabilistic inequality where the independent variable can range from zero to some value, a two-conditioned cumulative distribution function can be set up as the difference between two cumulative distribution probabilities. These two probabilities are (1) the chance a score would be zero, representing the total number of successful launches out of the sample space and (2) the chance a score would be up to a nonzero score. These two bounding scores can be designated as "*a*" and "*b*." Subtracting these two probabilities would yield the probability that a conceptual design would be both a failure and have a score comparable to historic cases with similar severity and/or diversity of human causal factors. The probability of failure of a conceptual vehicle system would then be of the form given by the expression $P\{\omega \mid a < X(\omega) \leq b\}$, where the probability ($P$) of failure event ($E$) is a cumulative distribution ($F$) of ($\omega$) possible cases, and ($X$) is the random variable of interest (the total score of root causes), which can take on a value greater than ($a$) but less than ($b$). It is important to realize that the summation of the number of cases corresponding to the scores ($a$) and ($b$) are used to calculate the probabilities (and not the scores themselves).

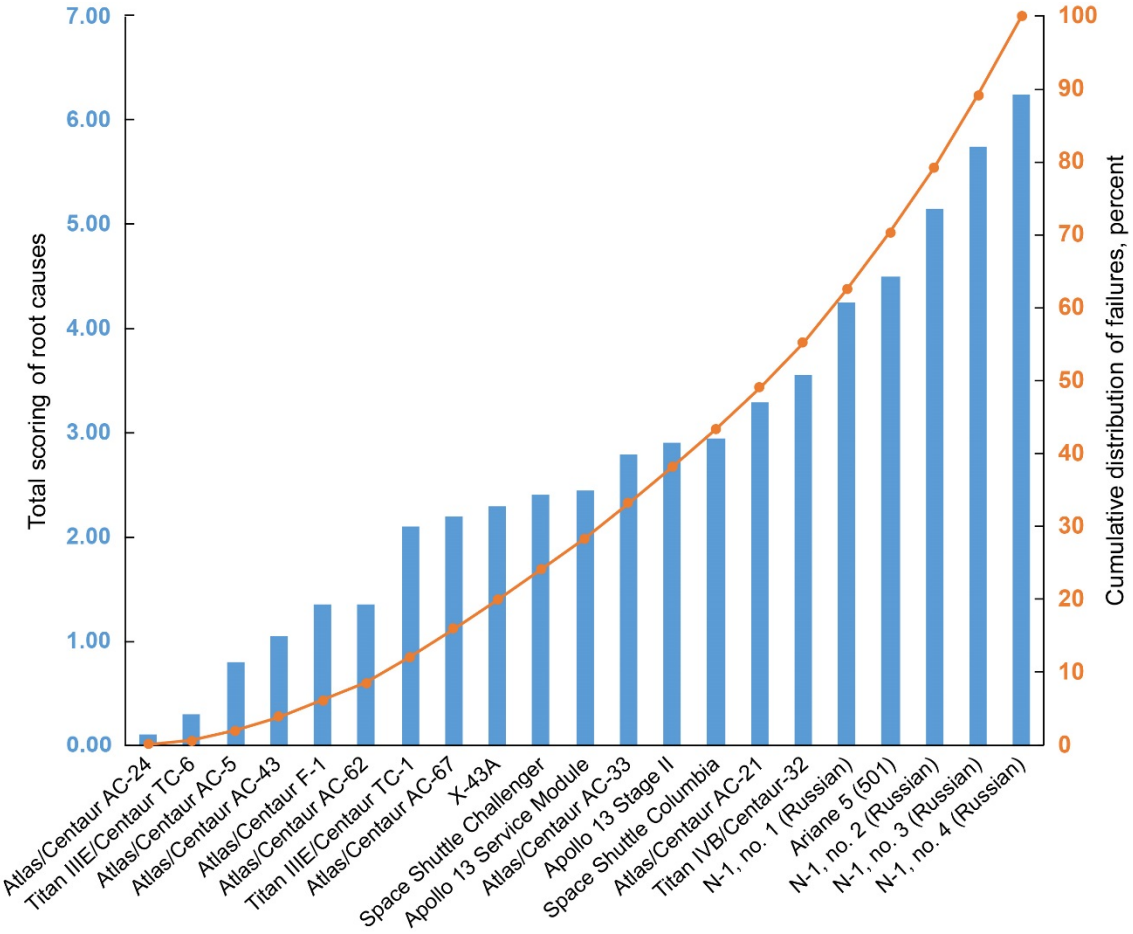


Figure 6.—Root cause totals per failure case and their cumulative percentage distribution.

Below is the derivation of the cumulative probability distribution function to be used to estimate the probability of launch vehicle system failure for future concepts. The cumulative distribution function $F$, where the random variable of interest $X$ is the total of the dozen causal sources of failure (and whose maximum numerical value $b$), is given by

$$F_X(b) = P\left(E_b^X\right) = P\{\omega \mid X(\omega) \le b\} \tag{1}$$

The probability of a successful case (i.e., score = 0) is expressed as

$$F_X(a) = F_X(0) = P\{\omega \mid X(\omega) \le 0\} \tag{2}$$

The number of case studies considered (the sample space Ω) is 359 (Table I). Within this sample space, there were 21 failures (i.e., 338 successes). Therefore, the probability of success of the entire sample space (where the maximum numerical value $a = 0$) is given by

$$F_x(0) = \frac{(359-21)}{359} = 0.9415 \tag{3}$$

The corresponding chance of failure is given by

$$(1-0.9415) = 0.0585 \tag{4}$$

which is approximately 1 chance of failure out of 17 attempts.

The probability of failure for a conceptual vehicle is the difference between the probability associated with its nonzero score and that of a zero score (i.e., success). For example: a concept with a score of 3.60 would lie between failed case no. 16 (score = 3.55) and failed case no. 17 (score = 4.25). There were 16 failures out of 338 + 16 = 354 launches whose scores were less than 3.60. The probability that a case is a failure and its score was less than 3.60 is given by Equation (5), where $F_x(a)$ corresponds to $F_x(0)$—the probability of a successful launch.

$$P(\omega \mid a < X(\omega) \le b) = F_x(b) - F_x(a) = \frac{[(359-21)+16]}{359} - 0.9415 = 0.9861 - 0.9415 = 0.0446 \tag{5}$$

This is approximately 1 chance of failure out of 22 attempts (i.e., a corresponding launch success of 95.5 percent).

## 5.0 Testing for Reasonableness of Probability of Failure Prediction

In order to test the reasonableness of the predictions of this methodology, a comparison with actual ground and flight test data from real vehicle systems was needed. It is important to underscore that this is an assessment of the total vehicle system (not a single failed subsystem, as in Sec. 4.6) prior to operation compared to its actual total failure record at the conclusion of its program. Admittedly, this is difficult to do in retrospect. The following example attempts to do just that. To test reasonableness of this failure probability prediction methodology, the assessment described in Section 4.3 had to be performed on a comprehensive system description of sufficient technical depth. One optional, but recommended, part of the scoring was the inclusion of comments and source references for each score given. Although similar comments were not provided in the scoring done in Section 4.6, this example contains these comments as a means to substantiate the rationale of the score assigned.


Figure 7.—G-Prime upper stage applications. (a) Shuttle/Centaur. (b) Titan IV launch vehicle.

The Shuttle/Centaur upper stage was a joint NASA–USAF program in 1981 to 1986 to develop two new configurations of the Centaur upper stage ("G" and "G-Prime") capable of launch from an orbiting space shuttle (Figure 7(a)). Although the program was cancelled only months prior to its first launch because of the aftermath of the Space Shuttle Challenger disaster, the essentially complete G-Prime configuration was immediately adopted by the USAF's new Titan IV booster program. Eventually, the G-Prime was launched 16 times on Titan IV from 1994 to 2003 (Figure 7(b)).

The vehicle's highly compressed original development schedule was driven by the requirements of its first two missions: both were to fly interplanetary trajectories whose 1986 launch windows could not be missed. The Shuttle/Centaur Preliminary Design Review (PDR) was followed by a Critical Design Review (CDR) only 9 months later. The aggregate data in those PDR and CDR packages were the most concise and comprehensive technical description of the program (Refs. 15 and 16). In addition, a book documenting the history of the Centaur upper stage had a comprehensive discussion of the technical problems encountered during its development (Ref. 17). These three sources served as the basis for scoring using this methodology.

Table III is the scoring of the Centaur G-Prime. It is seen here that although the Centaur was managed by the NASA Lewis (now Glenn) Research Center (LeRC), much of the NASA Johnson Space Center (JSC) management's actions and decisions negatively impacted Centaur development. Many of these JSC-initiated impacts are reflected in the scoring. There were several potential root causes of failure noted in the scoring, but the leading problems originated with the disparate approach to safety by the two managing NASA centers of the shuttle and Centaur stack (JSC and LeRC, respectively), which was due to concerns over the large cryogenic propellant upper stage in the cargo hold of the manned space shuttle. The significant score in Ineffective System Engineering, specifically in resolution of critical problems, stemmed from the fundamental disagreement between management of LeRC and JSC on critical fluid dumping requirements in case of an abort. These significant, major system changes driven by safety concerns continued throughout the development and even as final launch preparations began. A score of 0.70 was given because it continued to be a source of several prominent problems that required significant (and quick) resolutions. Safety problems were exacerbated by poor team communication, largely due to organization-to-organization cultural differences. While LeRC continually sought to resolve technical

TABLE III.—SCORING OF SHUTTLE/CENTAUR G-PRIME UPPER STAGE FAILURE

| Cause | Score | | Supporting references[a] |
|---|---|---|---|
| | Subroot cause[b] (qualitative) | Root cause (quantitative) | |
| Insufficient testing (integrated system) | | 0.50 | |
| Lack of prudent integrated system testing | | | No altitude propulsive stage test at 109%<br>Propellant-level indicating system mount failures<br>Centaur integrated support structure, erratic operation of propellant valves (Ref. 17, p. 206) |
| Not pursuing "test as you fly; fly as you test" | | | Structural dynamic test campaign, system integration laboratory for avionics hardware and software System Level III/IV |
| Insufficient understanding of interactions within entire system | | | Most of Centaur adopted or leveraged from existing, long-heritage Atlas/Centaur program |
| Lack of test data of functioning system while in relevant environment | | | Most of Centaur adopted/leveraged from existing, long-heritage Atlas/Centaur program |
| Engineering errors | | 0.00 | |
| Faulty hardware design, fabrication | | | System Level III/IV Program PDR (March 1983) and CDR (Dec. 1983) reports (Refs. 15 and 16) |
| Incorrect analytical modeling<br>Computational errors | | | System Level III/IV Program PDR (March 1983) and CDR (Dec. 1983) reports (Refs. 15 and 16) |
| Ineffective systems engineering (SE) | | 0.70 | |
| Inadequate SE, engineering judgment, or understanding; or failure to resolve critical problems | | | Repeated JSC safety-driven changes in critical fluid dump system interface between shuttle and Centaur |
| Insufficient meaningful reviews | | | System Level III/IV Program PDR (March 1983) and CDR (Dec 1983) reports (Refs. 15 and 16) |
| Failure to challenge analyses, heritage, assumptions | | | Repeated LeRC challenging of astronauts' liquid hydrogen concern with Centaur vs. shuttle external tank (Ref. 17, p. 197) |
| Analytic models uncorrelated with actuals<br>Ill-scaled or questionable validity | | | Modal survey performed on test article, trajectory design code based on past Atlas/Centaur flight data, and others |
| Insufficient testing (components, subsystems) | | 0.00 | |
| Lack of prudent component and subsystem testing | | | System Level III/IV Program PDR (March 1983) and CDR (Dec. 1983) reports (Refs. 15 and 16) |
| Verification by analysis or comparison with requirements only | | | System Level III/IV Program PDR (March 1983) and CDR (Dec. 1983) reports (Refs. 15 and 16) |
| Heritage hardware/software: not validating for new application | | | System Level III/IV Program PDR (March 1983) and CDR (Dec. 1983) reports (Refs. 15 and 16) |
| Not establishing instrumentation needs | | | System Level III/IV Program PDR (March 1983) and CDR (Dec. 1983) reports (Refs. 15 and 16) |
| Process errors | | 0.30 | |
| Fabrication, test, integration, or launch process not followed | | | Observed lower quality manufacturing, transport, and contractor staff actions (Ref. 17, pp. 209 to 210) |
| Nonstandard events, or work-arounds not incorporated into process | | | None identified |

TABLE III.—Concluded.

| | | | |
|---|---|---|---|
| Hardware failure (flight or ground) | | 0.20 | |
| Poor quality or statistically out-of-tolerance component | | | NA |
| Multiple unforeseen program or environment changes, or secondary effects | | | Change from "Element" to "Payload" designation drove critical hardware changes late in development |
| Faster, Better, Cheaper | | 0.50 | |
| Overworked staff due to imprudently short schedule | | | Contractor, LeRC leadership 50- to 70-hr weeks year after year (Ref. 17, pp. 196 to 198), short schedule in 1986 (p. 205) |
| Imprudently low funding | | | Joint NASA and U.S. Air Force funding at ~$2B (current-year funding over 4.5 years) (Ref. 18) |
| Poor program management | | 0.60 | |
| Lack of leadership integrity | | | LeRC securing 109% Space Shuttle Main Engine throttle baseline (Ref. 17, pp. 205, 208, 209) |
| Inattentiveness to (or ineffectiveness in) managing problems | | | JSC integration staff rather than JSC engineering staff delayed technical and safety responses (such as fill drain dump) |
| Normalization of deviance (unexpected deviation, revised expectation) | | | JSC shuttle lift capability versus commitment |
| Software failure (flight or ground) | | 0.00 | |
| Differences between functional specifications and true requirements | | | System Level III/IV Program PDR (March 1983) and CDR (Dec. 1983) reports (Refs. 15 and 16) |
| Insufficient (or no) independent verification and validation (IV&V) | | | System Level III/IV Program PDR (March 1983) and CDR (Dec. 1983) reports (Refs. 15 and 16) |
| Poor team communication | | 0.90 | |
| Organization-to-organization differences | | | JSC unresponsive to LeRC technical data requests due to differences in JSC staff cultures (integration group vs. engineering group) |
| Insufficient formality between working groups | | | Sufficient technical working groups between LeRC and General Dynamics Space Systems Division |
| Insufficient use of independent review team guidance | | 0.50 | |
| Absence of independent assessment | | | No Non-Advocate Review convened<br>Continued safety concerns by astronauts (Ref. 17, pp. 197 to 199 and 206 to 207) |
| Failure to heed or fully implement recommendations | | | NA |
| Others | | 0.00 | |
| International pressures | | | NA |
| Loss of key leader without comparable replacement | | | NA |
| Others | | | NA |
| **Total** | | **4.20** | |

[a]PDR refers to the Preliminary Design Review; and CDR, the Critical Design Review. JSC is the NASA Johnson Space Center; and LeRC, the NASA Lewis (now Glenn) Research Center.

[b]Green indicates there were no or minimal problems; yellow, moderate problems impacting program viability; and red, serious problems threatening program existence (see Figure 5).

problems stemming from the need to rapidly and safely dump propellants in the case of an abort, JSC was frequently nonresponsive to requests for technical data. Further, because of the designation of Shuttle/Centaur as a "Payload" rather than an "Element," it was the JSC integration staff rather than their engineering staff, who provided responses to LeRC. These responses were frequently unsatisfactory to help resolve engineering problems at the Shuttle-to-Centaur interface and so were a continuous source of major problems; thus, a score of 0.90 was assigned.

More moderate problems existed in four other areas that may or may not have been resolvable within the existing program budget and schedule. No propulsive altitude testing of the entire stage of Shuttle/Centaur was performed (Ref. 17). Propellant system failures and erratic behavior became apparent late in the development, as exhibited by the Propellant Level Indicating System mount failures and Centaur Integrated Support System propellant valve operation, respectively. There was no non-advocate review prior to the program start, which presumably would have surfaced some of the liquid hydrogen safety issues. In the area of poor program management, while LeRC management was proactive and determined to resolve intractable problems, the evidence of repeated delays, unresponsiveness to data requests, and inappropriateness of integration rather than engineering staff involvement on the part of the JSC management warranted at least a 0.60 score. Further, in the area of normalization of deviance, it had become commonplace for JSC to issue shuttle lift commitments that were not documentable and indeed incapable of being technically substantiated. This resulted in serious problems performing trajectory design and performance analysis by the Shuttle/Centaur program staff at LeRC. This also contributed to the 0.60 score. Lastly, the Shuttle/Centaur program achieved an admirable feat by going from proposal material to complete flight-configured stages at the Cape, being prepared for launch in a mere 4.5 years. The impressive technical progress in such a short period of time was evident in the major review documentation (Refs. 15 and 16). However, this was accomplished with considerable overtime by most of the leadership and many of the staff (Ref. 17). The favorable scores of zero (engineering errors, component testing, and much of the system-level testing) could be attributed to considerable contractor and NASA center technical expertise brought in from the operational Atlas/Centaur system to staff the new program.

The resultant total system score of 4.20 produced a probability of failure of 4.46 percent. The final record of the Centaur G-Prime upper stage on the Titan IV booster was 14 successes, 1 failure, and 1 "no-trial" (failure prior to Centaur phase). Thus, the actual system failure rate of 6.67 percent compared reasonably well with the predicted value. However, the most important result was the significantly different qualitative scoring of almost every subroot cause when compared to the Titan IVB/Centaur-32 failure, even though the G-Prime upper stages were essentially the same. A likely explanation was the change in organizations. The Shuttle/Centaur of the 1980s was developed by NASA LeRC and General Dynamics, while the failure in 1999 came after the transfer to USAF Space Division and the Lockheed Martin Corporation purchase of General Dynamics Space Systems Division. This methodology presumes consistency in organization and staff. When major corporate changes occur, this new method may not sufficiently account for that change.

## 6.0 Potential Future Applications

The Defense Advanced Research Projects Agency (DARPA) Experimental Spaceplane Program (XSP) is a currently in-development reusable booster. It is intended to be capable of 10 suborbital flights in 10 days, as well as hypersonic cruise missions up to Mach 10. It must also be capable of accommodating an expendable upper stage to perform low-Earth orbit missions. It has a cost-per-flight requirement of $5M (amortized over a reasonable, finite period). This program was the original impetus

for the development of the quantitative scoring methodology. It is currently under consideration for incorporation to some extent in order to further increase the likelihood of launch success.

The promising new commercial launch vehicles such as SpaceX's Falcon 9 and Blue Origin's New Glenn could also profit from this approach, since infant mortality still appears to be a factor. The existing legacy expendable launch vehicles (Atlas V and Delta IV) continue to fly and still undergo modifications and could also benefit. NASA's current Space Launch System and Orion programs have been repeatedly delayed, and costs continue to escalate (Refs. 19 and 20). This new methodology could help direct changes to improve their likelihood of success. Finally, this method can be generalized and applied to different types of space propulsive systems (such as in-space electric propulsion).

## 7.0 Caveats and Concerns

There were several concerns raised about this methodology by staff of the NASA Headquarters Safety Center; the NASA Glenn Research Center Safety, Reliability and Quality Assurance Branch; ARES Corporation; and The Aerospace Corporation. While generally acknowledging the shortcomings of the more traditional methods and the need for a method such as this one in principle, they urged caution in several areas. The authors have accepted many of their suggestions and introduced solutions into the methodology as a result. Other concerns were either rejected or merely noted, with reasons given here.

It was pointed out that successful launches, if subjected to this assessment, would likely result in nonzero scores as well. That is, no successful launch is exactly nominal, and failing to incorporate these "nonzero score successes" into the cumulative distribution function is not strictly correct. Although this is true, the source database did not contain evaluations of successful missions. Thus, this methodology produces a "floor" to the probability of failure rather than a "ceiling." To address this concern, scoring the 338 postflight reports of successful missions would be needed, just as in the cases of the accident investigation board reports of the 21 failed missions. This would require a considerable amount of effort.

This technique (like most that were discussed in Secs. 3.1 and 3.2) focuses on "errors"—the negative actions taken (or not taken). Positive actions (adaptations to new information or feedback loops in decision making) by people are typically not incorporated into these methodologies, yet are important in the correct representations of what actually takes place. Adaptations and feedback loops (internal and external to systems) are widely acknowledged as essential for successful outcomes, and their omission represents a meaningful modeling deficiency in assessments of probability of failure. "Failure to consider successful versus unsuccessful adaptations prevents comprehensive understanding of human behavioral variability" (Ref. 21).

It was asserted that the sample set was incomplete. That is, it should have also included launch scrubs and delays rather than just failures. This assertion was rejected due to the added seemingly infinite amount of "what if" speculation that would follow. What if a delay was followed by another delay of no attribution to the system, which results in a failure? Is that the fault of the system or not? Which indirect delays should be attributed to the system?

"Color coded" results were generally thought helpful, but the numerical scoring was thought by some to imply a precision that did not exist or was largely subjective. As a result, both scoring methods were retained.

It was pointed out that existing methods such as Failure Modes Effects Analysis, Fault Tree Analysis, Human Reliability Analysis, and others can already accommodate human factors and should be sufficient to address human causal issues. However, these methods were rejected after consideration because of their anticipated resource-intensive needs (people, time, and funding) if used to evaluate an entire launch vehicle system.

Another concern was the small sample size of 21 launches used as the basis for this method. While this suggests a moderately significant statistical error, it should be recognized that the 16 spacecraft missions exhibited similar failures for comparable root causes. Thus, a larger sample size of 21 + 16 = 37 might be inferred. Further, these are not all identical vehicles, but rather similar vehicles flying different spacecraft on different missions. Statistical methods predicated on samples taken from identical elements within a sample space may not be appropriate. What is important is a large enough sample space of failures so that no category of root causes was overlooked.

The scoring was greatly influenced by sample space definition: the greater the number of failures considered and included in the source data described in Section 4.6, the greater the range of potential scores and range of probabilities of failure. In this sample space, the greatest probability of failure was 5.85 percent (corresponding to a score of >6.25). Some "infant mortality" cases were not included, which likely reduced the range of potential failure probabilities. The scoring could be made more representative of history by including those cases.

As was discussed in Section 5.0, a potential major weakness can arise when there is a change in the organization that either leads the development or performs the launch operations (or both) between the time of application of the method and the launch system's operation. Implicit in this method is the presumption that there is minimal change in the organization. Negating that presumption could greatly compromise the prognostication.

Lastly, the greatest vulnerability to criticism for this methodology might be "20-20 hindsight bias" in the scoring. Comprehension of the circumstances surrounding the failure is even more important than judging past actions as imprudent or insufficient. Failure (mishap) reports frequently do not describe in great detail the various options available to the launch directors, their knowledge, or various competing issues that are being struggled with during the pressure-intensive countdown. The obvious poor decision in hindsight frequently appears to have been the correct decision in the heat of the moment. Because of this, reliance on (even) complete accident investigation board reports and experts with impressive comprehensive experience can still be subject to serious, credible criticism (Ref. 21).

# 8.0 Summary and Conclusions

A considerable number and variety of analytic methodologies exist to forecast the probability of failure for a major engineering system. Most of these methods are focused on component hardware and are statistical in nature. However, it has been shown repeatedly that the root causes of the overwhelming majority of past launch vehicle failures are human causal factors, not hardware unreliability, that are manifested in a statistical way. Although Probabilistic Risk Assessments, particularly when augmented with Human Reliability Analyses are effective, established methods to determine causes of failure for specific subsystems, they can be unwieldy and resource intensive if used system-wide to predict all likely means of failure for a launch vehicle system still in development.

A practical, prognostic method based on actual root causes of past failures has been created that can be applied to an entire launch vehicle system. Even though it is lacking in precision and strict statistical orthodoxy, it is relatively easy to use to generate either qualitative or quantitative results. Its baseline formulation is predicated on data from past accident investigation board reports and judgment by two nationally recognized experts in launch vehicle development and operations. A cumulative, probabilistic distribution function was generated from that analysis. Using that function and scoring based on proven human-centric root causes, the method's resultant predictions of probability of failure for an example case was shown to be in reasonable agreement with demonstrated actual performance of the completed launch vehicle program. However, the qualitative scoring of the predicted subroot causes of failure were

significantly different compared to the actual causes of a failed mission. This was attributed to the significant changes in Government and industry leadership and execution of the program, which took place between the time of prediction and time of failure.

This new methodology is currently under consideration by a DARPA launch vehicle development program. It could be used in other Government and commercial launch vehicle programs now in varying stages of development or upgrading, to assist program management in mitigating the true root causes of launch vehicle system failure. Although a numerical score from a failure risk assessment will never be actually verified because of the relatively small number of space launches (unlike aircraft or other vehicles), the enhanced focus on actions to mitigate human casual factors identified through this method should meaningfully improve reliability of future launch vehicle concepts.

## References


1. Hamlin, T., et al.: Shuttle Risk Progression: Use of the Shuttle Probabilistic Risk Assessment (PRA) to Show Reliability Growth. AIAA 2011–7353, 2011.
2. Office of Safety and Mission Assurance: NASA Procedural Requirements for Mishap and Close Call Reporting, Investigating, and Recordkeeping. NASA Procedural Requirements NPR 8621.1C, 2016.
3. Titan IVB-32/Milstar-3 Accident Investigation Board Report. USAF Form 711, USAF Mishap Report, 1999.
4. Chandler, Faith T., et al.: Human Reliability Analysis Methods: Selection Guidance for NASA. NASA/OSMA Technical Report, 2006.
5. Stamatelatos, Michael; and Dezfuli, Homayoon: Probabilistic Risk Assessment Procedures Guide for NASA Managers and Practitioners. Second ed., NASA/SP—2011-3421, 2011. http://ntrs.nasa.gov
6. Leveson, Nancy G.: Engineering a Safer World: Systems Thinking Applied to Safety. MIT Press, Cambridge, MA, 2011, pp. 169–249 and 469–493.
7. Huang, Zhaofeng; Fint, Jeffry A.; and Kuck, Frederick M.: Key Reliability Drivers of Liquid Propulsion Engines and a Reliability Model for Sensitivity Analysis. AIAA 2005–4436, 2005.
8. Gernand, Jeffrey L., et al.: Constellation Ground Systems Launch Availability Analysis: Enhancing Highly Reliable Launch Systems Design. AIAA 2010–2180, 2010.
9. Morse, Elisabeth L.; Fragola, Joseph R.; and Putney, Blake: Modeling Launch Vehicle Reliability Growth as Defect Elimination. AIAA 2010–8836, 2010.
10. Guikeme, Seth D.; and Pate-Cornell, M. Elisabeth: Bayesian Analysis of Launch Vehicle Reliability. AIAA 2003–1175, 2003.
11. NASA Reliability and Maintainability (R&M) Standard for Spaceflight and Support Systems. NASA–STD–8729.1A, 2017.
12. Nieberding, Joe; and Ross, Larry: Lessons Learned Applied To Space System Development. Presentation, Vol. 1, Version 1.0, Aerospace Engineering Associates LLC, Bay Village, OH, 2006.
13. Nieberding, Joe; and Ross, Larry: Mission Success First: Lessons Learned. Aerospace Engineering Associates LLC, presented in conjunction with the “Jornadas sobre Calidad Espacial” Class #100, Buenos Aires, Argentina, 2016.
14. NASA Mishap Investigation Board: Astro-E2/Suzaku X-ray Spectrometer. Mishap Report Type A Mishap, IRIS No. 2005–273–00003, Table 7–1, 2005. Available from the Suzaku Mishap Investigation Board.
15. Shuttle/Centaur Level III/IV Program PDR at LeRC. General Dynamics Convair Division, San Diego, CA, 1983.

16. Shuttle/Centaur Level III/IV Critical Design Review at LeRC. General Dynamics Convair Division, San Diego, CA, 1983.
17. Dawson, Virginia P.; and Bowles, Mark D.: Taming Liquid Hydrogen: the Centaur: Upper Stage Rocket, 1958–2002. NASA SP–2004–4230, Ch. 7, 2004, pp. 189–219. http://ntrs.nasa.gov
18. NASA Glenn Research Center: STS/Centaur Program Funding, internal presentation, Dec. 9, 1986.
19. NASA's Plans for Human Exploration Beyond Low Earth Orbit. NASA Office of Inspector General Report No. IG–17–017, 2017.
20. NASA Human Space Exploration. U.S. Government Accountability Office, Report to Congressional Committees GAO–17–414, 2017.
21. Lilley, S.: NASA Safety Center, Cleveland, OH, personal communication, Dec. 21, 2017.