

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

Craig H. Williams ¹

NASA Glenn Research Center, Cleveland, OH, 44135, USA

Lawrence J. Ross ² and J. Joe Nieberding ³

Aerospace Engineering Associates LLC, Bay Village, OH, 44140, USA

Lessons learned from past failures of launch vehicle developments and operations were 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 were 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 were 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 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: a numerical score of the probability of failure or 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 (USAF's Titan IV with the Centaur G-Prime upper stage) was used as a test case. Although the theoretical vs. actual probability of failure was found to be in reasonable agreement (4.46% vs. 6.67% respectively) the underlying sub-root 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.

I. 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)
Ω	=	sample space
ω	=	possible cases

II. 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 are generally human-caused, rather than hardware component failure or other manifestations of poor quality control.

¹ Aerospace Engineer, Propulsion Division, Associate Fellow AIAA

² Chief Executive Officer

³ President, Senior Member AIAA

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. While Probabilistic Risk Assessments (PRA) 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 PRA's performed a dozen years after the first Shuttle launch (and also after the Challenger 51-L failure) produced failure probabilities many times more favorable than actual risk progression (calculated after Shuttle retirement) [1]. As will be discussed, much of the optimistic assessments could be attributed to the lack of addressing human-causal factors.

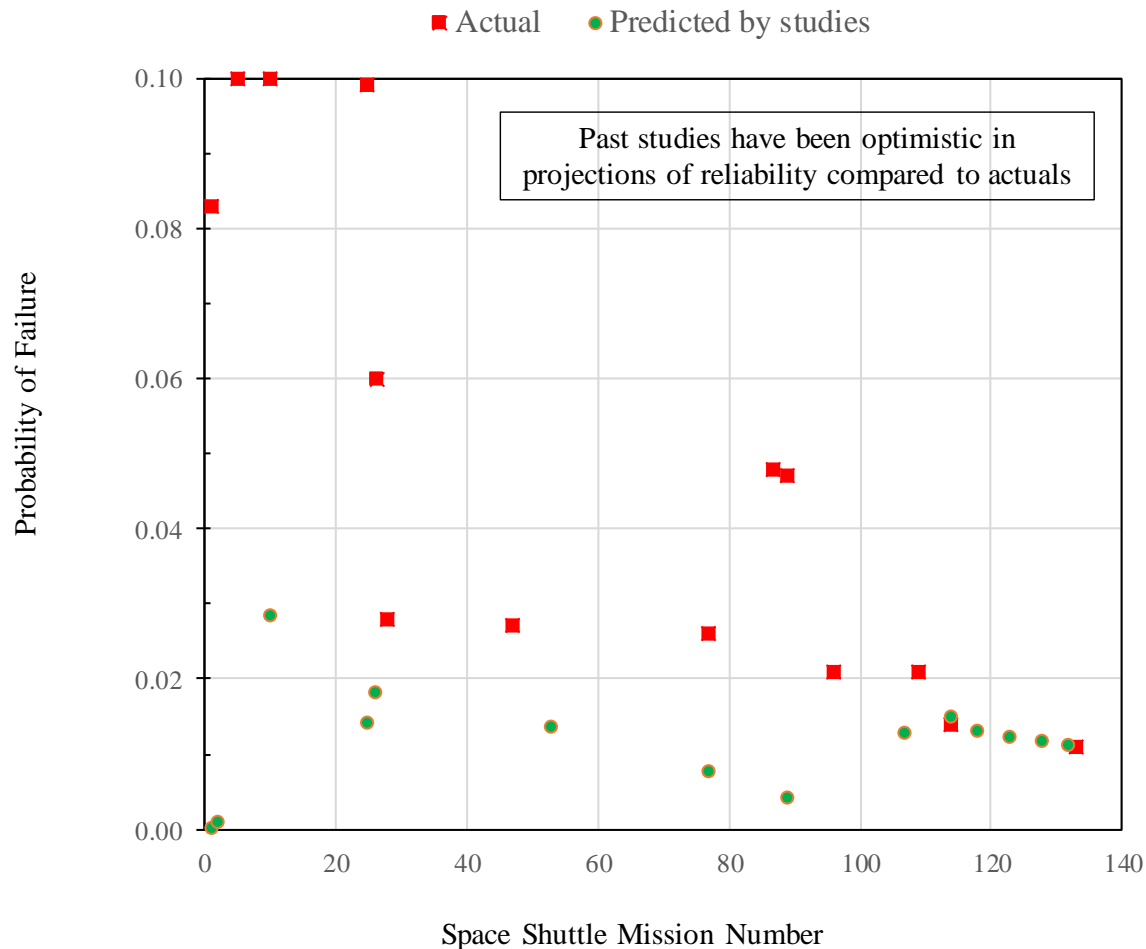


Figure 1: Actual vs. Predicted Probability of Failure for Space Shuttle System

III. “Proximate” vs. “Root” Causes of Failure

The distinction between “proximate” and “root” causes is given in NASA’s Procedural Requirements [2].

“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).”

“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 Titan IVB/Centaur launch of a Milstar spacecraft failure in 1999, the vehicle tumbled during 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” [3]:

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

IV. Existing Methods to Assess Probability of Failure

It was reasonable to assume that such a mature field would have created methods to assess probability of failure for entire aerospace systems which 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, as well as literature searches performed on the subject. Two comprehensive documents were identified and reviewed.

A. NASA/HQ Study of Human Reliability Analysis Methods

There are approximately fifty different methods to assess and predict complex system probability of failure developed over the past half-century [4]. Most of these methods were created to assist the nuclear power industry and are largely hardware-centric. Out of these fifty methods, fourteen were selected by NASA Headquarters (HQ) for further study on their applicability for launch vehicle failure assessments [4]. This subset was predicated on methods which 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 labs, and the private sector. (Note that the existing NASA PRA guidance provides a method similar to those practiced by industry [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. But the HRA process (problem definition, task decomposition, and the identification / representation / quantification of human error) was most readily applied to ‘bottoms-up’ initial design, analysis of individual tasks, and operating specific components/systems. The two initial HRA steps can become quite complex if not applied to clearly defined problems which 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) [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 sub-systems which have a limited number of operation steps. HF PFMEA methodology then defined all possible combinations of acts which a person could make in order to correct ‘bad’ sequences of events. This produced a considerable number of possible scenarios and actions making it unwieldy for system-wide application on a conceptual design. In addition, HF PFMEA did not calculate human error probabilities (a primary reason for it not being further considered in the HQ study).

The HQ’s study chose four of the fourteen 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 HSA/PRA process became even more tedious and complex for problems beyond a finite sub-system 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.

B. STAMP and STPA

In a comprehensive assessment external to NASA, the Systems-Theoretic Accident Model and Processes (STAMP) was found to be a viable prospect using an all-encompassing accident model based on systems theory [6]. STAMP both analyzed accidents after they occurred and created approaches to prevent them from occurring in developing systems. This method was not focused on failure prevention per se, but rather than 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 took exception to making a distinction between “proximate” and “root” causes, maintaining that the distinction was at least artificial and at most an obstacle to discovering the true causes of unsafe operations. STAMP’s top-down approach guided the user to produce ‘safety constraints’ and ensure that they were 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 (Fig. 2). STAMP was very comprehensive, analyzing not just immediate causes of the failure but also the societal, organizational, cultural, and governmental environments surrounding it.

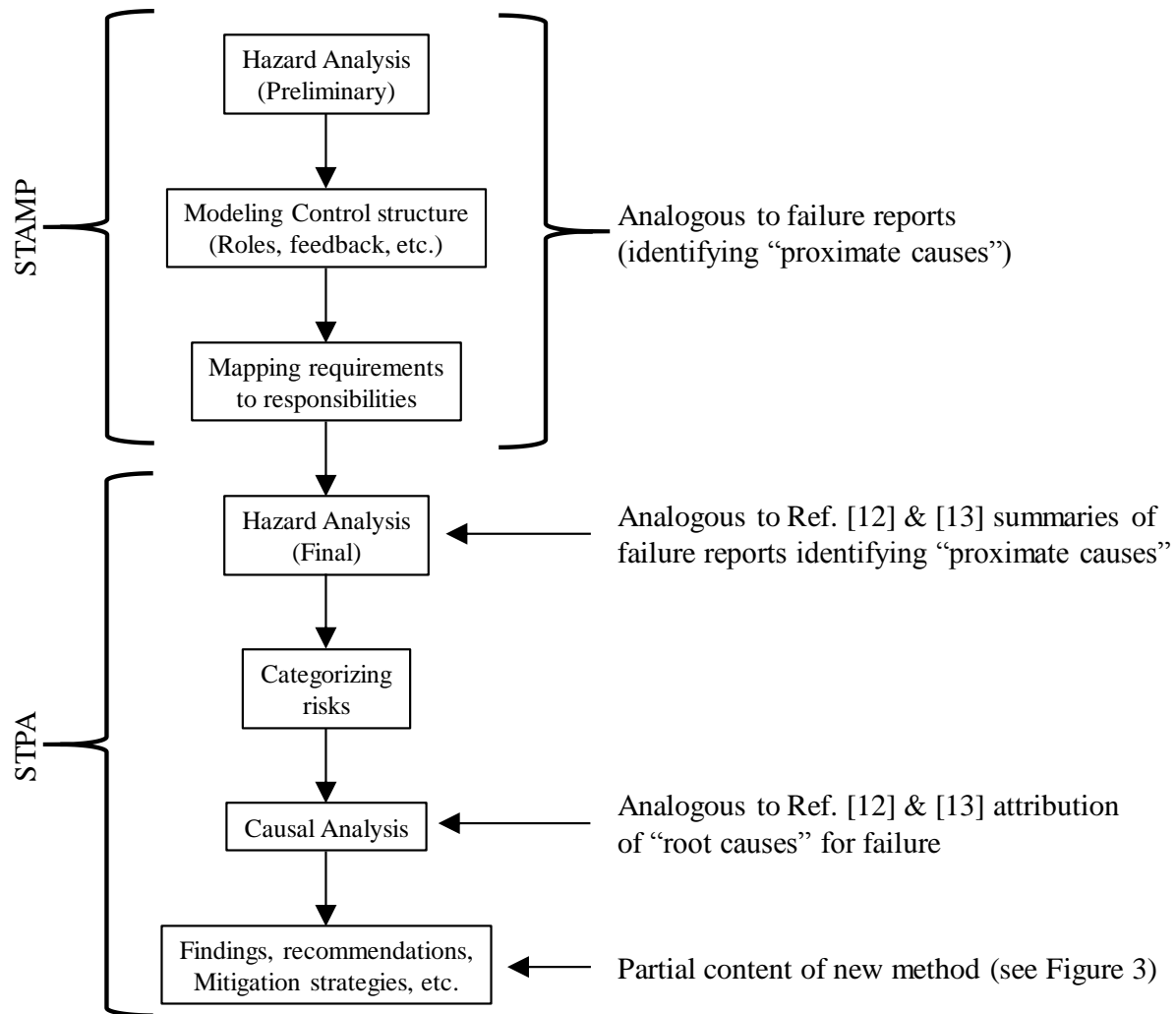


Figure 2: Comparison of STAMP/STPA to New Method

The predictive part of the problem (a "hazard analysis") was built on STAMP using it as a 'front end'. 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 [6]. A primary output of this method were "functional control diagrams" rather than component (i.e. hardware) diagrams. The two overall aspects of STPA were: 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 [6].

STAMP/STPA guided the design process rather than require the design to exist before hand, thus made it attractive for conceptual design applications. Thus the composite STAMP/STPA method, based largely on human factors, might be used for more accurately predicting system-wide probability of failure. However, as with the four methods assessed in the NASA HQ study, the STAMP/STPA's exhaustively detailed nature could drive analysis to become unwieldy if not narrowly tailored to a specific sub-system. For example: Ref. [6] applied STAMP to the failed Titan IVB/Centaur-32 launch, narrowly focused only on the proximate cause of the failure: faulty flight software development and insufficient testing. Yet the resulting STAMP processes for assessing the development, launch operations, ground staff, and software required thirty pages of text to summarize [6]. If this approach were used to assess the entire launch vehicle system, it would be overwhelming. Other examples were provided in the same reference which were similarly focused on the specific cause of failure, not a system-wide analysis. While 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 [6]." Thus for similar reasons as the NASA HQ study, it did not appear that

STAMP/STPA could be concisely applied to perform major system-wide assessment of probability of failure of a conceptual design. Reference 6 went further, suggested that attempting to quantify future human actions impacting system reliability may not even be possible due to the unpredictability of human interaction with the surrounding conditions (questioning this paper's premise) [6]. We rejected that assertion and proceeded to create a new method.

C. 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 [7]. Here assumed sub-system 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 sub-systems [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 [9]. More sophisticated prognostication methods of this approach also exist [10].

Various NASA program standards and guidelines now proscribe that some type of human factors assessments be performed, with no preferred practices. Both Human Error Risk Assessments and Human Factor Task Analysis of some type are required to be performed, where the latter must, "Analyze and list all the things people will do in a system, procedure, or operation" [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 predicated on human causes.

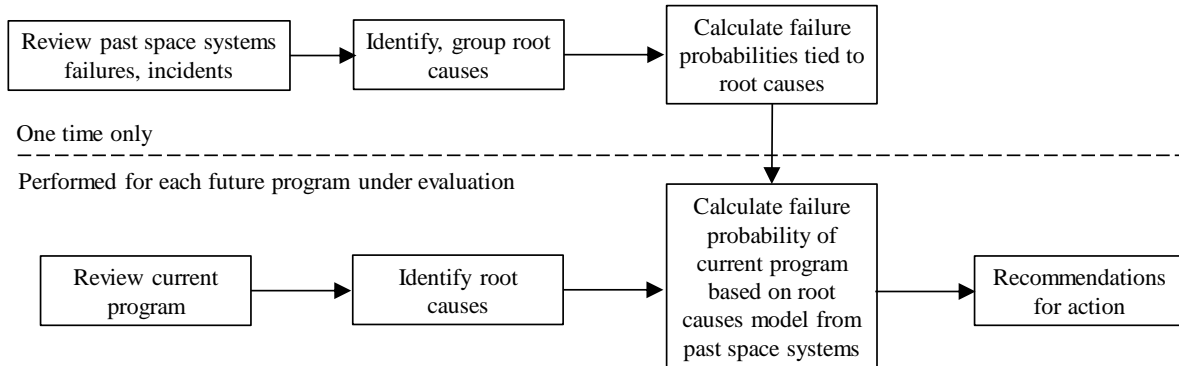


Figure 3: Approach of New Method

V. Proposed New Method

Since an existing, straightforward technique could not be found which 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 test 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) Establishing new method's basis
 - a. Review of past proximate causes of launch vehicle failures
 - b. Establishing root causes of past launch vehicle failures based on expert judgment
 - c. Categorizing, consolidating similar root causes into finite categories
 - d. Establishing baseline model using root causes of past launch vehicle failures
 - i. Selection of cases to be used
 - ii. Scoring of root and sub-root causes
 - iii. Plotting resultant data
 - e. Derivation of function for probability of failure of launch system
- 2) Application of 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 “Establishing new method’s basis” is a one-time only effort, reliant on the co-authors’ experience and judgement. The second part of this process “Applying 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 Fig. 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 the Fig. 2 (Findings, Recommendations, etc.) corresponds with (part of) this new methodology.

A. 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) [11, 12]. They are the co-authors of this paper. Over the course of their thirty-plus year careers, they successfully led launch vehicle development programs and actively served in leadership roles on more than sixty launch teams. It is this comprehensive experience which was fundamental to establishing the credibility of this new method. At AEA, they reviewed and assessed over fifty NASA and international case studies of launch vehicle and spacecraft failures, and 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, 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.

B. Establishing Root Causes of Past Launch Vehicle Failures Based on Expert Judgment

After analyzing the failures and their proximate causes, the co-authors developed specific actions to remedy the mistakes. It is the absences of these specific actions that can be viewed as the root causes of the failures. Among their findings was that the type of aerospace system (launch vehicle, spacecraft, major ground test site, etc.) did not matter with regard to the nature of the root causes. Also, root cause types did not change with time. What did matter was that human-root causes dominated failed hardware-root causes. Indeed, they state, “An examination of space mission and other mishaps finds human error to be a dominant factor [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: while methods to assess probability of failure tend to be hardware focused, the root causes tend to be human-centric. While 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, so 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.

C. Categorizing, Consolidating Similar Root Causes into Finite Categories

There have been efforts in the past to categorize and consolidate similar root causes. The report of the Mishap Investigation Board of the ASTRO E-2 mission in 2005 had a graphic which illustrated “Recurring Project Themes from Previous Mishaps”, which documented 28 distinct possible root causes [14]. The co-authors of this study have published an earlier presentation with (only) four distinct causation categories (one of which was sub-divided into six sub-categories) [13]. Upon reflection of the results in Section V-B 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 which mattered: insufficient testing, reliance on prior similar design requirements, erroneously assuming that implicit limits did not apply, etc. Testing was separated into two categories: system and sub-system/component. This was because system testing is designed to pick up integration and ambient environment issues, while sub-system/component testing is largely focused on individual self-functionality. Hardware and software failure root causes (the type which receives 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 sub-groups 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 so-called “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 increase significantly. Test data of an operating system while in relevant environment (thermal, vacuum, vibration, etc.) is 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 are 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 base decisions. Analytic models not correlated with actuals, 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 sub-system testing is also a major root cause of failures. Prudent testing prior to integration permits discovery while each sub-system/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/software may appear to save money and effort, but not validating it for new application, range of operation, or a new environment can risk significant cost and schedule downstream. Lastly, forgoing lower level testing can miss the opportunity to establish instrumentation needs which are typically first brought to light during sub-system level testing.
- 5) Failure to follow established processes (or errors in processes) span fabrication, test, integration, and launch operations. Non-standard events, loosely controlled changes, and workarounds not formally incorporated into 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, 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 1990's 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 which 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 judgement, 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 continued 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 which later proved program-lethal.
- 11) Independent reviews are intended to surface problems which are complex, cross many department/systems lines, too subtle for all but the most experienced staff to identify, and/or have escaped all customary checks and balances. Sometimes reviews take on an air of 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. 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. 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.

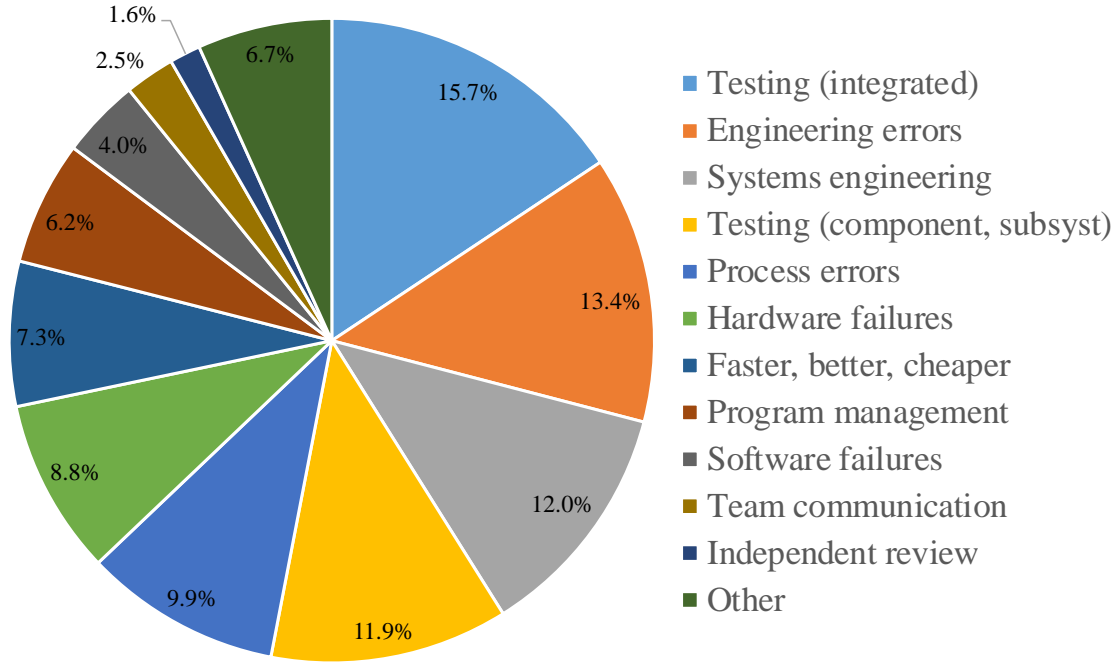


Figure 4: Distribution of Root Causes in Launch Vehicle System Failures

D. Assessment of Root Causes

The root causes of each failure case were grouped into the twelve categories described above as similarities became apparent. The groupings were then tallied and are shown in Fig. 4. There was a fairly even distribution among the human failing-root causes with no clearly dominant category. The leading root causes were lack of sufficient testing (both integrated system and component), lack of appropriate systems engineering, and engineering & process errors, altogether totaling 63%. All human-factor root causes amounted to 87% while hardware failures, by contrast, contributed less than 9%. In fact, there was only a single case where random part failure was the sole root cause [13]. These results indicated 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 confirmed that a multi-faceted approach focused on a variety of human-causal factors was needed.

E. Down Selection of Cases to be Used for the Basis of the Method

Not all cases that were assessed in References 12 & 13 were used. 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 V-H): how to account for the numerical total of spacecraft in the sample space? While 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 Intelsat's be grouped together or just within series? How should 'one-of-a-kind' interplanetary spacecraft such as Galileo be treated? While 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 were a small but adequate number launch vehicle cases, a practical decision was made to exclude spacecraft in the analysis-space. Another concern over which launch vehicle cases should be included was raised with respect to using only "operational" vehicles and avoid "Test or R&D infant mortality". But 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 & 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 1).

Table 1: Selected Failure Case Studies of Launch Vehicle Systems

		Mission	Problem	Result	Number in Series	Description of Total Number in Series
Research & Development						
	1	Atlas/Centaur F-1	Premature sheild seperation	Loss of mission	8	Test flights: 7 LeRC led + F-1
	2	Atlas/Centaur AC-5	Premature booster engine shutdown	Loss of mission, pad		See A/C F-1
	3	N-1 #1 (Russian)	Stage 1 failure	Loss of mission	4	Four N-1's in series
	4	N-1 #2 (Russian)	T - 0 explosion	Loss of mission, pad		See N-1 #1
	5	N-1 #3 (Russian)	Uncontrolled roll	Loss of mission		See N-1 #1
	6	N-1 #4 (Russian)	POGO	Program termination		See N-1 #1
	7	Titan IIIC/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	LOX tank explosion	Loss of mission	20	Total Service Module flights
	2	Apollo 13 Stage II	POGO	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 seperation failure	Loss of mission	61	Total non-test flight A/C up to 1990 (AC-69)
	5	Atlas/Centaur AC-24	Avionics hardware failure	Loss of mission		See A/C-21
	6	Atlas/Centaur AC-33	Loss of control	Loss of mission		See A/C-21
	7	Atlas/Centaur AC-43	Booster engine failure	Loss of mission		See A/C-21
	8	Atlas/Centaur AC-62	Loss of control during coast	Compromised mission		See A/C-21
	9	Atlas/Centaur AC-67	Lightning strike	Loss of mission		See A/C-21
	10	Space Shuttle Challenger	SRM 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 IIIC/Centaur TC-6	Stage 2 LOX 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
					359	

F. Scoring of Root and Sub-Root Causes with Requisite Expertise

The scoring for the first part of the methodology was done based on judgement of the root causes identified in References 12 & 13. “Expert judgement” credibility of both co-authors 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) which 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 “sub-root 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 sub-root 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 sub-root 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’ 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).

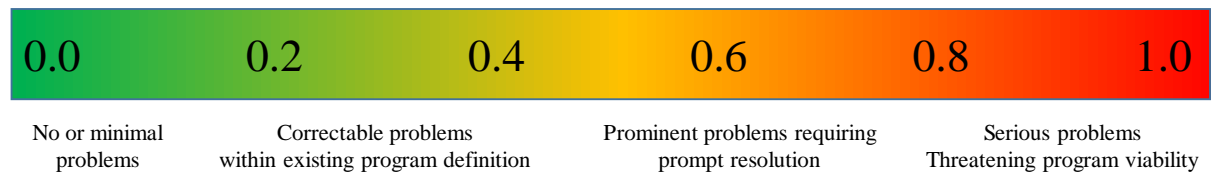


Figure 5: Scoring Scale for Root Causes of Past Launch Vehicle Failures

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 desire was for a numerical result similar to the product of the method which was currently in use (predicated on the mean time between failures of components used as input to a statistical analysis). 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 zero and one, with the color-coded sub-root causes used as a guide: (0.0 ≤ ‘green’ < 0.3), (0.3 ≤ ‘yellow’ < 0.7), and (0.7 ≤ ‘red’ ≤ 1.0) (Fig. 5). An example of the scoring done on one of the 21 failure cases is in Table 2: an assessment of the Titan IVB/Centaur-32 failure. The duality of this scale (color and numeric) allowed for either subjective or objective scoring. While both means of scoring are admittedly subjective, it should be kept in mind that what is being attempted to measure are human errors --- which are by definition subjective. While each root cause was distinct, it was recognized that they were not necessarily independent of the others. 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 (Fig. 4), merely summing the individual root cause scores appeared reasonable.

Table 2: Scoring of Root Causes of Titan IVB/Centaur - 32 Failure

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

G. Plotting of Resultant Root Cause Scores from Historical Launch Vehicle Data

Each of the failure cases listed in Table 1 were scored according to the method described in Section V-F. The resultant total root scores were plotted in the order of increasing total score of root causes (Fig. 6). Scores ranged from 0.10 (for Atlas/Centaur-24) to 6.25 (for Russian N-1 #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. While 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.

H. Derivation of the Cumulative Distribution Function to Calculate the Probability of Failure

Because every non-zero score represented a case of a failed launch, and increasing non-zero 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 V-F above, 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 Fig. 6 were to be expanded to include all of the successful launches, then the first part of the curve (as well as the corresponding bar chart) would be identically zero for all these cases.

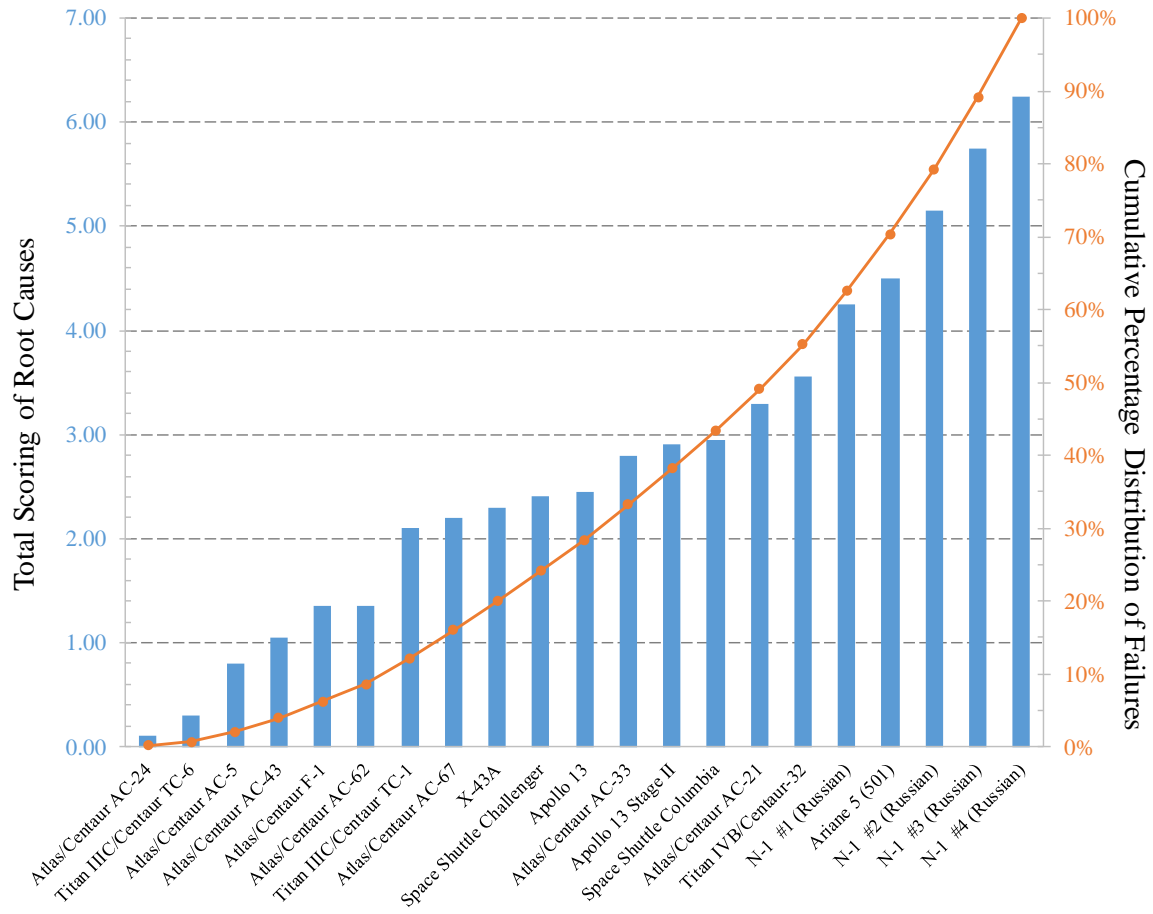


Figure 6: Root Cause Totals per Failure Case and Their Cumulative Percentage Distribution

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: the chance a score would be zero (representing the total number of successful launches out of the sample space) and the chance a score would be up to a non-zero score (of the conceptual design). 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. These two bounding scores can be designated as ‘a’ and ‘b’. 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 (ω) 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).

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:

$$F_x(b) = P\{E_b^X\} = P\{\omega \mid X(\omega) \leq b\} \quad (1)$$

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

$$F_x(a) = F_x(0) = P\{\omega \mid X(\omega) \leq 0\} \quad (2)$$

The number of case studies considered (the sample space Ω) is 359 (Table 1). 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:

$$F_x(0) = (359-21)/359 = 0.9415 \quad (3)$$

The corresponding chance of failure is given by:

$$(1 - 0.9415) = 0.0585 \quad (4)$$

Which is approximately one chance of failure out of ~ 17 attempts.

The probability of failure for a conceptual vehicle is the difference between the probability associated with its non-zero 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 #16 (score = 3.55) and failed case #17 (score = 4.25). There were sixteen 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 Eq. (5), where $F_x(a)$ corresponds to $F_x(0)$ --- the probability of a successful launch.

$$P\{\omega | a < X(\omega) \leq b\} = F_x(b) - F_x(a) = ((359-21) + 16)/359 - 0.9415 = 0.9861 - 0.9415 = 0.0446 \quad (5)$$

This is approximately one chance of failure out of ~ 22 attempts (a corresponding launch success of 95.5 %).

VI. 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/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 sub-system, as in Section V-F) prior to operation compared to its actual total success/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 V-C 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. While similar comments were not provided in the scoring done in Section V-F, this example contains these comments as a means to substantiate the rationale of the score assigned.

The Shuttle/Centaur upper stage was a joint NASA-USAF program in 1981-1986 to develop two new configurations of the Centaur upper stage (“G” and “G-Prime”) capable of launch from an orbiting Space Shuttle (Fig. 7a). Although the program was cancelled only months prior its first launch due to 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 sixteen times on Titan IV from 1994-2003 (Fig. 7b).

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 nine months later. The aggregate data in those PDR and CDR packages was the most concise and comprehensive technical description of the program [15, 16]. In addition, a book documenting the history of the Centaur upper stage had a comprehensive discussion of the technical problems encountered during development [17]. These three sources served as the basis for scoring using this methodology.

Table 3 is the scoring of the Centaur G-Prime. (Note: while the Centaur was managed by Lewis Research Center (LeRC), much of the Johnson Space Center (JSC) management 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 & Centaur stack (JSC and LeRC). This was due to the 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 NASA 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 which 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 the resolve technical 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, due to 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 were a continuous source of major problems, thus a score of 0.90 was assigned.

More moderate problems existed in four other areas which may or may not have been resolvable within the existing program budget and schedule. No entire stage propulsive altitude testing of Shuttle/Centaur was performed [17]. Propellant system failures and erratic behavior became apparent late in the development, 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 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 JSC staff’s repeated delays, unresponsiveness to data requests, and inappropriateness of integration rather than engineering staff involvement all 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 which were not documentable and indeed incapable of being technically substantiated. This resulted in serious problems in 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 ½ years. The impressive technical progress in such a short period of time was evident in the major review documentation [15, 16]. However, this was accomplished with considerable overtime by most of the leadership and many of the staff [17]. The zero (i.e. favorable) scores (engineering errors, component testing, and much of 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 %. The final record of the Centaur G-Prime upper stage on the Titan IV booster was 14 successes, one failure, and one “no-trial” (failure prior to Centaur phase). Thus the actual system failure rate of 6.67 % compared reasonably well with the predicted value. However, the most important result was the largely inverse qualitative scoring of almost every sub-root 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 1980’s was developed by NASA GRC & General Dynamics, while the failure in 1999 came after the transfer to USAF Space Division and Lockheed/Martin purchase of General Dynamics SSD. Effects of major changes in organizations can be a weakness of this methodology.



Figure 7a): Shuttle / Centaur G-Prime Upper Stage and Figure 7b): Titan IV Launch Vehicle

Table 3: Scoring of Shuttle/Centaur G-Prime Upper Stage Failure

	Sub-root cause	Root Cause	Total Score
	Qualitative Scores	Qualitative Scores	4.20
Insufficient testing (integrated system)		0.50	
Lack of prudent integrated system testing			No altitude propulsive stage test at 109%; PLIS mount failures; CISS prop valves erratic ops p. 206
Not pursuing "test as you fly; fly as you test"			Structural dynamic test campaign, system integration facility (for avionics, S/W, others) System Level III/IV
Insufficient understanding of interactions within entire system			Most of Centaur adopted/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
Incorrect analytical modeling or computational errors			System Level III/IV Program PDR (March 1983) and CDR (Dec 1983) reports
Ineffective Systems Engineering		0.70	
Inadequate SE / engr judgment / understanding, resolving crit problems			Repeated JSC safety-driven changes in critical fluid dump system interface between Shuttle & Centaur; "fu
Insufficient meaningful reviews			System Level III/IV Program PDR (March 1983) and CDR (Dec 1983) reports
Failure to challenge analyses, heritage, assumptions			Repeated LeRC challenging of astronauts' LH2 concern with Centaur vs. ET (TLH p. 197)
Analytic models uncorrelated w/ actuals, ill-scaled, or questionable validity			Modal survey performed on test article, trajectory design code based on past Atlas/Centaur flight data, etc
Insufficient testing (components, sub-systems)		0.00	
Lack of prudent component, sub-system testing			System Level III/IV Program PDR (March 1983) and CDR (Dec 1983) reports
Verification by analysis or comparison with requirements only			System Level III/IV Program PDR (March 1983) and CDR (Dec 1983) reports
Heritage hardware/software: not validating for new application			System Level III/IV Program PDR (March 1983) and CDR (Dec 1983) reports
Not establishing instrumentation needs			System Level III/IV Program PDR (March 1983) and CDR (Dec 1983) reports
Process errors		0.30	
Fabrication, test, integration, or launch process not followed			Observed lower quality manufacturing, transport, and contractor staff actions (TLH p. 209-210)
Non-standard events, work-arounds not incorporated into process			None identified
Hardware failure (flight or ground)		0.20	
Poor quality or statistically out of tolerance component			n/a
Multiple unforeseen program/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 p. 196-198; short sched in 1986 p. 205
Imprudently low funding			~\$2B current year funding over 4.5 years; Joint NASA & USAF funding
Poor program management		0.60	
Lack of leadership integrity			LeRC securing 109% SSME throttle baseline (TLH p. 205, 208, 209)
Inattentiveness to (or ineffectiveness in) managing problems			JSC integration staff rather than JSC engineering staff; delayed tech responses, safety issues (such as FDC
Normalization of Deviance (unexpected deviation, revised expectation)			JSC Shuttle lift capability/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
Insufficient (or no) IV&V			System Level III/IV Program PDR (March 1983) and CDR (Dec 1983) reports
Poor team communication		0.90	
Organization-to-organization differences			JSC unresponsive to LeRC technical data requests; difference in Center cultures, JSC Integration vs. Tech
Insufficient formality between working groups			Sufficient technical working groups between LeRC and GDSSD
Insufficient use of independent review team guidance		0.50	
Absence of independent assessment			No NAR convened ; continued Safety concerns by astronauts p.197-199 and 206-207
Failure to heed or fully implement recommendations			n/a
Others		0.00	
International pressures			n/a
Loss of key leader without comparable replacement			n/a
Others			n/a

VII. Potential Future Applications

The Defense Advanced Research Projects Agency's (DARPA) Experimental Spaceplane Program (XSP) is a currently in-development reusable booster. It is intended to be capable of ten suborbital flights in ten days, as well as hypersonic cruise missions up to $M = 10$. 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 this methodology, which 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 which could also benefit. NASA's current Space Launch System and Orion programs have been repeatedly delayed and costs continue to escalate [18, 19]. 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).

VIII. Caveats and Concerns

There were several concerns raised about this methodology by staff of the NASA Headquarters Safety Center and the NASA Glenn Research Center Safety, Reliability and Quality Assurance Branch. 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 non-zero scores as well. That is: no successful launch is exactly nominal, and failing to incorporate these “non-zero score successes” into the cumulative distribution function is not strictly correct. While true, the source data base 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 post flight 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 discussed in Sections IV-A and -B) focuses on “errors” --- 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 [20].”

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 which 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 due to their anticipated resource-intensive needs (people, time, 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 V-F 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 % (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 the Section VI, a potential major weakness can arise when there is a change in the organization which 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 organizations. 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 are 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, and various competing issues all being struggled with during the pressure-intensive countdown. The obvious poor decision in hindsight frequently appears to be 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 [20].

IX. 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 component hardware-focused and 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 manifested in a statistical way. While Probabilistic Risk Assessments, particularly when augmented with Human Reliability Analysis, are effective, established methods to determine causes of failure for specific sub-systems, 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 which can be applied to an entire launch vehicle system. While 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 judgement 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 sub-root causes of failure were significantly inverse 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. While a numerical score from a failure risk assessment will never be actually verified due to 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.

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'Conner, Mr. Steve Lilley, and Mr. Keith Knudson of the NASA HQ 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 DARPA/TTO Experimental Spaceplane Program staff and management, particularly XSP managers Mr. Jess Sponable and Mr. Scott Wierzbanski, and Chief Engineer Vijay Ramasubramanian.

References

- [1] Hamlin, T., et al., "Shuttle Risk Progression: Use of the Shuttle Probabilistic Risk Assessment (PRA) to Show Reliability Growth", AIAA paper 2011-7353, September 2011.
- [2] "NASA Procedural Requirements for Mishap and Close Call Reporting, Investigating, and Recordkeeping", NPR 8621.1C, 19 May 2016.
- [3] "Titan IVB-32/Milstar-3 Accident Investigation Board Report", USAF Form 711, USAF Mishap Report, date unknown.
- [4] Chandler F.T., et al., "Human Reliability Analysis Methods Selection Guidance for NASA", NASA HQ/OSMA Technical Report, July 2006.
- [5] "Probabilistic Risk Assessment Procedures Guide for NASA Managers and Practitioners", NASA/SP-2011-3421 Second Edition, December 2011.
- [6] Leveson N. G., "Engineering a Safer World", MIT Press, Cambridge, MA, 2011, Chapters 6, 7, and 8, pp. 169 – 249, and Appendix B pp. 469 – 493.
- [7] Huang, Z., et al., "Key Reliability Drivers of Liquid Propulsion Engines and a Reliability Model for Sensitivity Analysis", AIAA paper 2005-4436, July 2005.

- [8] Gernand, J. L., et al., “Constellation Ground Systems Launch Availability Analysis: Enhancing Highly Reliable Launch Systems Design”, AIAA paper 2010-2180, April 2010.
- [9] Morse, E., L., et al, “Modeling Launch Vehicle Reliability Growth as Defect Elimination”, AIAA paper 2010-8836, Sept 2010.
- [10] Guikeme, S., D., et al, “Bayesian Analysis of Launch Vehicle Reliability”, AIAA paper 2003-1175, January 2003.
- [11] “NASA Reliability and Maintainability (R&M) Standard for Spaceflight and Support Systems”, NASA Technical Standard 8729.1A, June 13, 2017.
- [12] Nieberding, J. J.; Ross, L. J., “Lessons Learned Applied to Space Systems Developments”, presentation, Aerospace Engineering Associates LLC, Bay Village, OH, Vol 1, Ver. 1.0, 2006.
- [13] Nieberding, J. J.; Ross, L. J., “Mission Success First: Lessons Learned”, Aerospace Engineering Associates LLC, Bay Village, OH, Class #100 presentation, 9-10 November 2016.
- [14] ASTRO-E-2 Mishap Report (Appendix), Table 7-1, NASA Safety Center, Mishap Investigation Board, Type A Mishap, IRIS No. 2005-273-00003, 2005.
- [15] Shuttle/Centaur Level III/IV Program PDR at LeRC, General Dynamics Convair Division, San Diego, CA, March 1983.
- [16] Shuttle/Centaur Level III/IV Critical Design Review at LeRC, General Dynamics Convair Division, San Diego, CA, December 1983.
- [17] Dawson, V., Bowles, M., “Taming Liquid Hydrogen: the Centaur: Upper Stage Rocket, 1958-2002”, NASA History Series, NASA SP-2004-4230, Washington, DC, 2004, Chapter 7, pp. 189 – 219.
- [18] “NASA’s Plans for Human Exploration Beyond Low Earth Orbit”, NASA Office of Inspector General, Report No. IG-17-01713, Washington, DC, April 2017.
- [19] “NASA Human Space Exploration”, U.S. Government Accountability Office, GAO-17-414, Washington, DC, April 2017.
- [20] Lilley, S., NASA email, NASA HQ Safety Center, Cleveland, OH, personal communication 21 December 2017.