

A Measurement Framework for Air Traffic Safety Metrics

Jacques Press
NAS System Engineering Group

December 2003

DOT/FAA/CT-TN04/10

Document is available to the public
through the National Technical Information
Service, Springfield, Virginia 22161



U.S. Department of Transportation
Federal Aviation Administration

William J. Hughes Technical Center
Atlantic City International Airport, NJ 08405

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof. The United States Government does not endorse products or manufacturers. Trade or manufacturer's names appear herein solely because they are considered essential to the objective of this report. This document does not constitute FAA certification policy.

This report is available at the Federal Aviation Administration, William J. Hughes Technical Center's full text, technical reports web site: <http://actlibrary.tc.faa.gov> in Adobe Acrobat portable document format (PDF).

Technical Report Documentation Page

1. Report No. DOT/FAA/CT-TN04/10		2. Government Accession No.		3. Recipient's Catalog No.	
4. Title and Subtitle A Measurement Framework for Air Traffic Safety Metrics				5. Report Date December 2003	
				6. Performing Organization Code ACB-210	
7. Author(s) Jacques Press				8. Performing Organization Report No. DOT/FAA/CT-TN04/10	
9. Performing Organization Name and Address Federal Aviation Administration NAS System Engineering Group William J. Hughes Technical Center Atlantic City International Airport, NJ 08405				10. Work Unit No. (TRAIS)	
				11. Contract or Grant No.	
12. Sponsoring Agency Name and Address Federal Aviation Administration System Engineering Division William J. Hughes Technical Center Atlantic City International Airport, NJ 08405				13. Type of Report and Period Covered Technical Note	
				14. Sponsoring Agency Code ACB-200	
15. Supplementary Notes					
16. Abstract This technical note identifies cautionary steps required when measuring aviation safety, air traffic safety in particular. Measuring safety in objective terms (i.e., metrics) is widespread today. The aviation community practices it not only to promote a universal understanding of safety but also to act upon what the metrics seem to indicate. However, safety is a naturally abstract concept, making its measurement difficult to characterize and interpret. To minimize this difficulty, safety metrics should be bounded by some sort of rational framework. Without it, metrics can be derived haphazardly, causing their dismissal downstream for lack of meaning. This note introduces a measurement framework as three fundamental principles. Collectively, they supply common sense guidance during metric selection. Within this spirit, the framework stresses measurement prerequisites like definition, context, scope, and intent. The document illustrates the framework with examples, several of which derive from real but de-identified aviation safety analysis results. The note also provides three systematic methods to derive appropriate metrics.					
17. Key Words safety, measurement, metrics				18. Distribution Statement This document is available to the public through the National Technical Information Service, Springfield, Virginia, 22161	
19. Security Classif. (of this report) Unclassified		20. Security Classif. (of this page) Unclassified		21. No. of Pages 31	22. Price



Table of Contents

	Page
Executive Summary	v
1. Introduction.....	1
1.1 Safety in Air Traffic Control	1
1.2 The Movement Towards Metrics.....	1
1.3 Contrast.....	2
1.4 Caution.....	3
1.5 From Measurement Ideology to Framework	3
2. Measurement Ideology.....	4
2.1 Historical Perspective	4
2.2 Current Status	4
2.3 A Modern Obstacle	5
2.4 Measurement Standards and Numerical Assignments.....	5
2.5 Important Safety Measurement Questions.....	6
3. A Proposed Measurement Framework.....	8
3.1 What Is a Framework.....	8
3.2 Sample Frameworks	8
3.3 Why a Measurement Framework.....	8
3.4 A Measurement Framework	8
4. Definitions and Classification.....	10
4.1 Fundamental Definitions.....	10
4.2 Products, Processes, and Resources.....	10
4.3 Measurement Artifacts.....	11
5. The Representational Theory of Measurement.....	15
5.1 Measurement Conditions	15
5.1.1 Representation Condition.....	15
5.1.2 Uniqueness Condition.....	15
5.1.3 Meaningfulness Condition.....	16
5.2 Caution.....	17
5.3 A Compromise.....	17
6. Measure Selection.....	18
6.1 Systematic Measure Selection	19
6.2 Measure Qualities	20
7. Concluding Remarks.....	23
7.1 How Safety Becomes Synonymous With Its Measurement	23
7.2 From Many Measures to a Few Standards.....	23
References.....	24

List of Illustrations

Figures	Page
Figure 1. Two conceptual views of ATC flight safety.....	2
Figure 2. Measurement artifact pyramid.....	11
Figure 3. Metric example: Nationwide airspace violation counts.	12
Figure 4. Runway incursions at major airports.	13
Figure 5. Pilot duty times frequency distribution.	19

Tables	Page
Table 1. A Measurement Framework	9
Table 2. Measurement Scale Typology According to Stevens (1959)	16

Executive Summary

Today's aviation community strives to capture the concept of safety in quantitative terms as a way to promote a universal understanding of it. Within this spirit, it is no surprise to find safety measures (commonly called metrics) frequently used in air safety analysis. Yet, safety is a naturally abstract concept, making its measurement difficult to characterize and interpret. To minimize this difficulty, safety metrics should be bounded by some sort of rational framework. That is, without certain principles frameworking (limiting) the measurement process, metrics can be derived haphazardly, causing their dismissal downstream for lack of meaning.

There are many foundational reasons to be cautious with metrics. For instance, metrics may turn out to be meaningless when they involve transformations (like the mean, median, mode, etc.) derived without concern for the measurement scale typology. Within representational measurement theory, scale typology is identified as nominal, ordinal, interval, or ratio, and this classification limits our metric applications to only certain privileged transformations.

Starting with a layperson description of measurement's foundations, this document introduces the framework in terms of several principles. They supply common sense guidance during metric selection. Towards this goal, the framework approach adopted here stresses measurement prerequisites like definition, context, scope, and intent. Also, the document provides three methods as systematic ways to derive appropriate metrics.

Although presented in tutorial format, the forthcoming discussion does not remain abstract. It illustrates the framework with examples, several of which derive from real but de-identified safety analysis results. The discussion adopts a universal theme expecting to attract an audience wider than that of air traffic control. The framework is generic enough to extend to other metrics, like those envisioned for airport safety, aircraft fleet maintenance and reliability, air traffic capacity, and so on.



1. INTRODUCTION

The aviation community relies increasingly on measures, labeled metrics¹, to monitor its vital signs. Flight safety is among the most sensitive signs because it concerns risk to human life. Measuring safety requires a most careful treatment; its metrics must be exceptionally clear and precise before we make safety claims based on them. Frameworking¹ metrics within rational principles ensures this analytical rigor.

1.1 Safety in Air Traffic Control

Relative to other domains, Air Traffic Control (ATC) enjoys a superior safety record. So superior, it engenders the belief we have finally solved the air traffic problem: safely separating flights. Indeed, rare are the mishaps involving ATC flaws. However, we must also acknowledge we have been infusing much technology into ATC, and, with technology comes a host of complexity, interfacing, and self-autonomy unknowns. In turn, these unknowns can bring surprise; they might paradoxically increase rather than decrease safety risk.

Example: A once-proposed ATC concept called Conflict Resolution Advisories (CRA) contains sophisticated logic supplying real-time advisories to controllers who must resolve flight conflicts (Wasser, Hauser, & Press, 1989). With a design promising to reduce operational errors, CRA software represents autonomy at a quantum level beyond the simpler computer information that controllers see today. So modern is the concept that they might hesitate accepting this artificial advice unconditionally. One reason is they would want assurance that CRA accounts for all possible conflict situations without creating secondary new ones in the process. Thus, how far CRA's autonomy should safely match the human mental process remains in question. As a result, CRA implementation awaits more definitive times. Technology unknowns like those in CRA explain why safety-sensitive agencies proceed cautiously with ATC modernization. Cases like CRA also explain why safety analysis must continue as an important hazard preventive, even though ATC safety prevails in today's skies.

1.2 The Movement Towards Metrics

ATC safety analysis made its formal debut about mid-20th Century when Collision Risk Modeling (CRM) formulations appeared in publication. For a brief history see Machol (1995). In general, CRM invokes principles from geometry, dynamics, and probability to express flight separation in equation form. These physical equations contribute to the understanding of safety. However, being complex, they must undergo much validation and experimentation before we can use them directly, say towards improving ATC's rules and regulations.

Physical equations aside, decision-makers are steering recently towards more direct methodologies to measure safety. Metric-Based Analysis (MBA) is one of them. MBA relies on collected information to derive measures (metrics) of safety. This information finds its roots in data sources maintained throughout aviation. Sources include records of defects, errors,

¹ Metrics and frameworking are defined formally later in the document.

incursions, violations, near-misses, incidents, even accidents. Normally less mathematical than CRM, MBA serves, nonetheless, an important purpose. Metrics show they are useful when we seek to summarize the notion of safety into key numerical indicators. Once properly aggregated, these numbers help us track objectively how well we are meeting safety goals in a most global sense.

1.3 Contrast

CRM focuses on separation dynamics between a few flights or ground obstacles at a time; however, MBA views traffic as a continuous flow-stream subject to safety deviations, as contrasted conceptually in Figure 1. MBA isolates not only deviation counts but also their types and rates as key variables summarizing traffic's "health."

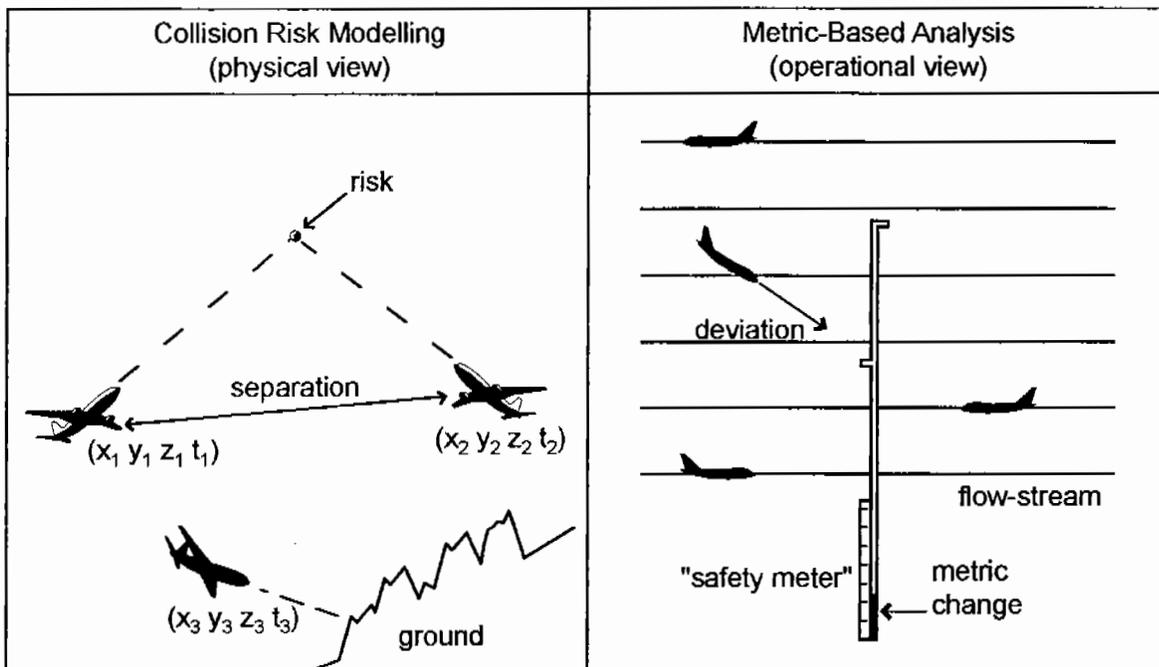


Figure 1. Two conceptual views of ATC flight safety.

Analogy: Two views describing fluid dynamics show analogy to the CRM-MBA views. One view, due to the mathematician Lagrange (1736-1813), leads to motion equations by assuming individual fluid particles are the main players. Analogous to CRM, the Lagrangean approach characterizes particle dynamics by tracing their pathlines. The other view, due to the mathematician Euler (1707-1783), matches closer MBA's mode. It assumes fluid motion as a field with streamlines instead of particle pathlines. Eulerian measurement would probe the

flow-stream (similar to an aircraft pitot tube) to sense pressure changes, much like MBA would sample ATC deviations nationwide to sense any changing trends as symbolized by the “safety meter” in Figure 1.

1.4 Caution

After several decades past inception, CRM’s foundations should be well established on the analytical landscape. Unlike CRM, which depends substantially on mathematical derivations, MBA gains its validity mostly from the veracity of available information. However, MBA’s direct approach belies any simplicity to it. Metrics must be substantially representational, unique, and meaningful of the measured target, according to a modern measurement theory. Otherwise, metric-based statements like

- runway incursions reduced by one third,
- operational errors increasing 50%, and
- safety testing 90% complete

prove to be ambiguous or worse, misleading. Numbers can mislead unless preceded by careful derivation. Underscoring this caution, several authors warn us of analytical surprises where numerical irrationality confounds even the best intentions (see Barnett, 1994; Dewdney, 1996; Huff, 1993; Paulos, 1990).

1.5 From Measurement Ideology to Framework

This document offers an overview of measurement ideology leading to important practical questions. To address them, the document proposes a measurement framework for ATC safety. The framework is generic enough to extend to other metrics, like those envisioned for airport safety, aircraft fleet maintenance and reliability, air traffic capacity, and so on.

2. MEASUREMENT IDEOLOGY

2.1 Historical Perspective

Scholars would most likely place measurement among key practices advancing human knowledge. Progressive societies not only promoted this practice, but also philosophized about its meaning (see Encyclopedia Britannica; Finkelstein, 1982; Klein, 1974; Narens & Luce, 1986). History finds early thinkers questioning how to size physical objects meaningfully, according to logical formalism (see Savage & Ehrlich, 1992). Later, their successors deliberated why certain quantities (length, area, volume) were additive for measurement purposes, whereas human qualities (like virtue) were not readily so (see Michell, 1990). Came late 19th Century, modern measurement theory was initiated by Hermann von Helmholtz (1821-1894). It was later complemented by others, like Otto Hölder (1859-1937). Early 20th century saw axioms, postulates, and hypotheses pluralizing measurement thought into several competing doctrines or theories (Churchman, 1959; Michell, 1986; Narens & Luce).

2.2 Current Status

Doctrines range from the restrictive classical theory embracing only physical measurements (speed, altitude) to the more flexible representational theory where abstractions (intelligence, quality) are also assumed measurable. Being flexible, the representational theory might succeed formalizing the measure of safety.

Discourse in the literature portrays measurement playing an important role in many research endeavors, be it in traditional fields like physics or newer ones like software engineering. Yet, because researchers focus on their own particular field, they accept measurement as a tool rather than a discipline central to their objectives. As a result, they do not dwell much on measurement theory or assumptions. Some sources however treat measurement more universally, as a stand-alone discipline complete with its own generic formalism (Berka, 1983; Ellis, 1966; Krantz, Luce, Suppes, & Tversky, 1971, 1973; Kyburg, 1984; Narens & Luce, 1986; Pfanzagl, 1968; Savage & Ehrlich 1992). Within such formalism, these scholars continue to deliberate ideologies championed by various schools of measurement thought.

This deliberation would lead one to believe a universal measurement understanding established here by now, based on a single theory supporting it. This theory would be framed permanently in the annals of human knowledge, much like the theories frameworking so many other disciplines. Unfortunately, philosophical knots still fragment this wishful rationalism, limiting the prospects for a universal foundation. Elegant mathematical formalisms have been proposed to bring measurement into a same theoretical context but discourse about them continues among scholars in various disciplines, particularly in psychology where abstract notions abound (Michell, 1990). Sources illustrating such discourse include Carnap (1966), Churchman (1959), Michell (1986), Narens and Luce (1986), and Schwager (1991).

Measurement doctrines anchor themselves around differing philosophical beliefs; therefore, they persist being inherently developmental. Yet, society, driven by practical needs, has been sizing and tracking successfully all sorts of concepts. It continues to do so today without waiting for definitive philosophical closure.

2.3 A Modern Obstacle

An obstacle restricts the act of measurement itself; what exactly is being measured proves open-ended, particularly in the case of modern abstractions like safety. Measuring means ascertaining or estimating (usually quantitatively) the state of something. By no means restrictive, this definition should apply to concrete concepts like aircraft speed, runway length, or flight time, but also extends desirably to notions like accident risk, aircraft reliability, and air traffic safety. Ideal measurement seeks to determine the ratio of an object's quantity to an official constant. An aircraft flight time of 5 hours simply means a quantity observed 5 times an official constant, the standard clock hour. This comparison helps rank objects of interest, according to some global scale. Logical as it sounds, measurement is not so straight in practice. It stays subject to differing opinions when we try to measure broadly defined concepts like controller performance or aviation safety. By human nature, the broader the concept, the more controversial its measurement becomes.

Example: We can easily deduce that Aircraft A is 5 times faster than B because we can clock both according to a standard time definition, but we would have difficulty concluding just as firmly that A is 5 times safer than B. The argument vanishes only if "safe" is defined in an absolute way, with the same specific attributes each time. Yet, to be truly absolute, safe must carry its own standard for all parties to accept in the same, universal sense, as explained next.

2.4 Measurement Standards and Numerical Assignments

Physical concepts like length and speed are concrete enough to be readily understood through observation (even though their epistemology may have been tossed around by philosophers). We can easily observe one runway being longer than another. Therefore, we would naturally expect that any numerical assignments we make to length would also be different from runway to runway, based on some sort of measurement standard, say the yardstick. After centuries of experimentation and debate, society now accepts readily numerous measurement standards like the yard, degree centigrade, and mile per hour. These are cases where cultures first accepted intuitively what a yard, mile, or hour "should be like" (say, relative to the earth's geometry or astronomical clockwork motion). Only afterwards did they establish more formal standards for them.

Less can be said about the mental picture scientists have concerning modern extra physical concepts, like quality, reliability, or safety. Imagine how long our checklist must be before engineers could declare an aircraft design 100% safe to fly, truly an impossible number to reach. Or, think how complex the definition of reliability must become before we can claim a universal standard on how to view it, let alone measure it. We just do not hold enough of a collective understanding of reliability, even though we claim to recognize it on an individual basis. Yet society often seeks to measure interesting notions like reliability and safety. These concepts are abstract, lacking a standard about which we can all "feel" the same way. However, this argument does not claim dogmatically they should stay non-measurable. Indeed, reliability and safety are measurable in practice, even if the results prove approximate. This document claims, however, that a collective understanding and definition (through some agreeable framework) must precede any meaningful measurement of these abstract concepts.

Example: Measuring meaningfully the safety of ATC software presents astounding difficulties. For instance, traditional reliability analysis methods may not apply to software, let alone software safety in ATC systems. Reliability focuses on mishaps precipitated by component failures. This premise works when individual components are hardware showing well-defined functional boundaries among themselves. Less can be said about software, particularly where system interfaces (human, software, hardware, environmental), rather than components, are seen as frequent sources of failure, according to Leveson (2000). Moreover, reliability measurement works well for repetitive cases where hardware design and manufacturing remain constrained by physical laws. Unlike hardware, software is abstract, allowing for creative, revolutionary applications with fewer design limits or identical precedents. Therefore, methods more specialized than those suggested by classical reliability theory must be advanced to measure this new dimension of safety. To meet this need, the classical hardware frame of mind should be replaced with an entirely new paradigm before safety measurement can emerge meaningful in the case of software.

2.5 Important Safety Measurement Questions

The preceding argument leads to questioning the serious task of assessing safety. To size it, managers might propose intuitive attributes like number and type of precautions taken to preserve it, such as the number of flaws discovered during inspection, cumulative days without controller errors, days between incidents, accident severity, fatalities, and number of discovered software defects, all reasonable choices on the surface. However, applying a mental notion (measurement) to symbolize an empirical phenomenon like safety fails to be so direct. Measurement requires careful thinking before selecting any measure. In the process, we are obligated to answer introspective questions like:

1. How much do we need to understand collectively about the notion of safety prior to measuring it in a certain way (say, representing it according to the same universal attributes)?

Example: Two nations rate runway incursions by severity levels, each using a separate classification criterion. One nation uses four levels, the other five, and they are all defined differently. Because of the difference, combining or comparing the two nations' runway safety progress in terms of a same severity metric would be misleading if not accounting for the difference in definitions and levels.

2. Do the chosen metrics represent uniquely our concept of safety?
3. Can we claim we have properly measured (captured) safety? For instance, can our chosen safety attributes be used assuredly as precursors to predict accidents?
4. What kind of meaningful statements can we make about safety once we measure it? That is, what kind of numerical operations (addition, multiplication, division, etc.) are we allowed to make on the attributes we choose to represent safety? Going further, can we summarize safety using statistics like proportions, averages, medians, or standard

deviations, all of which involve addition, multiplication or division? For instance, can we introduce an “average safety” for a nation’s airspace? How precise must this average be?

These questions carry no simple answers. Their difficulty justifies why we seek a measurement framework. However, frameworking does not pretend to answer them squarely. Instead, it offers a disciplined way to handle their underlying presence in practice.

3. A PROPOSED MEASUREMENT FRAMEWORK

3.1 What Is a Framework

A framework is a rational process governed by principles and rules formalizing how to view a concept of interest. For example, principles can be conjectured to integrate safety measurement (the concept of interest) into a mental frame. By imposing coherent rules and definitions, framing consolidates the concept, making it easier for all to understand and accept. Framing stems from many sources: experimentation, theory, insight, lessons learned, and so on.

Frameworks, despite their formality, are not physical objects. They are only hypotheses of our empirical or theoretical beliefs. As a result, frameworks become subject to revision when new findings or practical lessons appear along the way.

Example: Newton established concrete principles (laws) frameworking the physical world. But this framework, the starting point for much of physics, was eventually displaced by a more accurate framework based on relativity theory soon after measurement confirmed the case.

In non-physical cases (safety, for example), where established theory is weak or absent altogether, we expect these conjectures to be even more transient than in physical science. However, we must start tentatively somewhere, otherwise in no way can we measure these abstractions.

3.2 Sample Frameworks

Examples of frameworks applied to abstract notions include a comprehensive one by Fenton et al. (1997). With it, he proposes formalizing software measurement and selecting the right software metrics. Also, see Zachman (1987) and Evernden (1996), both of whom describe how to framework information system architectures. Frameworks are also adopted to clarify abstractions like human behavior and performance, notably in psychology, sociology and education assessment.

3.3 Why a Measurement Framework

Frameworking the measurement process serves as a baseline, a starting point leading to the right measures. Without this frame of reference, we cannot be sure we are later successful during metrication (the process of selecting and applying measures, metrics, etc.). That is, without reference to measurement principles, it is not really possible to know if we are measuring safety in a rational, defensible way.

3.4 A Measurement Framework

This document proposes a framework consisting of two foundational principles (rules). See Column 1 in Table 1. They supply the analytical basis for rational measurement. The framework also furnishes an operational principle (Table 1, Column 2), which details metrication rules that complement those in Column 1.

Table 1. A Measurement Framework

Foundational Rules	Operational Rule
<p><u>Rule 1</u></p> <p>Measurement practice should aim for clear definitions and classification. As prerequisites to measuring ATC safety, defining and ordering the basic constructs of measurement ahead of time lessens confusion as to what is being measured downstream. (See Section 4.)</p>	<p><u>Rule 3</u></p> <p>Safety measures must be selected according to a systematic process, within the rigorous spirit of Rules 1 and 2. Enforcing this rigor, we should size candidate measures using criteria addressing context, purpose, quality, scope, and maturity, for example. (See Section 6.)</p>
<p><u>Rule 2</u></p> <p>Measurement practice should depend on an underlying measurement theory (like representational theory) to formalize it. For example, if representational theory is adopted, then any resultant metric proves formal if it is sufficiently representational, unique and meaningful of the object of interest. (See Section 5.)</p>	

4. DEFINITIONS AND CLASSIFICATION

Clear and consistent terminology not only reduces subjectivity but also encourages universal acceptance, both important framework objectives. Accurate definition and classification, the hallmarks of most sciences, work together towards these objectives.

4.1 Fundamental Definitions

During measurement, numbers or symbols are assigned to attributes of entities to describe them in a useful way.

- An *entity* means an object, person, event, even an entire system targeted for measurement; examples include flight plans, pilot deviations, controllers, displays, and documents.
- An *attribute* means a property, feature, or characteristic that we seek to measure for each entity; examples include runway incursion severity, incident frequency, pilot deviation count, software error count, software quality, display reliability, and so on.

Entities and attributes are generic constructs of any measurement, be it in air traffic safety or otherwise. Note, however, that we measure the attributes of entities rather than the entities themselves.

Example: We should “measure the reliability of the ATC display” (say, in terms of several attributes), rather than “measure the display” itself. The second choice of words implies vague intent. This simple example shows how the proper terminology reduces subjectivity. Terminology forces the measurer to have clear purpose, measure reliability, not any other aspect. In more complex cases, we might find ourselves wanting to measure an Air Route Traffic Control Center (ARTCC), tower, or some other similarly large entity for safety. However, without defining clearly the proper representational attributes upfront, we cannot be sure what is being measured later on.

4.2 Products, Processes, and Resources

Fenton and Pfleeger (1997) recommend that entities should be classified as products, processes, or resources, also in precedence to measurement. This premise makes sense because these three terms form the essential classes of our work experience, with or without measurement.

Example: Defining and measuring safety attributes of a radar antenna (a product) should obviously differ than those of ATC personnel (a resource). One reason: the decision path to deploy a more reliable radar antenna would likely differ from the path to hire more controllers. In the antenna case, “time between failures” might represent an important measure. In the second case, we might select training and experience rather than the absurd metric “controller time between collisions”. This example illustrates how a measure appearing to work for a product might fail for a resource.

Although the preceding example seems trivial, there are subtle cases where we seek to measure things without identifying their entity class first. Without classification, the measures we choose might turn out to be surprisingly incompatible with intended objectives.

4.3 Measurement Artifacts

Beyond entities, the proposed framework offers to distinguish between the basic measurement tools (artifacts), according to their intended operational use. For this purpose, the framework singles out four artifacts: (a) measures, (b) metrics, (c) indicators or (d) models, all conceived in some quantitative, qualitative, or even graphical form. Before we select any of them, the framework requires they be defined and ranked, specifying how each serves a separate purpose. Towards this end, the following conceptual pyramid (Figure 2) lists them in ascending order of sophistication.

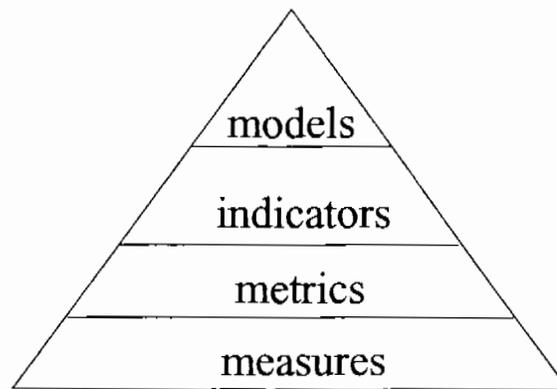


Figure 2. Measurement artifact pyramid.

a. Measures:

At the bottom, measures represent the simplest artifact. A measure states results obtained from one or more soundings but leaves the audience to induce any trend beyond the numbers.

Example: Measure M1 states 12, 15, and 20 runway incursions at Chicago O’Hare on August 2nd, 7th, and 8th, with no measurement frequency or trend stated explicitly beyond the counts.

b. Metrics:

They belong further up the pyramid. Unlike a measure, a metric stipulates a set of repeated, periodic measures to track patterns, trends, or correlations.

Example: Extending the M1 example, the metric M2 consists of a chart showing an increasing number of incursions, going from 12 to 20, during the last three periodic readings.

Metric, deriving from *metricus* (Latin for measurement), means, in our context, something tracked with a meter. Meter implies a tool that measures or

registers, usually periodically, something of interest. Guided by this etymology, we adopt the term metric instead of measure when the latter becomes subject to a formal underlying metering process, a rigorous way where we do not select measures casually nor use them occasionally. Instead, they are picked carefully and then metered (tracked) repeatedly. This way, they might point to patterns, correlations, and similar trends useful in safety assessments, predictions, and decisions.

Example: Figure 3 shows a metric tracking nationwide airspace violations counts. These violations, derived from 12,513 pilot deviations recorded periodically by a national airspace information monitoring system, confirm a substantial increase in faulted flights since the 9/11/01 catastrophic day in the United States. Possible reasons for this change include

1. increase in the number of newly defined restricted airspaces,
2. widen existing restricted airspaces,
3. more vigilance on the part of controllers monitoring their airspace,
4. higher, security-driven motivation to report violations, and
5. time lag between airspace restriction publication and pilot awareness.

The sudden peak in Figure 3 occurred during the months immediately following September 2001. The count goes down afterwards. However, on the average, it stays noticeably higher than before 9/11/01.

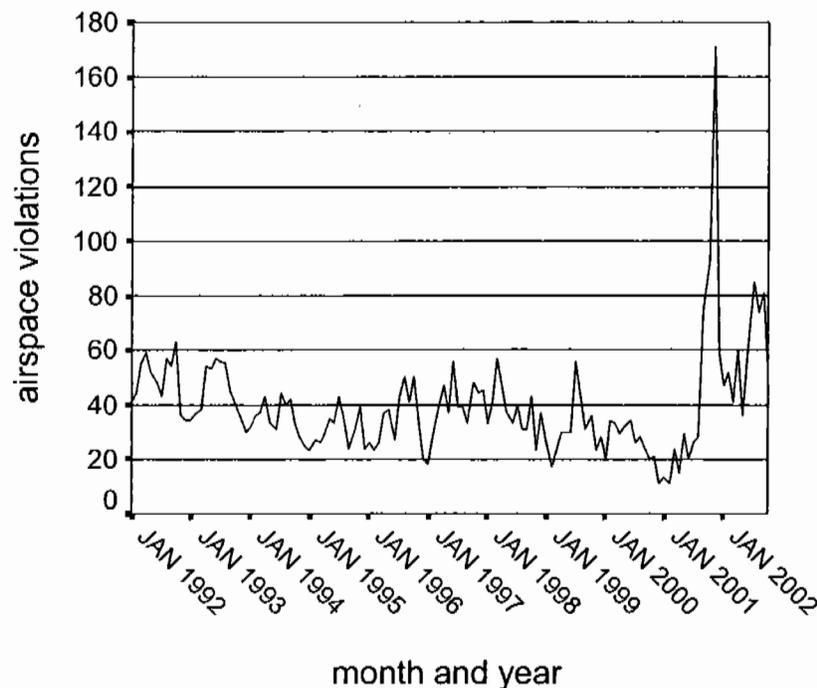


Figure 3. Metric example: Nationwide airspace violation counts.

c. Indicators:

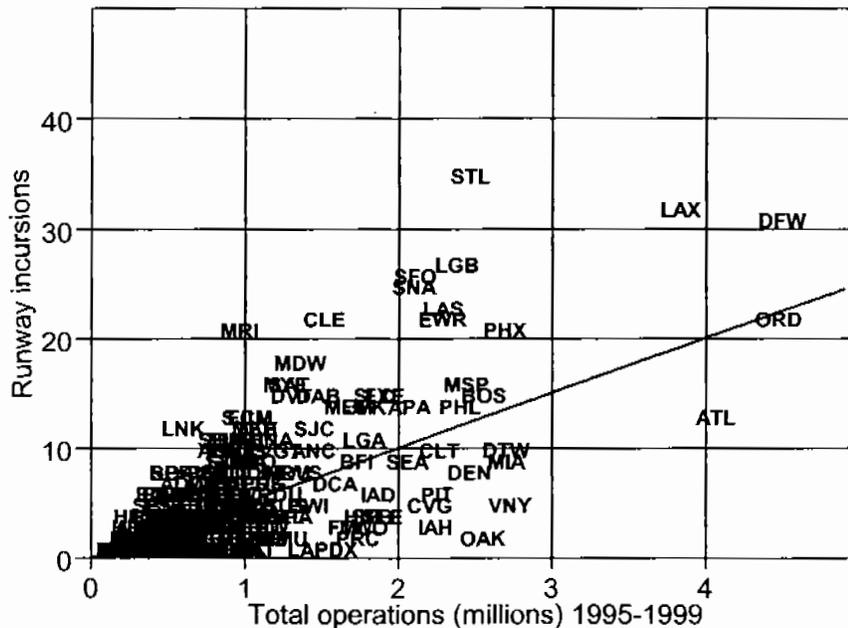
Indicators claim even more sophistication. They represent a metric with an upper and/or lower limit(s) assigned to their trend. The limits warn us to take action on those data points falling outside the limits, as would be shown on a typical quality control chart.

Example: Indicator M3 shows defects to be 12, 15, and 20, increasing beyond the indicated maximum limit of 10.

d. Models:

Highest on the hierarchy, models symbolize a process, product, or resource. Models connect two or more attributes of interest in some rational fashion. Thus, according to this hierarchy, models represent enriched composites of two or more measures, metrics or indicators.

Example: There are several ways to determine how runway incursions are distributed nationwide. We could start with simple measures: austere lists of airport names and their corresponding counts, even pie or bar charts. Incursions can also be depicted with more depth, using a multi-attribute statistical model, as follows. The data scatter and the related regression line in Figure 4 shows how various airports (labeled with 3-letter acronyms) rank among each other, not only in incursions (see vertical axis) but also in total operations (see horizontal axis) within a single diagram. The regression line, based on weighted least squares, serves as a rough model where the slope approximates five incursions per million operations nationwide.



Most models come in two broad categories: mathematical or simulation. In the mathematical case, equations are derived into a specific form based on some theoretical or empirical (observational, statistical) knowledge, as done in the preceding example. For a general account of mathematical models, see Saaty (1984). Simulation models are more concerned with the realistic flow and connection among various entities and attributes. Analysts resort to simulation when mathematical models prove difficult to represent realistically the object or situation of interest. For more depth on simulation modeling, see Law (1994).

5. THE REPRESENTATIONAL THEORY OF MEASUREMENT

Representational measurement establishes a correspondence between

a set of manifestations and the relation between them in the empirical world (say, 235 airspace violations at Center X, versus 195 at Y),

and

a set of numbers and the relation between them in the abstract mathematical world (say, the integers $235 > 195$, where the symbol “>” means “greater than”).

5.1 Measurement Conditions

A typical measurement theory imposes specific logical conditions to frame measurement practice concisely. Accordingly, measurement scholars (Finkelstein, 1982; Krantz et al., 1971, 1973; Narens & Luce, 1986) provided the following three conditions when they defined representationalism as a mapping from the two previous examples:

- Representation
- Uniqueness
- Meaningfulness

They are addressed individually in the next three sections.

5.1.1 Representation Condition

This condition addresses whether or not the conjectured measures based on measurement theory reflect faithfully the concept considered for measurement.

Example: Suppose quantitative measure, M , is selected to represent the “amount” of safety, S . Then, if M is proposed as a representational measure of S , say, $M(S)$, we declare item X having more S than Y if, and only if, the measure of X , labeled $M(X)$, is seen greater than $M(Y)$. This statement can be expressed compactly as

$$M(X) > M(Y) \text{ implies and is implied by } S(X) > S(Y)$$

If this condition is enforced for all items beyond X and Y , then the concept of S in the empirical (observational) sense is truly represented by M in the mathematical (quantitative) sense.

5.1.2 Uniqueness Condition

Uniqueness concerns how close a measure comes to being alone in symbolizing an attribute. That is, if M is truly unique, then there should be no other measure that gives the same results. Otherwise M is not unique in measuring S .

Example: Non-uniqueness implies several measures ($M, M', M'',$ etc.) of S can produce the same measurement results, a state of affairs difficult to reconcile if we have to interpret S in terms of more than one measure. Without this condition, we cannot tell which is actually relevant.

5.1.3 Meaningfulness Condition

The meaningfulness condition helps determine whether statements we make about measurement observations are true, regardless of the scale we use to symbolize attribute measurement. Stine (1989) addresses this condition comprehensively.

Example: Say M is a measure of length, in feet, of taxiway X being longer or shorter than Y. Suppose we change (transform) this scale into the decimal system (say, meters) by a constant multiplier. Then we can still claim that X is longer or shorter than Y by the same equivalent amount. In this case, we say that the results based on M hold true, regardless of scale change. Therefore, they prevail interpretable and useful (meaningful) under any scale.

Specifying the meaningfulness condition would be incomplete without relating it to a scale system. Stevens (1959) proposed four basic types of scales within the realm of representational measurement. They are ranked in Table 2, from most (top) to least restrictive in use. Also shown are the corresponding arithmetic and statistical transformations allowed for each.

Table 2. Measurement Scale Typology According to Stevens (1959)

Scale	Allowable Arithmetic Operation	Examples	Permissible Measures of Location	Permissible Measures of Dispersion
Nominal	only equalities allowed ($=$), tallying units into class A or B.	numbering runways (22R, 18L); labeling airports LAX, ORD, ATL, etc.	mode	
Ordinal	only ranking allowed, greater than, equal, or less than ($<, =, >$)	runway incursion severity (A,B,C,D)	median	percentiles
Interval	differences allowed ($+, -$)	temperature (Celsius), calendar days	arithmetic mean	standard deviation variance
Ratio	all arithmetic operations allowed: $+, -, \times, /$ proportions, percentages, rates.	rates, length, density, position, elapsed clock time, temperature (Kelvin), brightness	geometric mean harmonic mean	percent variation (coefficient of variation)

Example: Suppose we want to compute an average severity for runway incursions at a particular airport by examining the counts accumulated in A, B, C, and D severity levels. For the moment, A-D levels are distinguishable but only subjectively. The A-D classification is ordinally scaled from most to least severe, therefore, only the median is computationally permissible (see Table 2). The arithmetic average stays undefined unless differences between A, B, C and D are known numerically.

5.2 Caution

Recently, the Stevens (1959) typology is seen careening into controversy. Arithmetic restrictions imposed by the typology are rousing several scholars wishing to have more mathematical freedom in choosing measures. Writers like Chrisman (1998), Michell, (1986, 1990), and Velleman and Wilkinson (1993) cast doubt on the absoluteness of the scale categories in Table 2. Their discontent illustrates why measurement theory continues to lack full closure.

5.3 A Compromise

Satisfying all the theory's conditions defies practicality. If taken dogmatically, they would prevent finding any acceptable measure. Yet, their radical dismissal would limit how far we define formally our measurement process. Consequently, a reasonable compromise between these two extremes should help adapt this theory to practice. The previous conditions represent an ultimate standard for measuring safety. Therefore, practical measurement should attempt coming reasonably close to this theoretical standard instead of meeting it exactly. For further background and assessment of this theory, see Schwager (1991).

6. MEASURE SELECTION

Choosing the wrong measures might fray our measurement integrity unexpectedly. To safeguard this integrity, the framework's third principle asks us to select measures carefully so they match effectively their target.

Example: Among important aviation issues, rule makers are seeking to shorten air carrier flight crew duty time to safer limits. They believe lengthy duties increase accident risk due to fatigue. To better understand the fatigue-accident link, analysis might naturally focus on pilot duty and flight times spent prior to accidents or incidents. Searching data sources for a relevant duty time measure (metric) points to a reliable database identifying pilot deviations from flight rules as a collection of air traffic incidents. In that source, duty time is found listed within each deviation record, based on a pilot's subjective interview. Excluded from the search are deviations with non-pilot causes (like equipment failure) to implicate the fatigue factor even further. Figure 5 summarizes these reduced data by showing how 1,529 air carrier deviations are distributed by duty times in the 24 hours prior to incident. The distribution has a mean of 7.62 hours and a standard deviation of 3.77 hours. From these statistics, we estimate the average air carrier deviation occurs after about 6 to 8 duty hours. However, the data carry enough drawback to generate a weak metric based on the following argument. Duty times may not really reflect accident risk due to fatigue because, in 24 hours, there is enough slack (sliding) time for a full 8-hour rest period. A better measure would have been "continuous duty time immediately prior to incident." Without "continuous" and "immediately," pilot recuperation time is unknown for each incident. Also not known is whether the answering pilots understood that the duty time question meant to include those two key words. Because of these unknowns, only the tail end of the distribution can be useful, say past 16 hours (24 hours minus 8 for rest). Unfortunately, there are much less data beyond 16 hours. Although the duty times in Figure 5 might be interesting as a metric, they do not match the targeted fatigue-accident factor sufficiently well. This example illustrates why careful measure selection must precede any ensuing analysis. Without precise knowledge of the data source, measure-target mismatch might lead inadvertently to the wrong conclusions. To prevent mismatches, the proposed framework suggests guiding rules, which help identify the right measures according to a systematic process and several criteria.

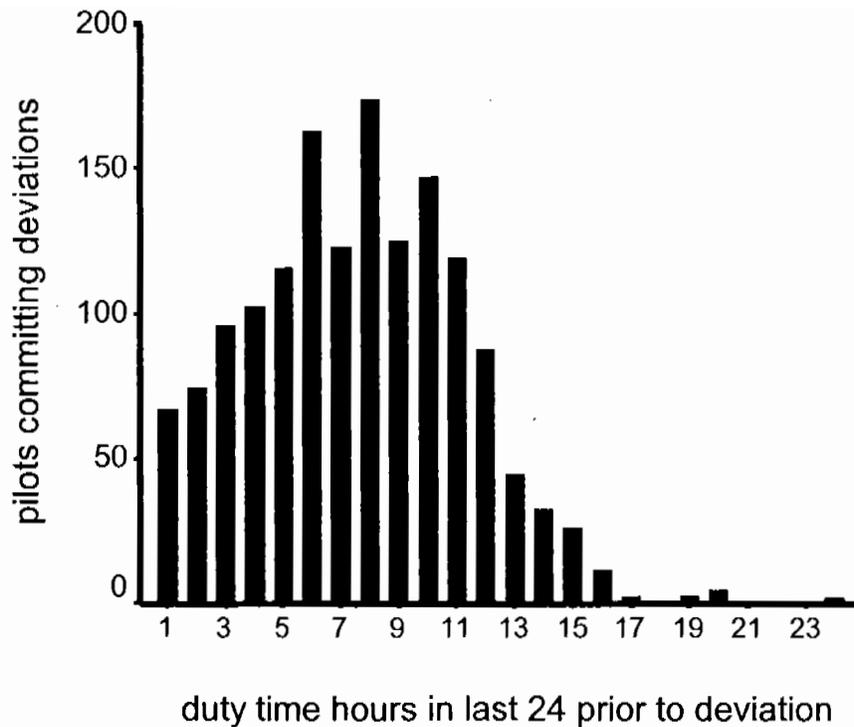


Figure 5. Pilot duty times frequency distribution.

6.1 Systematic Measure Selection

There are many ways to derive measures. Some ways might be shallow, even haphazard, due to prevailing circumstances, like insufficient measure search time because of short deadlines. Others are more formal. Among the formal ones, several originate in software engineering, a young discipline already rich in measurement methodology. Reaching into software engineering knowledge and experience shows relevancy to our framework because the discipline concerns software, a concept similarly abstract as safety. One source (Selby, 1990) points to three systematic strategies for selecting software metrics. By far, the first one is most popular in practice, as seen in the literature. Being generic, all adapt readily to safety measure selection, as summarized in the following paragraphs.

The goal-question-measure strategy requires defining first one or more broad goals spelling out the needs of the organization. These goals are then brainstormed and characterized into a set of finite questions. Next, the analyst searches for information necessary to answer them. At the same time, the analyst identifies specific measures to characterize that information.

Example: A goal might be to quantify the reliability of ATC system XYZ. Questions corresponding to this goal might be “What is the average down time for XYZ? What is the number of defects detected? What is the time between failures?” Candidate measures providing answers to these questions might be: “average of the down times of components A, B and C making up system XYZ, or their number of hardware failures by priority.”

According to the factor-criteria-measure strategy, the customer identifies several factors of interest (reliability, economy, adaptability, etc.). The analyst then interprets each factor as one or more specific criteria driving key elements in the application environment. Measures are then derived to address each criterion.

Example: A factor might be quality of ATC service. Criteria addressing this factor might consist of responsiveness, courtesy, or correctness. Corresponding measures might be chosen to be response time to fix a problem, customer rating of courtesy, or number of times the problem was fixed correctly without need for customers to re-question them.

In the classification trees and networks strategy, we first classify critical system components in terms of their individual likelihood of requiring a measurement for defect detection, quality inspection, and so on. Then, we identify measures addressing these components.

Example: ATC display panel A has been manufactured and used many times over. In contrast, panel B has just been designed. Consequently, we devise safety measures pertaining more to B than A.

6.2 Measure Qualities

Measures must show context, purpose, scope, quality, and maturity. The following list explains these qualities.

- Context spells out the specific safety facets and levels to be measured by identifying where or when measure M expects to take place in the application environment.

Example: To reach a composite safety measure, an ARTCC averages individual operational errors committed in its sectors. If so defined, this measure lacks sufficient context because some sectors have more exposure to traffic and maneuvers. Thus, they carry more risk for committing error. Think of low altitude sectors accepting and receiving dense airport traffic versus high altitude sectors handling few overflights, therefore requiring little ATC service. Normalizing the error counts with traffic load and sector type remedies this disparity.

- Purpose (intent) tells why there is need to measure certain products, processes, or resources, and how the need ties to higher (say, nationwide) goals. Purpose helps clarify the several kinds of safety goals envisioned in practice. For instance, if the goal is to size a nation's safety record, then statistical safety analysis is in order. Counting past accidents serves as a valid metric, in this case. However, if the goal is to prevent accidents, then preventive safety analysis is more appropriate. For metrics in this case, we would seek data on accident precursors rather than on the accidents themselves.

Example: Air carrier accidents, despite being rare events, draw prime interest when rulemaking decisions need to be made. As a result, there is widespread belief that the number of accidents ranks highest as a metric of choice in reflecting aviation safety. Yet their close kin, incidents, are not

only observed more frequently but also implicated as precursors to accidents. An incident is an occurrence other than accident associated with the operation of an aircraft, which affects or could affect the safety of operations. Based on this premise, incidents deserve the same intensive analysis as accident records when seeking to prevent operational safety gaps, for instance. For prevention purpose, incident metrics bring added value because they originate most likely from sample data much larger and varied than that of a few accidents.

- Scope decides how far measure M should extend economically, in terms of time and resources, vis-à-vis, the risk involved if M is not applied sufficiently.

- Quality assesses M in terms of desirable external and internal features, as follows.

(1) Desirable external features are those where

- (a) M is well documented, with language and semantics known and recognizable in the safety community. Obscure notation and mathematical expressions are low. Also, expertise is available to answer questions about the measure;
- (b) M is flexible enough to be modified to fit into a larger context of safety analysis; and
- (c) M can stand alone, results-wise, or has low or no analytical dependency on other successor or predecessor measures to be useful.

(2) Desirable internal features are those where

- (a) M carries analytical objectives, assumptions, and limitations, which are coherent without contradiction or deviation from each other;
- (b) M has low design complexity, or, if high, documentation is properly modularized and understandable. Also results are clearly displayed and easily accessible for interpretation; and
- (c) accuracy is explicit because M makes provisions for reporting error tolerance levels in the results.

- Maturity identifies the operational state of candidate measure M. This criterion is particularly useful when importing existing measures into an application.

(1) The criterion might consist of the following four levels, from least to most mature:

- (a) M remains in research, in one or few incubators, with minimal validation and verification, with some broadcast in the literature but no plans for implementation;
- (b) M is beyond the research stage, but remains in prototype form only, at very few test sites, with limited evidence of validation and with only broad plans for broadcasting and deployment;

- (c) M is implemented for restricted use, with limited validation and short past track record. Computation is reasonable and feasible; data are available or easily obtained to mechanize it; and
- (d) M shows a reliable past track record, with ample evidence of validation, widespread use, and documentation.

7. CONCLUDING REMARKS

Frameworking the measurement process leads to useful outcomes. Two such outcomes are envisioned in the following sections where measurement and concept join towards a beneficial end. Both are scenarios postulating how frameworking the measurement process characterizes safety into an increasingly formal sense.

7.1 How Safety Becomes Synonymous With Its Measurement

Question 1 in Section 2.5 within this document implies we cannot measure things before they are sufficiently defined. However, with safety as the concept being considered, no measurement can ideally take place because its abstractness would discredit any proposed measure. To simplify, we can start with a primitive, empirical definition of the elusive concept along with corresponding rudimentary measures. With time, our definition matures as we measure deeper the same or additional attributes to the point where higher meanings are replacing the original simplistic definition. Finally, at a certain stage, the safety community begins to accept comfortably the measures as defining collectively the concept itself. This scenario suggests an evolving ontology where the meaning of safety emerges defined not only in the empirical sense (in terms of its physical manifestations) but also in a quantitative sense (in terms of measures, metrics, indicators, or models.) To be realizable, this scenario depends on measurement experience shared among major players and interest groups in the aviation community. Information sharing is already taking roots on an international scale. Started in the 1990s, The Global Aviation Information Network serves as a key example (GAIN, 1997).

7.2 From Many Measures to a Few Standards

The artifact pyramid in Section 4.3 suggests measurement should be viewed as an economic-like process where its own enrichment leads progressively to higher maturity levels. This enrichment spirals up the pyramid as follows. At the bottom, numerous measures are used to size safety in the application environment. Some are temporal, serving an immediate, narrow purpose with no follow-up. Others might get promoted to metrics once additional trend data become available during the measurement process. Here again, some metrics or indicators become part of models as more attributes enter the picture. Models are subject to paradigms. Some paradigms fall aside, but others succeed in bringing measurement into proven practice. As practice widens, more organizations are encouraged to adopt these artifacts as standards. Although the aviation community might initiate numerous safety measures in specific subdomains, it should expect to end up with a few standards defining safety in the long run.

References

- Barnett, A. (1994). How numbers can trick you. *Technology Review* (pp. 38-45).
- Berka, K. (1983). *Measurement : Its concepts, theories, and problems*, Reidel, MA: D., Hingham.
- Carnap, R. (1966). *An introduction to the philosophy of science*, M. Gardner, Ed., New York: Dover Publications.
- Chrisman, N. R. (1998). *Rethinking levels of measurement for cartography, cartography and geographic information systems*, Vol. 25, No. 4, pp. 231-242.
- Churchman, C. W. (1959). Chapter 4: Why measure? In C. W. Churchman & P.Ratoosh, (eds.) *Measurement, definitions and theories*. New York: J. Wiley and Sons.
- Dewdney, A.K. (1996). *200% of Nothing: An eye-opening tour through the twists and turns of math abuse and innumeracy* New York: J. Wiley and Sons.
- Ellis, B. (1966). *Basic concepts of measurement*. Cambridge: Cambridge University Press.
- Encyclopedia Britannica, entries for *Measurement systems and measurement theory*.
- Evernden, R. (1996). The information framework. *IBM Systems Journal*, 35, pp. 37-68.
- Fenton, N., & Pfleeger, S. L. (1997). *Software metrics, a rigorous and practical approach*. Boston, MA: PWS Publishing Co.
- Finkelstein, L.(1982). Chapter 1: Theory and philosophy of measurement. In P. H. Sydenham, Editor, *Handbook of Measurement Science*, Vol. 1, New York: J. Wiley and Sons Ltd,.
- GAIN, Global analysis and information network. (1997). *Proceedings, 27-28 May*. London, UK: Royal Aeronautical Society.
- Huff, D. (1993). *How to lie with statistics*. New York: W. W. Norton and Co.
- Klein, H. A. (1974). *The science of measurement, a historical survey*. New York: Dover Publications.
- Krantz, D. H., Luce, R., Suppes, P., & Tversky, A. (1971, 1973). *Foundations of measurement, volumes I, II, III*. New York: Academic Press.
- Kyburg, H. E. (1984). *Theory and measurement*. Cambridge: Cambridge University Press.
- Law, A. (1994). *Simulation modeling and analysis, Second Edition*. New York: McGraw-Hill.
- Leveson, N. G. (2000). System safety in computer-controlled automotive systems. *Society of Automotive Engineers Conference, 8th Congress*, March 2000.

- Machol, R. E. (1995). *Thirty years of modelling midair collisions, interfaces (25) 5*, pp 151-172.
- Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin, Vol. 100*, No. 3, pp. 398-407.
- Michell, J. (1990). An introduction to the logic of psychological measurement. Hillsdale, NJ: L. Erlbaum Associates, Cambridge: Cambridge University Press.
- Narens, L., & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin, Vol. 99*, No. 2, pp. 166-180.
- Paulos, J. A. (1990). *Innumeracy: Mathematical illiteracy and its consequences*. New York: Vintage Books.
- Pfanzagl, J. (1968). *Theory of measurement*. New York: J. Wiley.
- Saaty, T. (1984). *Thinking with models*. Oxford: Pergamon Press.
- Savage, C. W., & Ehrlich, P. Eds. (1992). *Chapter 1, Philosophical and foundational issues in measurement theory*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Schwager, K. W. (1991). The representational theory of measurement: An assessment. *Psychological Bulletin, Vol. 110*, No. 3, pp. 618-626.
- Selby, R. (1990). Extensible integration frameworks for measurement. *IEEE Software*, pp. 83-on.
- Stevens, S. S. (1959). Chapter 2: Measurement, psychophysics, and utility. In C. W., Ratoosh and P. Churchman (Eds). *Measurement, definitions and theories*. New York: J. Wiley and Sons.
- Stine, W. W. (1989). Meaningful inference: The role of measurement in statistics. *Psychological Bulletin, American Psychological Association, Vol. 105*, No. 1, pp. 147-155.
- Velleman, P. F., & Wilkinson, L. (1993) Nominal, ordinal, interval and ratio typologies are misleading. *The American Statistician, Vol. 47*, No. 1, pp. 65-72.
- Wasser, B., Hauser, S., & Press, J. (1989). Evolution of a conflict resolution advisories function. *ATCA Proceedings*, Arlington, VA, pp. 173-178.
- Zachman, J. (1987). A framework for information systems architecture. *IBM Systems Journal, 26*, pp. 276-292.

