Evaluating Risk: A Revisit of the Scales, Measurement Theory, and Statistical Analysis Controversy

J.D. Solomon, PE, CRE, CMRP

Daniel Vallero, PhD, M.ASCE

Kathryn Benson, PE, CMRP

Key Words: Risk, Risk Analysis, Human Factors Engineering, Prioritization, Asset Health Management, Aging Infrastructure

SUMMARY & CONCLUSIONS

The risk management community has had a limited focus on the scales-measurement theory and typically applies all forms of parametric analysis on ordinal data. In many cases the resulting conclusions that are developed are flawed and misleading. Risk professionals should be cognizant of the impacts of the underlying theory of scales and measurements. It is the authors' position that the application of best practices that comprehensively mitigate impacts of the scales-measurement controversy are necessary for avoiding significant negative impacts that lead to misguided decision making.

1 INTRODUCTION

Risk assessment is a very inclusive and general term. In fact, the term often embodies a three-step process, beginning with a credible, scientifically sound risk assessment, which is followed by risk management and risk communication. Risk management includes engineering and the design professions. Indeed, every engineering discipline is called upon to manage risks, e.g. designing and applying pollution control technologies, incorporating prevention and green chemistry into process and product life cycles, and designing end-of-product life measures" [4].

Risk communication is actually not a single phase that follows assessment and management. It co-occurs and complements both. In the early phases of a project, potential users as well as those who may be affected by a project or product must participate in decision making. Thus, all three aspects of risk-based decision-making, i.e. assessment, management and communication, must be built upon reliable data and information. Gathering and assuring the quality and representativeness of physicochemical and social scientific data vary by type. For example, laboratory analyses and field studies usually provide ratio and interval data, whereas estimates on how devices will be used or whether actions are indeed useful may have to rely on ordinal data. It is this last dataset that requires social science instruments, e.g. questionnaires and evaluation tools.

Since its separation from risk management, scientific risk assessment has almost exclusively fallen into the domains of the physical and biological sciences [1]. This is understandable, given that the process for health risks begins with identifying the hazard and determining the likelihood of one being exposed to that hazard. However, this latter aspect of risk assessment depends on more information than is contained in the traditional biomedical and environmental sciences. Proper exposure estimates and recommendations also depend on reliable information from the social sciences [2]. For example, estimates of exposure to pesticides and many chemical ingredients in consumer products is largely determined by product use, which is the domain of human factors engineering and the behavioral sciences [3].

Mathematicians, statisticians, and behavioral scientists have hotly debated the role of measurement theory and its impact on the evaluation of opinion-based data that is typical in many surveys and qualitative or semi-quantitative assessments. This type of data is also common to many forms of risk assessments. Some of the uses may not be apparent. For example, during triage in the emergency room of a hospital, the health practitioner may ask the patient about pain, which is a risk assessment step. This is highly subjective and of little predictive value. However, if the patient is given a scale, with accompanying emotions, this ordinal information becomes more objective. Such ordinal scales are also useful in deciding whether to even undertake a design project, e.g. would it be adopted if built? The modern basis of the scientific use and evaluation of such opinion-based data can be traced in the western hemisphere to the late 1800's. Many best practices developed from this time period are still applicable today

Rensis Likert [5] is credited with the creation of one of the first ordinal data instruments, which employs the 5-point scales that are currently used today in most opinion-based surveys, including those used to evaluate risk. By example, Likert scales can be described as a survey where participants normally select a value that equates to a value or attitude. In their original form, the scale was: 1=strongly approve, 2=approve, 3=undecided, 4=disapprove and 5=strongly disapprove. Today, these scales exist in 5-, 7-, and 9-point versions and have a wide range of descriptors. It should be noted that all such scales produce only ordinal data. That is, the size of the ranges within each category will vary. All that can be derived is the order of opinion, e.g. a score of 1 from the respondent is more acceptable than a score of 2. This is why statisticians have warned against means and other measures of central tendency and are restricted to only interval and ratio data. The correctness of this view will be considered in the next section.

This paper appears in the proceedings of the international Reliability and Maintenance Symposium (RAMS 2017).

The modern measurement-scale debate had its true origins to the turn of the 20th century. Psychologists and sociologists in the era of Sigmund Freud were motivated to make their science more measureable and quantifiable, similar to movements championed by mathematicians and statisticians led by Karl Pearson in the natural sciences and economic sciences. Pearson took major exceptions to many of the non-parametric methods championed by psychologists such as Charles Spearman. When understanding the modern scales-measurement controversy that often impacts risk management, the juxtaposition of Karl Pearson's impact on engineering and financial education and the "softer", more non-parametric approach used in psychology and human behavior education should not be lost.

2 THE SCALES-MEASUREMENT CONTROVERSY

Measurement is a process of assigning numbers to objects in such a way that unique relationships of the objects are reflected as well as the unique properties of the numbers themselves. Two fundamental theories of measurement are representativeness, meaning that objects can actually be assigned to a number, and uniqueness, meaning that a number cannot be assigned to the same object by different measurers.

A central question in the analysis of opinion-based data similar to that used in the majority of risk assessments is whether the measurements and their underlying scales are both representative and unique.

One viewpoint maintains that the scales on which the measurements rests are both representative and unique, and the nature of the scale determines the statistics that should be used. The opposing viewpoint is that within the measures there is at best only loose relationship to both representatives and uniqueness, and therefore there is no relationship to the scales and the statistics that should be used [6]. The former positon dictates that the manner in which the scale is framed, administered, and analyzed, either parametrically or non-parametrically, matters. The latter positon maintains that just about anything goes in terms of the manner in which the data are analyzed. This is further complicated by the fact that risk assessments often include every type of data, from ratio data from physical and biological sciences, interval data from social sciences and ordinal data from opinions and other societally-derived information, e.g. surveys and judgment samples. Each explains a facet of potential risks, but adequate assessment of these data requires that they be combined and integrated in order to support decisions by evaluating and optimizing among available options.

This scales-measurement question is important in risk assessment and risk management because of its implications on resources, validity of conclusions, and risk mitigation planning.

2.1 *S.S. Stevens and the Opposing View(s)*

S.S. Stevens in his classic 1946 paper "On the Theory of Scales of Measurement" developed the predominate arguments for the measurement scale position [7]. He described four types of scales – nominal, ordinal, interval and ratio. Ratio scales, which are continuous, are the only type of scale that all types of statistical analysis are valid; for interval scales, a limited number of classic statistical analyses are valid. For ordinal scales Stevens states "ordinary statistics like the mean and standard deviation ought not be used"; the practical implication is that only non-parametric approaches are applicable to ordinal data. Herein lies the controversy with most survey-type data that is common in social sciences and many forms of risk assessment.

A number of well-known behavioral scientists and statisticians have taken the opposite position, which simply stated as a quote from attributed to renowned statistician Jimmie Savage "I know of no reasons to limit statistical procedures to those involving authentic operations consistent with the scale of the observed observations [8]." The controversy has taken many forms of debate over the years including whether there are only four scales as Stevens methodology suggests, whether certain parametric functions may or may not be applicable to ordinal data, and whether the role of distributions of the ordinal data makes classic statistical methods more applicable. More advanced debates concern transformations of the data itself. Many of the more famous behavioral scientists and statisticians of the second half of the twentieth century have staked a position on one side or the other of this debate.

2.2 *The Practical Perspective*

As mentioned, the dominant viewpoint is that classical parametric analysis is applicable to only ratio (continuous) data. Parametric analysis that is applied to ordinal data will yield misleading and sometimes disastrous results, as will be seen through some of the case examples provided in subsequent sections of this paper.

A practical perspective is that the negative impacts associated with application of parametric methods to ordinal data are less impactful, and can provide some positive insight, if the ordinal scales have even intervals, or essentially mimic interval data. This is not to say that the analysis is, for example, significant to multiple decimal places; it is only to say that from a practical perspective the results are not as wrong and provide some insight, which cannot be said for results created by parametric analysis on ordinal data.

The exclusive view is often violated.

3 BEST PRACTICES

Beyond applying reasonable default assumptions, the potential impacts of the issues related to scales-measurement and statistical analysis can be mitigated more directly and scientifically through survey design, administration, and analysis. This section will provide insights into these three areas dating back to Likert and several modern references directly applicable to reliability and risk professionals.

In addition to previously discussed concepts of representativeness and uniqueness that are the underlying aspects of measurement theory, the concepts of validity and reliability are introduced as key concepts that apply to surveys, including those associated with risk management. A survey is considered valid if it can be shown to measure the variable that it is intended to measure and not others. Survey reliability refers to the extent the same results are obtained with the same question when repeated to the same group of respondents.

3.1 Likert's Recommendations

Likert used two primary methods: the Sigma Method based on classic statistics for central tendency and variance; and the Simpler Method based on nonparametric methods and visualization. His primary purpose was not to create a new method but to compare results with Thurstone's more complex methodology of measuring opinion-based attitudes. Likert's measure of reliability of the results was verified using the Spearman-Brown methodology, a nonparametric test that in simple form compares the results from odd and even data sets.

The primary purpose of his work was to explore analysis aspects and appropriately that is where the most time in his work is spent. The Simpler Method was found to yield essentially the same results as either the Sigma method or Thurstone's method. Likert concludes that opinions and social attitudes are best analyzed by "clustering", and warns of the need to "cut through the statistical confusion" that was being created even is his time.

Likert provides the following recommendations concerning survey design:

- It is essential that all statements be expressions of value or desired behavior and not statements of fact;
- It is necessary for stating each question in clear, concise, and straightforward manner;
- The scale should be created where the modal reaction in in the approximate middle of the responses;
- Approximately one half of the questions should be in the affirmative (strongly approve) range and approximately half in the negative (strongly disapprove) range;
- The ONE is assigned consistently to the negative end of the scale and the FIVE to the positive end;
- In terms of measurement, it is immaterial what the extremes are called as long as the respondents' understanding of the description is the same;
- Internal consistency;
- Selection of the most differentiating statements for use in the final form of the survey.

Likert provides few recommendations in terms of survey administration. However, he clearly believes that the survey should be evaluated for reliability and used the Spearman-Brown method as a primary tool. He also used data from re-surveys of a given group of respondents up to 30 days following the initial survey.

3.2 Modern References for Best Practices

A number of modern era sources of best practices are available. In addition to the authors' own experiences, the basis for the ones provided in this paper relate closely to the field of reliability engineering and include The Handbook of Human Factors Testing and Evaluation [9], Institute for Defense Analysis [10], and the international Handbook of Survey Methodology [11].

3.2.1 Design

Perhaps as important as any aspect, the response scale and descriptor set will determine the form of the answer. This enables both representatives and uniqueness to be fulfilled from the scale measurement perspective. Some of these considerations, associated with the number and type of allowable questions include: the balance of potential answers around the mid-point, the polarity of the potential answers, and the number of values. Balanced scales are particularly valued by many analysts because they tend to produce normal distribution and in turn reduce the importance of selecting the most appropriate parametric or non-parametric statistical techniques. In terms of the number of values, it has been shown that clear discriminability can be obtained with not more than seven values (5- or 7-point Likert scales are sufficient in the majority of cases).

In terms of question construction, most modern references cite approaches that are very similar to that provided by Likert. It includes: speak to the level of the individuals who will be answering the questionnaire; avoid using jargon, acronyms, or overly technical terms that may be misunderstood by the respondents; Use positive phrases wherever possible; avoid posing two questions simultaneously; and avoid using leading or emotional questions.

3.2.2 Analysis

The Handbook makes the following statements on statistical analysis. "Obviously, questionnaire data are not a ratio measurement scale like meters or feet. Instead, questionnaire data are often ordinal measures, for example, very acceptable is better than effective. Under the best cases, where response scales and descriptor sets are based on normative data (as in the scales and descriptors recommended earlier), questionnaire data will approximate an interval scale. At their worst, questionnaires based on scales and descriptors with unknown properties represent qualitative, or categorical scales." The desire is ultimately for the construction to mimic an interval scale in order to provide the greatest flexibility of analysis. However "Questionnaire data, however, still should not be subjected to statistical analyses without careful consideration of the way in which the data are distributed [12]."

The primary means for analyzing the data will be to determine the frequency of responses for each point. The frequency of responses should then be plotted in a table or visually in a bar chart to obtain a quick and simple visualization of the results. Issues with the questionnaire, such as a potential tendency for all questions to be skewed in one direction, or in the responses themselves, such as bimodal distributions, where there are two different clusters of responses, can be easily flagged for more in-depth analysis.

In evaluating the data, a measure of central tendency and a measure of dispersion should be provided. The median should be used and it should be reported to a whole number in the case of the typical, non-continuous 5-point or 7-point Likert scale.

In the case of the potential variance, the Handbook recommends the 80th percentile. By way of example, a statement on variability may read something like "80 percent of the ratings fell at *effective* or better". When combined with the median, this is references as the 80-50 rule, which the Handbook describes being quite successful providing that questionnaire best practices are followed.

"As a source of supporting test information, they are unquestionably a cost-effective and highly flexible test methodology. Using questionnaire data as the principal evaluation item in a test, however, must be accompanied by a well-constructed and accepted criterion measure of success. Employing the principles and practices of good questionnaire design, administration, and analysis will help to ensure the success of your questionnaire ventures [13]."

3.2.3 Administration

Administration of the opinion-based surveys are a major factor of success since the human element is a major factor. Solomon and Benson use a number of best practices related to risk evaluations of infrastructure systems, some of which include: beginning and ending the session with "easier" subsystems to account for mental warm-up and to avoid the potential effects of fatigue; using a combination of paper and electronic forms; going through each question as a group to avoid the potential for respondents to feel as if they were in a mental race; and providing a comment section for each question to allow respondents to give more detail and to avoid potential frustration if a question is not well understood.

Other sources recommend: avoid giving surveys at the beginning or end of a day or work shift; avoid administering detailed surveys at the beginning or end of a work week; facilitate the questions orally; use facilitators that are available throughout the time(s) the survey is being completed to answer questions; provide written instructions or frequently asked questions so that each respondent has a common basis of knowledge; and in general any single session/scenario should be able to be completed by a respondent working independently in a 15-minute time period.

3.2.4 Hill's Criteria

Searching for clues to cancer etiology is an example of risk assessment based on disparate and seemingly unrelated data. Cancer epidemiology began with descriptive information, such as incidence and prevalence of various types of cancer in differing regions and demographic categories. This search can begin with weight of evidence and statistical association, but to determine causality, many other factors must be considered. Austin Bradford Hill [14] used a set of criteria that must be used in deciding causality: Strength of Association; Consistency; Specificity; Temporality; Biologic Gradient; Plausibility; Coherence; Experimentation; and, Analogy.

Hill's criteria can be essential to a cause, e.g. strength of association and temporality (i.e. the cause must precede the effect). Others, like coherence and plausibility, are often desirable, but may detract from the assessment if they wrongly resist paradigm shift (e.g. new hypotheses for diseases).

Likert's recommendations are in a number of ways similar to Hill's criteria. Both call upon a more complex view of disparate types of information to recognize patterns. Exclusive reliance on statistical association or ratio data can limit discovery. They are also examples of systems thinking and translational science. Risk is a complex system that often includes ill-posed problems. The analogy here is missing the forest for the trees.

3.3 Risk Matrices

The risk matrix is common tool within the risk management industry. In most forms it is a simple expansion of opinion-based surveys using Likert scales. In its most common form it incorporates either the x–axis or the y-axis with a 5-point ordinal or interval scale measuring the likelihood of occurrence and the other axis with the consequence of failure. Also known as "5x5's", risk matrices of this type are often used to produce "heat maps" where red areas can depict high risk and green areas depict low risk.

Another form of a risk matrix is in a decomposition table of either consequence or likelihood. In this form, components of consequences – say financial impacts, customer complaints, and regulatory compliance – can be shown of the left axis of the table and a 5-, 7-, or 9-point scale can be used across the top axis. By assigning a score to each component, the score can be added to produce a single composite score for the consequence of failure. The same can then be done for the likelihood.

The same foundational issues are present with different forms of risk matrices or decomposition tables. The underlying measurement and scales issues must be properly addressed to produce valid results. Perhaps the most common flaw is centered on properly handling ordinal and interval data, and in the special case of matrices the issue of scalar properties is of special importance. A second issue that is not as apparent is related to measurement properties, most notably whether a given score – say a "10" for financial impacts, customer complaint and regulatory compliance – have the same measurement meaning, especially when these scores will be combined to produce a final, singular risk score. Anthony Cox and others have numerous publications on this topic, and this paper will leave this issue with the foundational remark that the same scale, measurement, and statistics issues are relevant just as they are with more simplified approaches.

3.4 Risk Indices

Environmental risk indices combine attributes to determine a system's condition (e.g. diversity and productivity) and to estimate stresses. The original index of biotic integrity was based on the biological attributes of an aquatic system to estimate the efficiency with which the system can respond to a combination of stresses [15]. The integrity aspects are indirect indicators of a biological response to changes in structure and function, e.g. the abundance of game fish is directly related to dissolved oxygen concentrations.

3.5 Discovery and Prediction

Risk matrices and indices are particularly useful in comparative risk analysis (CRA). In any risk-based decision process, there are at least two choices, change versus status quo; but usually several alternative actions from which to choose. CRA tools aid in these decision, e.g. multi-criteria decision analysis (MCDA), analytical network process (ANP) and a multi-objective optimization by ratio analysis (MOORA). These CRA tools help to combine various types of data, less for absolute calculations than to compare among alternatives.

Comparative analysis is among several approaches to scientific discovery. Some of the most important scientific discovery depends on all types of data. Indeed, engineers depend on every form of reasoning, i.e. deductive, inductive and intuitive. Deduction draws conclusion from a general principle to a specific instance. Conversely, inductive decision-making draw general conclusions from specific observations. Much has been written on the perils of induction, but one can argue that statistical inference falls within this decision domain. Engineers, physicians and other professionals make use of both, with a healthy dose of intuition. Actually, professional intuition to a large extent is based on collective memory of past deductively and inductively derived decisions, both successes and failures. This is, in part, the reason most aspiring professionals do not actually begin to practice until they have spent years learning from a more seasoned professional who has accumulated knowledge in many forms [16].

For example, failure analysis and exposure reconstruction require not only root cause and other technical analyses, but complete analysis calls for "companion data" regarding the activities that took place leading to the failure or exposure to a hazard [17]. Risk management and risk communication usually require combinations of scientific risk assessment information with other information, including opinions about the extent to which a habitat or resource is to be protected. Thus, there is a cascade from highly reliable, quality assured measurements, which are in turn entered into models. The modeled and measured data are then input to indices. Next, surveys and other measurement-scale instruments are used to reflect societal needs. As representatives of the public safety, health and welfare, risk managers must integrate these data to build knowledge. This knowledge is the basis for risk-based decisions. The use of the term "risk" in this instance is inclusive of both evidence-based and precautionary principles. Evidence-based decisions consider a proposed action be accepted unless the decision maker can show that the risks are unacceptable. The precautionary approach puts the onus or burden of proof on the proposer of the action to demonstrate a priori that the action is safe. The precautionary approach is usually reserved for actions that can lead to severe and irreversible adverse outcomes [18]. Such predictions, such as whether a new chemical presents acceptable risks if it reaches the marketplace, require an extensive and reliable knowledge base, built from all types of data.

4 PRACTICAL IMPLICATIONS

4.1 Abuse of Survey Data

The first case comes from an evaluation conducted by a professional engineering association, which suspected it was not positioned correctly in the eyes of regulators, the public, and elected officials. Survey questions were developed and analyzed by a public relations firm and of course reviewed by licensed engineers. However, the data was treated as continuous and the mean result, extended to two decimal places, was used to reach sweeping conclusions. By way of example, a question was asked to most members of the state Assembly concerning the opinion of numerous professions on a 5-point Likert scale. The analysis yielded the following results: physicians (4.47); educators (4.33), research scientists (4.28), engineers (4.24), accountants (4.22); architects (4.21), home builders (4.05); contractors (3.91), and attorneys (3.60). The conclusion was the higher score indicated a more favorable impression and therefore engineers were better regarded than certain other professions; examining the data properly, it appears that all professions are regarded favorably. When reviewing the individual responses, the validity of the line of the questions is also of concern – none of the politicians seemed to dislike any profession as indicated by few individual scores of 1 or 2, and seemed to be have a favorable impression of all professions since each profession cluster around 4, or favorable.

While this example is one which led to limited negative consequences, except perhaps overconfidence, the same "mean taken to two decimal places" was used to force rank areas that needed improvement or were considered at risk. When analyzed properly and independently it was found that no significant differences existed. However, the leaders of the organization blindly followed the flawed recommendations. The end result was over $80,000 of expended resources on a public relations campaign to address problems that did not exist.

Another example comes from an international technical services company who began conducting 5-point Likert scale customer service studies. The results in all categories were impressive with customers indicating that they were "satisfied", or 4. However, the company's analysis used the mean rounded to two decimal places, and the erroneous conclusion drawn about needed areas of improvement compared to other areas. And worse, subsequent surveys were compared to the original survey and trend analysis was performed to two decimal places rather than properly considering the ordinal data as clustered around "4". Conclusions of progress and backsliding were drawn over the course of several years which the data when properly analyzed did not support; corporate resources were expended to address areas of perceived risk that did not exist.

A third example is cited from a regulatory-mandated management performance survey of a public utility. In this case the national consulting firm that conducted the survey measured the current and desired states of management practices using a 5-point Likert scale using the mean rounded to 3 decimal places. For 19 categories related to operations and maintenance practices, 10 categories were identified as needing improvement. When analyzed using clustering techniques, the number of areas of needed improvement dropped from 10 to 7.

4.2 Risk Evaluations

In seven of the ten risk matrix evaluations that were examined for this paper, the special case of applying scalar properties to ordinal data was found to be a major source of error. The ordinal data from 5- and 9-point Likert risk surveys was treated parametrically and in several cases multiplied as if the data were continuous. The proper treatment is for the scales to be handled non-parametrically, and if combined treated additively. A comparative analysis of the results indicated much different risk rankings and prioritization depending on the manner in which the data was treated.

From a practical perspective, the ordinal data can be treated as interval data if the scales, and their underlying attributes, are developed in this manner. The Likert data then can be treated as continuous from a practical standpoint, at least in terms of having limited impact on the risk rankings and prioritization. This was the case in 3 of the cases that were examined; however, the fact that a minority of the risk matrices were set up in this manner is consistent with the authors' experiences and underscores a significant practitioner problem.

Another major issue associated with eight of the risk matrices that were examined involved the concept of vertical consistency. By way of example, in one risk matrix involving the consequences of failure, the value associated with a human life was 10. However, a 10 was also correlated to a budget impact of $150,000 or a customer experiencing 5 days without utility service. Clearly these are not equivalent, and in this case the results were blindly aggregated without regard to vertical equivalency. Technically this is a violation of the measurement principal of representativeness. From a practical perspective, it skews many lesser important systems to be treated as equivalent to those where human health and safety should be dominant.

4.3 Practical Implications Summary

In most of the risk assessments cases, the risk analysis was technically performed wrong but the final grouping of asset and actions during risk evaluation yielded the correct results. It appears that common sense or maybe even luck overcame poor analytical techniques. In the remaining cases, the risk evaluation ended up including, or excluding, some wrong items. System values were in the range of half billion to more than one billion dollars, so the potential impacts were not insignificant.

In all three surveys, the ordinal data was treated as continuous data and parametrically. Meaningful distinctions were made in the means calculated to two to three decimal places, and significant investments were made in order to address perceived deficiencies. When the data was re-evaluated as ordinal data and non-parametrically, the need for investment in mitigation actions was reduced or eliminated.

The examples that were evaluated indicated that ordinal data is often treated poorly. In spite of this, items that ranked at the extremes, either high or low, were considered properly regardless of the strict correctness of the calculations. The items that fell in the middle were those most greatly impacted by the analysis. In terms of properly handling risk, it is less about the correct percentage of correct results and more about the specific items treated – in prioritization, the order does matter [18].

5 CONCLUSIONS

In the new age of meta-data, the authors' believe that there are greater opportunities than at any time in our history to leverage data and to utilize this resulting information to yield improved risk assessments. All data potentially leads to knowledge, and knowledge can lead to greater understanding.

Leveraging data requires being able to properly apply the correct analytical tools. In some cases, like mixed data types, this requires properly identifying and caveating underlying assumptions. To do so effectively requires better training related to the measurement-scale controversy, better integration between the social and physical sciences with respect to data treatment, and a greater focus on parametric and non-parametric statistical courses at the undergraduate level. Reliability and maintenance professionals require the profuse use of data and statistics, and should be leaders in this effort.

REFERENCES

1. National Research Council (1983). "Risk Assessment in the Federal Government: Managing the Process". Washington, DC, National Academy Press.
2. D.A. Vallero, (2015). "Exposure space: Integrating exposure data and modeling with toxicity information". OpenTox USA 2015, Baltimore, MD.
3. D.A. Vallero, S. S. I., V. G. Zartarian, T. R. Mccurdy, T. Mckone, P. Georgopoulos, C. C. Dary (2010). *Modeling and Predicting Pesticide Exposures: Hayes Handbook of Pesticide Toxicology*. R. Krieger. New York, NY, Elsevier Science: 995-1020.
4. S. Matos and J. Hall (2007). "Integrating sustainable development in the supply chain: the case of life cycle assessment in oil and gas and agricultural biotechnology." Journal of Operations Management 25(6): 1083-1102.
5. Rensis Likert, R.S. Woodruff, editor, Archives of Psychology, Columbia University, New York, Volume XXII, Nos. 146-146, 1932-1933, pp.4-43.
6. J.T. Townsend and F.G. Ashby, "Measurement Scales and Statistics: The Misconception Misconceived", Quantitative Methods in Psychology, 1984.
7. S.S. Stevens, "On the Theory of Scales of Measurement" Science, Vol. 103, No. 2684. June 7, 1946, pp. 677-680.

8. P.F. Vellman and L. Wilkinson, “Nominal, Ordinal, Interval, and Ratio Typologies are Misleading”, The American Statistician, 1993, 47:1, 65-72
9. *Handbook of Human Factors Testing and Evaluation, 2nd edition*, edited by S.G. Charlton and T.G. O’Brien, publishers Lawrence Erlbaum Associates, Inc., 2002.
10. Institute of Defense Analysis, “ICH Q9 Briefing Pack II”, July 2006.
11. European Association of Methodology (EAM), *International Handbook of Survey Methodology*, edited by E.D. de Leeuw, J.J. Hox, and D.A. Dillman, 2008.
12. Handbook of Human Factors Testing and Evaluation, 2nd edition, edited by S.G. Charlton and T.G. O’Brien, publishers Lawrence Erlbaum Associates, Inc. 2002, p 239.
13. Handbook of Human Factors Testing and Evaluation, 2nd edition, edited by S.G. Charlton and T.G. O’Brien, publishers Lawrence Erlbaum Associates, Inc. 2002, p 245.
14. A.B. Hill, (1965). "The environment and disease: association or causation?" proceedings of the Royal Society of Medicine 58(5): 295.
15. J.R. Karr, (1981). "Assessment of Biotic Integrity Using Fish Communities." Fisheries 6(6): 21-27.
16. D.A. Vallero, (2015). *Environmental Biotechnology: A Biosystems Approach*, Academic Press.
17. Y.M. Tan, et al. (2012). "Reconstructing human exposures using biomarkers and other clues." Journal of Toxicology and Environmental Health, Part B 15(1): 22-38.
18. C, G. Turvey and E. M. Mojduszka (2005). "The precautionary principle and the law of unintended consequences." Food Policy 30(2): 145-161.
19. Daniel Kahneman, *Thinking, Fast and Slow*. Farrar, Straus and Giroux. 2011.

BIOGRAPHIES

J.D. Solomon, PE, CRE, CMRP
CH2M
3120 Highwoods Boulevard
Suite 214
Raleigh, NC 27604

E-mail: jd.solomon@ch2m.com

JD Solomon is a Vice President with CH2M. He serves as a senior consultant focusing on maintenance & reliability, asset management, financial management, strategic decision-making, and master planning. He is a Certified Reliability Engineer (CRE), Certified Maintenance and Reliability Professional (CMRP), is certified in Lean Management, and is a Six Sigma Black Belt. JD has a Professional Certificate in Strategic Decision and Risk Management from Stanford, an MBA from the University of South Carolina, and a BS Civil Engineering from NC State University.

Daniel Vallero

Duke University
Pratt School of Engineering
Department of Civil & Environmental Engineering
121 Hudson Hall
Durham, NC 27708

e-mail: dav1@duke.edu

Dan Vallero is an internationally recognized expert in environmental science and engineering. For four decades, he has conducted research, advised regulators and policy makers, and advanced the state-of-the-science of environmental risk assessment, measurement and modeling. He has worked in both the executive and legislative branches of U.S. government on the most important environmental problems, including global scale atmospheric problems such as persistent, bioaccumulating oxins (so-called PBTs), ecosystem response to climate change and acid rain, and human risks posed by chemicals, such as cancer and endocrine disruption. He is the author of thirteen textbooks addressing pollution engineering, environmental disasters, biotechnology, green engineering, life cycle analysis and waste management.

Kathryn Benson
CH2M
3120 Highwoods Boulevard
Suite 214
Raleigh, NC 27604

E-mail: kathryn.benson@ch2m.com

Kathryn Benson is a project manager based in CH2M’s Raleigh office with 16 years of progressive experience in evaluating risk and reliability for the built and natural environments. She specifically focuses on O&M Optimization and the development and implementation of Asset Management Programs. Kathryn also has experience with ISO 9001 and ISO 14001 environmental management systems. Kathryn is a professional engineer in North Carolina, a Certified Maintenance and Reliability Professional, has undergraduates degrees from the University of South Carolina, and a master degree from the University of Texas.

This paper appears in the proceedings of the international Reliability and Maintenance Symposium (RAMS 2017).