DECISION SCIENCES
A JOURNAL OF THE DECISION SCIENCE INSTITUTE

Decision Sciences
Volume 42 Number 3
August 2011

© 2011 The Authors
Decision Sciences Journal © 2011 Decision Sciences Institute

An Alternative FMEA Method for Simple and Accurate Ranking of Failure Modes

James R. Bradley†
Mason School of Business, College of William and Mary, Williamsburg, VA 23187-8795, e-mail: james.bradley@mason.wm.edu

Héctor H. Guerrero
Mason School of Business, College of William and Mary, Williamsburg, VA 23187-8795, e-mail: hector.guerrero@mason.wm.edu

ABSTRACT

Failure modes and effects analysis (FMEA) is a methodology for prioritizing actions to mitigate the effects of failures in products and processes. Although originally used by product designers, FMEA is currently more widely used in industry in Six Sigma quality improvement efforts. Two prominent criticisms of the traditional application of FMEA are that the risk priority number (RPN) used to rank failure modes is an invalid measure according to measurement theory, and that the RPN does not weight the three decision criteria used in FMEA. Various methods have been proposed to mitigate these concerns, including many using fuzzy logic. We develop a new ranking method in this article using a data-elicitation technique. Furthermore, we develop an efficient means of eliciting data to reduce the effort associated with the new method. Subsequently, we conduct an experimental study to evaluate that proposed method against the traditional method using RPN and against an approach using fuzzy logic. [Submitted: July 9, 2010. Revision received: October 14, 2010. Accepted: November 18, 2010.]

Subject Areas: Collaborative Decision Making, Group/Team Decision Making, Multi-Criteria Decision Making Methods, Organizational Decision Making, Process Improvement, Quality Management and Systems, Systems/Process/Service Design

INTRODUCTION

Multicriteria ranking problems are ubiquitous in industry (Arrow & Raynaud, 1986), and one such problem is to determine which failure modes in products and processes should be mitigated first based on multiple characteristics of the failure modes. Failure mode and effects analysis (FMEA) is the most prevalent method used for this problem, and it was first applied to prioritize potential failures of product designs. Formal procedures have been defined in some industries, including the defense industry (Department of Defense, 1980) and the automobile industry (AIAG, 2008), and an FMEA is sometimes contractually required.

†Corresponding author.

Sellappan and Sivasubramanian (2008) provide a comprehensive list of industrial FMEA standards. More recently, FMEA has been used to prioritize failures in processes, and it has become an integral tool in Six Sigma process improvement. FMEA can even be simultaneously applied to product design and the production process as Wu, Kefan, Gang, and Ping (2010) have discussed in the context of concurrent engineering. The widespread adoption of Six Sigma in industry implies its frequent use by nontechnical managers in improving processes in all functions of a company including operations, sales and marketing, accounting, and information technology.

FMEA ranks potential failure modes based on three criteria: frequency of occurrence, severity of failure, and difficulty of detection. (We will use the shorthand occurrence, severity, and detection to refer to these criteria, respectively.) Each failure mode is assigned scores for each of these three criteria on an ordinal scale, where a larger ordinal score indicates a less-desirable circumstance. In traditional FMEA, a risk priority number (RPN) is then computed for each failure mode by multiplying the three ordinal scores. The failure modes are subsequently prioritized by RPN: efforts to mitigate failures are focused on the failure modes with the greatest RPN values.

A number of issues have arisen with FMEA and, in particular, with the computation of RPN values as cited in Ben-Daya and Raouf (1996), Braglia, Frosolini, and Montanari (2003), Chang, Wei, and Lee (1999), Devadasan, Muthu, Samson, and Sankaran (2003), Franceschini and Galetto (2001), Gilchrist (1993), Jenab and Dhillon (2005), Pillay and Wang (2003), Rhee and Ishii (2002), Sankar and Prabhu (2001), Sharma, Kumar, and Kumar (2005), and Wang, Chin, Poon, and Yang (2008). We summarize those criticisms below:

(i) RPN is a product of ordinal measures and is, therefore, not a meaningful measure.

  (a) Multiplication is an arbitrary choice for combining three criterion scores.

  (b) Generating an RPN on $[1, \ldots, 1000]$ from criteria scores on $[1, \ldots, 10]$ generates a fictitious increase in measurement resolution.

(ii) The method for computing RPN does not assign weights to severity, occurrence, and detection; all components are assumed to be of equal importance.

  (a) The RPN computation assumes that the scales for occurrence, severity, and detection are equivalent (i.e., a 3 on the severity scale represents the same significance level as a 3 on the occurrence or detection scoring scales).

(iii) The single RPN score does not sufficiently describe the three criteria scores.

Thus, the predominant criticisms of RPN are with its validity as a measure, its incapability to weight the three criteria, and its incapability to comprehend the full complexity of the ranking problem. Support of the first criticism can be found

in widely accepted measurement theory, which holds that the multiplication of ordinal measures results is a meaningless measure. Also, the literature suggests that the deficiency with regard to weighting is significant: Arrow and Raynaud (1986) note the frequent desire of decision makers to weight decision criteria.

A vast literature is motivated by these shortcomings of RPN where researchers have applied fuzzy logic to generate FMEA rankings with mathematically valid operations while weighting the decision criteria. We will focus on the seminal paper in this area by Yager (1981) who employs fuzzy logic to construct a general ranking procedure, and Franceschini and Galetto (2001) who applied Yager's procedure to FMEA. Since Franceschini and Galetto, many papers have also used fuzzy logic to resolve possible ranking inaccuracy due to the invalid RPN computation and the inability of RPN to weight criteria.

Although this recent research has focused on improving FMEA accuracy, the literature suggests another important characteristic of an effective FMEA, namely its simplicity. Franceschini and Galetto (2001), for example, use simplicity as a criterion to justify their approach. Tay and Lim (2010) refer to ranking by RPN as "simple and well accepted." Yang, Bonsall, and Wang (2008) also cite the ease of a traditional FMEA ranking with RPN, but they note that fuzzy approaches compromise the simplicity and transparency of the traditional approach. Yang et al., thus, signal a potential trade-off between simplicity and accuracy.

The increasing frequency of FMEA deployment motivates us to resolve this trade-off between accuracy and simplicity by defining a FMEA ranking method that is both simple and accurate. We strive for accuracy by developing a ranking method that adheres to measurement theory, provides the capability to weight decision criteria, and adheres to other technical requirements such as Pareto optimality. Our strategy is to allow arbitrary weighting of criteria by directly eliciting decision makers' relative priority for all 1,000 criteria score combinations. A ranking method must also be as simple as possible if it is to be adopted, and so we develop a method to reduce the amount of data needed from a decision maker. We develop an algorithm that requires only a subset of the 1,000 ranking scores, from which it interpolates for the unspecified data. We also supply a mathematical proof that guarantees that the data thus determined by the algorithm satisfy technical requirements (i.e., Pareto optimality). To further simplify the task of eliciting data, we develop a software tool that eases the decision maker's task of specifying ranking data by providing visual cues. We build theoretical hypotheses regarding the comparative accuracy of the ranking methods (RPN, Yager's method, and the proposed method), as well as the simplicity of the RPN method compared with our proposed method. We then test those hypotheses experimentally. Our contributions are, thus, the development of a rigorous FMEA ranking method, the development of a data-elicitation tool to support the method, and testing of the proposed method.

In the sections that follow, we first discuss the details of FMEA, and then in the subsequent section we discuss attributes of a rational ranking scheme that our proposed method must satisfy. In the following section, we describe Franceschini and Galetto's application of Yager's method to FMEA. Next, we develop our interpolation algorithm for determining a set of ranking data from partial data. The subsequent section draws upon preceding sections as it develops our hypotheses

regarding the comparative accuracy and simplicity of the RPN ranking, Yager's method, and our proposed method. The next section describes results of our experimental study that evaluate those hypotheses. The last section in this article discusses those results in detail.

FMEA IN DETAIL

As described in the previous section, traditional FMEA involves three steps: (i) determining three criteria scores for each failure mode, (ii) computing RPN scores, and (iii) ranking the failure modes in descending order by RPN. The most common ordinal scales for occurrence, severity, and detection include 10 levels represented by the integers 1 through 10: we will use the algebraic symbols O, S, and D to represent these values. Each ordinal scale value for O, S, and D is defined in either qualitative (linguistic) or quantitative terms. One can observe considerable variation in scales depending on the particular problem at hand, the individuals performing the analysis, and the organization in which the analysis is situated (cf. Gilchrist, 1993; Ben-Daya & Raouf, 1996; Chang et al., 1999; Franceschini & Galetto, 2001; Puente, Pino, Priore, & de la Fuente, 2001; Sankar & Prabhu, 2001; Pillay & Wang, 2003; Sharma et al., 2005).

The RPN is traditionally computed as the product of the three criteria scores:

$$g(O, S, D) = O \times S \times D,$$

where the function $g(\cdot)$ used to denote the RPN calculation can be written as a more general mapping from O, S, and D to *RPN* that might employ other methods besides multiplication to derive a single value representing the overall importance of a failure mode:

$$g : \{1, \ldots, 10\} \times \{1, \ldots, 10\} \times \{1, \ldots, 10\} \rightarrow \{1, \ldots, 1000\}.$$

We will refer to the output of this mapping as the importance score for a failure mode. We can write a still more general expression,

$$g : \mathcal{O} \times \mathcal{S} \times \mathcal{D} \rightarrow \mathcal{I},$$

where the importance score might be on some set $\mathcal{I}$ other than $\{1, 2, \ldots, 1000\}$ and the criteria scores might be assigned from sets $\mathcal{O}$, $\mathcal{S}$, and $\mathcal{D}$ other than $\{1, \ldots, 10\}$. For example, linguistic measures are sometimes used for importance scores, such as low importance, medium importance, and high importance. Although criteria scoring scales with 10 levels are most common, scales with fewer than 10 levels have been used (Sciemetric Instrument, Inc., 2005). For an in-depth description of FMEA, see Besterfield, Besterfield-Michna, Besterfield, and Besterfield-Sacre (2002).

CHARACTERISTICS OF RATIONAL AND PRACTICAL RANKINGS

Arrow and Raynaud (1986) define technical axioms that multicriteria decision algorithms must satisfy, which in the context of FMEA are:

Axiom 1: Pareto optimality: if for two failure modes m and n, $O_n \geq O_m$, $S_n \geq S_m$, and $D_n \geq D_m$, then $g(O_n, S_n, D_n) \geq g(O_m, S_m, D_m)$.

Axiom 2: Transitivity: for three failure modes k, m and n, if $g(O_n, S_n, D_n) > g(O_m, S_m, D_m)$ and $g(O_m, S_m, D_m) > g(O_k, S_k, D_k)$, then $g(O_n, S_n, D_n) > g(O_k, S_k, D_k)$.

Axiom 3: Independence of irrelevant alternatives: if $g(O_m, S_m, D_m) > g(O_k, S_k, D_k)$ for some failure modes $m, k \in \{1, \ldots, N\}$, then the relative ranking of m and k remains undisturbed when an additional failure mode $N + 1$ is introduced.

Pareto optimality is rational because higher criteria scores cannot lead to a lower importance score and a lower ranking. It is easily shown that the traditional RPN procedure is Pareto optimal, as is Yager's method. All methods considered in this article will also satisfy the remaining two axioms.

Arrow and Raynaud also introduce other axioms, which they argue are necessary for ranking accuracy and user acceptance:

Axiom 4: Methods should be versatile so that they do not diminish the role of the decision maker.

Axiom 5: Methods should not be so versatile that decision makers can fall victim to their biases.

Axiom 6: Methods should be easily understood by decision makers.

Whereas, Axioms 1 through 3 might be called axioms of rationality, Axioms 4 through 6 might be called axioms of adoption because the degree to which they are satisfied positively influences the probability of successful implementation. Also addressed in these axioms is the accuracy of a FMEA ranking. Specifically, the notion of versatility in Axiom 4 implies that a ranking method must allow a user the latitude to express their true preferences. In other words, the method cannot overly constrain a user's ranking to one that differs substantially from their true preferences. Conversely, Axiom 5 implies that some constraint must be present in the method to prevent an intentional or unintentional act of manipulating the rankings to match a user's *a priori* judgments. Constraint might alternately be implemented in a method such that, although the ranking mechanism is clear (Axiom 6), the mechanism is not so transparent that the results can be easily manipulated.

Other technical axioms have been advocated in the literature as well. Franceschini and Galetto (2001) have referred to "positive association of higher scores," which means that higher criteria scores should lead to higher overall importance scores for a failure mode; while related to Pareto optimality this is a stronger statement of the same notion. Additionally, Yager (1981) argues that, if criteria can be weighted, then the ranking should be influenced to a greater degree by the more highly weighted criteria.

To these axioms of adoption we would add that a method must be simple and efficient. The literature on FMEA posits that simplicity is one critical factor responsible for the widespread adoption of FMEA (Franceschini & Galetto, 2001;

Yang et al., 2008; Tay & Lim, 2010). We cannot expect users to adopt an onerous tool.

YAGER'S RANKING METHOD

Yager's ranking algorithm (Yager, 1981) was motivated by the need to weight decision criteria that are measured on ordinal scales in a manner that adheres to the rules of measurement theory. Yager's method can be used for any multicriteria ranking problem, and Franceschini and Galetto (2001) applied it to FMEA. Many papers followed Franceschini and Galetto in applying fuzzy logic and grey theory to FMEA, which are closely related to Yager's method. We demonstrate the characteristics of Yager's method in this section to motivate our proposed FMEA ranking method and to suggest hypotheses that we will test experimentally.

Summary of Yager's Method

Franceschini and Galetto use traditional FMEA criteria scoring scales of $\{1, 2, \ldots, 10\}$ in implementing Yager's method, which requires three weighting parameters defined on the same ordinal set as are the criteria scores, $W_O, W_S, W_D \in \{1, 2, \ldots, 10\}$. The weights indicate the relative importance of criteria, with larger weights reflecting greater importance. Denoting any arbitrary criterion score or weighting factor by $x, y \in \mathcal{X} = \{1, 2, \ldots, 10\}$, Yager's method uses three set operators:

(i) The intersection of x, y is defined as $x \cap y = \min(x, y)$

(ii) The union of x, y is defined as $x \cup y = \max(x, y)$

(iii) The complement of an element x is $x' = |\mathcal{X}| - x + 1$

where the cardinality of $\mathcal{X}$, $|\mathcal{X}|$, in Franceschini and Galetto's formulation is 10.

Yager's mapping of criteria scores to the importance score is

$$g_Y(O, S, D) = (W'_O \cup O) \cap \left(W'_S \cup S\right) \cap \left(W'_D \cup D\right). \tag{1}$$

Although many logic rules might be employed in such a mapping, the reader may refer to Yager's paper for the intuition reflected in Equation (1). Yager's method, importantly, adheres to the rules of measurement theory and it allows criteria to be weighted, thus addressing two main criticisms of RPN. Yager's method also satisfies the technical axioms of Pareto optimality, transitivity, and independence of irrelevant alternatives, as discussed earlier. The following subsection investigates Yager's method applied to FMEA in light of the remaining axioms.

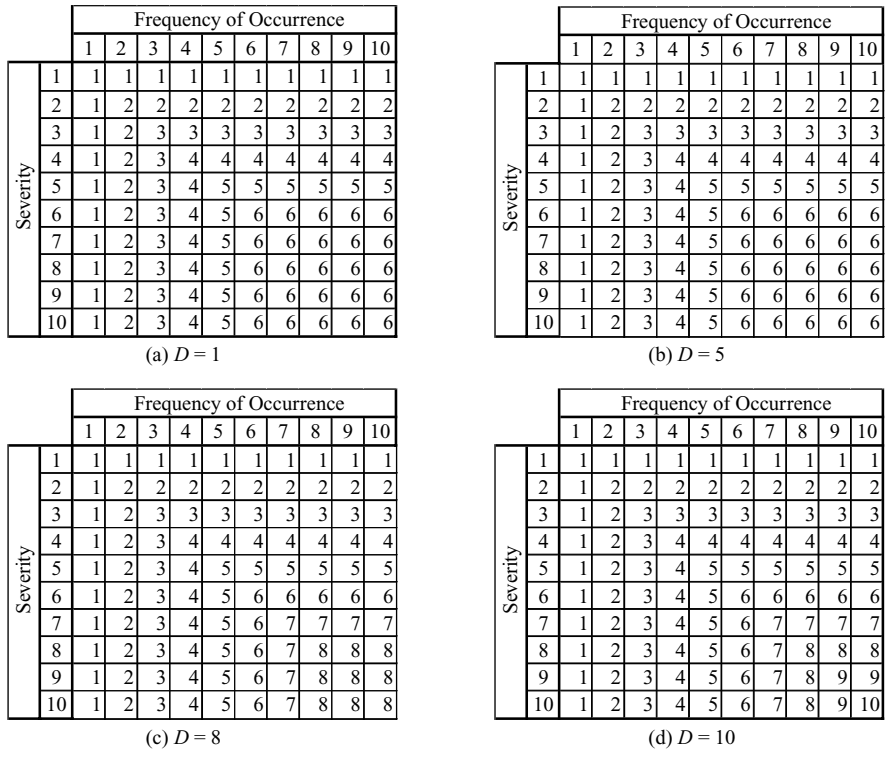
Characteristics of Yager's Method

As an example for discussion, Figure 1 shows importance scores determined by Yager's method for $D = 1, 5, 8, 10$ where $(W_O, W_S, W_D) = (10, 10, 5)$, from which we note:

(i) The importance scores are identical for $D = 1, 5$ (and also for $D = 2, 3, 4, 6$).

(ii) For $D = \{1, \ldots, 6\}$ and $S, O > W'_D = 6$, the importance score is invariant in S and O.

Figure 1: Yager weighted ranking scheme for $(W_F, W_S, W_D) = (10, 10, 5)$.



(a) $D = 1$

Severity \ Frequency of Occurrence	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	1
2	1	2	2	2	2	2	2	2	2	2
3	1	2	3	3	3	3	3	3	3	3
4	1	2	3	4	4	4	4	4	4	4
5	1	2	3	4	5	5	5	5	5	5
6	1	2	3	4	5	6	6	6	6	6
7	1	2	3	4	5	6	6	6	6	6
8	1	2	3	4	5	6	6	6	6	6
9	1	2	3	4	5	6	6	6	6	6
10	1	2	3	4	5	6	6	6	6	6

(b) $D = 5$

Severity \ Frequency of Occurrence	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	1
2	1	2	2	2	2	2	2	2	2	2
3	1	2	3	3	3	3	3	3	3	3
4	1	2	3	4	4	4	4	4	4	4
5	1	2	3	4	5	5	5	5	5	5
6	1	2	3	4	5	6	6	6	6	6
7	1	2	3	4	5	6	6	6	6	6
8	1	2	3	4	5	6	6	6	6	6
9	1	2	3	4	5	6	6	6	6	6
10	1	2	3	4	5	6	6	6	6	6

(c) $D = 8$

Severity \ Frequency of Occurrence	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	1
2	1	2	2	2	2	2	2	2	2	2
3	1	2	3	3	3	3	3	3	3	3
4	1	2	3	4	4	4	4	4	4	4
5	1	2	3	4	5	5	5	5	5	5
6	1	2	3	4	5	6	6	6	6	6
7	1	2	3	4	5	6	7	7	7	7
8	1	2	3	4	5	6	7	8	8	8
9	1	2	3	4	5	6	7	8	8	8
10	1	2	3	4	5	6	7	8	8	8

(d) $D = 10$

Severity \ Frequency of Occurrence	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	1
2	1	2	2	2	2	2	2	2	2	2
3	1	2	3	3	3	3	3	3	3	3
4	1	2	3	4	4	4	4	4	4	4
5	1	2	3	4	5	5	5	5	5	5
6	1	2	3	4	5	6	6	6	6	6
7	1	2	3	4	5	6	7	7	7	7
8	1	2	3	4	5	6	7	8	8	8
9	1	2	3	4	5	6	7	8	9	9
10	1	2	3	4	5	6	7	8	9	10

(iii) For each value of D, holding O constant, the ranking $g_Y(O, S, D)$ does not increase as S increases unilaterally for $S \geq O$. Similarly, $g_Y(O, S, D)$ does not increase when S is held constant and O increases unilaterally for $O \geq S$.

(iv) When a unit increase in a criterion score increases the importance score, the importance score always increases one unit, regardless of the criterion weight.

We now elaborate on the observations (i) through (iii) above by writing a more general and precise mathematical statement of the conditions when positive association with higher scores fails to hold. The proof is trivial and, therefore, is omitted.

Lemma 1: Denote x, y, and z as the three criterion scores, where $x \in \{O, S, D\}$, $y \in \{O, S, D\} \backslash x$, and $z \in \{O, S, D\} \backslash \{x, y\}$. Then, an increase in criterion score x by one level does not affect the importance score when either

(i) $x < W'_x$, or

(ii) $\min(x, W'_x) < y, z$

The statement holds similarly when restated for y and z.

Table 1: Frequency of positive association with higher scores (PAHS) violation among criteria weight combinations.

Criteria Scores Where PAHS Violated		Frequency of
Number	% of Scoring Combinations	Criteria Weights
516	57%	100
525	58%	100
543	60%	100
564	63%	100
612	68%	100
627	70%	100
708	79%	100
729	81%	101
801	89%	100
900	100%	99
	Total	1,000

Table 1 summarizes the frequency over all 1,000 combinations of criteria weights and shows how often at least one of the two conditions in Lemma 1 holds. In particular, Table 1 displays the number and percentage of possible criteria scoring combinations for which a unilateral increase in a criterion score does not cause an increase in the importance score, and the count of the criteria weights where each datum is observed. At a minimum, for any set of weights, an increase in a criterion score has no impact on the importance score for over 50% of the scoring combinations. For some weight combinations, a unilateral score change never impacts on the importance score or ranking of a failure mode. Positive association with higher scores should not be interpreted as the requirement that an increase in a criterion score should always cause the importance score to increase; it might be appropriate, for example, for a criterion score to increase by two levels before it induces an increased importance score. Still, Table 1 might suggest that Yager's method does not broadly reflect positive association with higher scores. This is caused by the minimum and maximum operators, which omit criteria scores so they do not affect the importance score rather than reflecting a combined effect of the criteria scores. In most situations, the nature of Yager's method to not reflect positive association with higher scores might imply the inability to model users' rational thought processes and, therefore, accurately represent their ranking preferences. Further, Jenab and Dhillon (2005) raise the issue of subjectivity in specifying criteria weights, which could affect ranking accuracy.

Consistent with argumentation in the literature, we might also desire that increases in the scores of more highly weighted criteria would cause the importance score to increase by a greater amount (point (iv) above). With Yager's method, however, when an increase in a criterion score does affect the importance score in Yager's method it always does so in the same manner, regardless of criterion weight: the importance score increases by one for every increment in criterion

score. This observation, again, raises the potential for insufficient rationality and inaccuracy.

Finally, Arrow and Raynaud suggest that a barrier to implementation might be raised if decision makers do not find the mechanics of a ranking method to be intuitive. Yang et al. (2008) state that fuzzy logic approaches to ranking inhibit clarity, thus raising an issue about adoption of Yager's and related methods.

PROPOSED RANKING METHOD

Interpolation Algorithm for Elicited Data

We are motivated by the criticisms of RPN and the preceding critique of Yager's method to develop an alternate ranking method. The design criteria are all the axioms of rationality and adoption, such that our method would be as accurate and simple as possible. We later evaluate this candidate method compared with the RPN method and Yager's method.

The method that gives a decision maker the greatest possible ranking flexibility specifies importance scores for the 1,000 scoring combinations (assuming $\mathcal{O}, \mathcal{S}, \mathcal{D} = \{1, 2, \ldots, 10\}$) with the requirement that importance scores satisfy Pareto optimality. This approach would allow an arbitrary weighting criteria and, therefore, could be adapted to any user's ranking preferences. Its versatility vis-á-vis Axiom 4 could impart greater ranking accuracy. This approach also, presumably, would be understandable to users: they would be specifying a ranking framework that would be deployed on a particular problem or problems (we will test for this hypothesis). The greatest barrier to implementing this approach is, perhaps, the burdensome task of specifying 1,000 data points. To remove this barrier, we develop a method for interpolating a full set of importance scores from a partial set specified by a decision maker. The method will be designed to satisfy Pareto optimality and the other technical axioms.

We can describe the problem of interpolating for unspecified data where importance scores are missing for one level of detection, $D = d$, as follows. Because we know $g(O, S, D)$ for $O, S \in \{1, 2, \ldots, 10\}$ and $D = \mathcal{D}/d = \{1, \ldots, d-1, d+1, \ldots, 10\}$, the task at hand is to infer $g(O, S, d)$ for $O, S \in \{1, 2, \ldots, 10\}$ such that Pareto optimality is maintained (we assume the specified data is Pareto optimal). Pareto optimality requires $g(O, S, d)$, for every combination of $O, S \in \{1, 2, \ldots, 10\}$ to be greater than or equal to each failure mode importance score that its arguments dominate, and less than or equal to the ranking of each instance where its arguments are dominated:

$$\begin{array}{l} g(O, S, d) \geq g(O', S', D) \text{ for all } O \geq O', S \geq S', \text{ and } d > D \\ g(O, S, d) \leq g(O', S', D) \text{ for all } O \leq O', S \leq S', \text{ and } d < D \end{array}. \tag{2}$$

The theorem below simplifies the Pareto optimality requirement from Equation (2), which we use in our algorithm (the proof is contained in the Appendix).

Theorem 1: The set of admissible values, $\mathcal{R}(O, S, d)$, for importance score $g(O, S, d)$ that preserve Pareto optimality for $O \in \mathcal{O}$ and $S \in \mathcal{S}$ is:

$$g(O, S, d) \in \mathcal{R}(O, S, d) \equiv \{x : x \in \mathcal{I}, g(O, S, d-1) \leq x \leq g(O, S, d+1)\}. \tag{3}$$

Note that the set $\mathcal{R}(O, S, d)$ may contain more than one admissible value. Also note that Theorem 1 assumes missing data for only one value of D but, in what follows, we iteratively apply Theorem 1 to determine $g(O, S, d)$ for multiple, consecutive missing values of D.

Simply put, Theorem 1 says that choosing an importance score $g(O, S, d)$ for any one (O, S) that is between the importance values for adjacent values of d, $g(O, S, d-1)$ and $g(O, S, d+1)$, preserves Pareto optimality between specified and interpolated importance scores. We must also ensure Pareto optimality within the interpolated data $g(O, S, d)$ for $D = d$ over all (O, S). For example, Figure 2(c) shows the admissible values $\mathcal{R}(O, S, d)$ for $d = 4$ and $d = 5$ based on interpolation between importance scores for $D = 3$ and $D = 6$, where the importance score for $(O, S, D) = (9, 7, 4)$ can be either 4 or 5. If 4 is selected then, by Pareto optimality, the importance values for $(O, S, D) \in \{(8, 7, 4), (8, 6, 4), (9, 6, 4)\}$ are constrained to 4 as well. Our algorithm, which we describe later, facilitates such Pareto optimality using the definitions below, where $\mathcal{R}(O, S, d_L, d_U)$ denotes the set of admissible importance scores between two matrices $g(O, S, d_L)$ and $g(O, S, d_U)$ with $d_U > d_L$, where importance scores are known for detection levels $D = d_L$ and $D = d_U$, and unknown for the integers in the interval $[d_L + 1, d_U - 1]$. The subscripts L and U can be thought of as lower and upper boundaries on the detection levels for which the importance scores are unknown. In Figure 2, $d_L = 3$ and $d_U = 6$. Note that there may be more than one level of detection where importance scores are unknown; that is, $d_U - d_L$ may be greater than 2.

Definition 1: A *subordinate cell* is a combination $(\overline{O}, \overline{S})$ with a cardinality greater than 1, $|\mathcal{R}(\overline{O}, \overline{S}, d_L, d_U)| > 1$, where by restricting from consideration some values from the set $\mathcal{R}(O, S, d_L, d_U)$ for at least one other combination $(O, S) \neq (\overline{O}, \overline{S})$ the cardinality of $\mathcal{R}(\overline{O}, \overline{S}, d_L, d_U)$ could decrease.

Definition 2: A dominant cell is a combination (O^*, S^*) with a cardinality greater than 1, $|\mathcal{R}(O^*, S^*, d_L, d_U)| > 1$, that is not subordinate to any other cell (O, S).

Thus, in Figure 2(c), $(O, S) = (9, 7)$ is a dominant cell and $(O, S) \in \{(8, 7), (8, 6), (9, 6)\}$ are subordinate to $(9, 7)$. Note that a dominant cell such as $(O, S) = (9, 2)$ in Figure 2(c) may have no cells subordinate to it. Denote by $\mathcal{C}_D(d)$ the cells (O, S) for a particular level of detection d that are dominant cells and $\mathcal{C}_S(d)$ the set of subordinate cells for $D = d$.

A mechanism must be defined to choose among the values in the set $\mathcal{R}(O, S, d)$ for each $(O, S) \in \mathcal{O} \times \mathcal{S}$ when it contains more than one element. Our algorithm below interpolates between importance scores at two levels of detection, $g(O, S, d_U)$ and $g(O, S, d_L)$, for $D = d$ where for $d_L + 1 \leq d \leq d_U - 1$,

Figure 2: Example interpolation of unknown importance scores for two levels of detection.

Severity(S) \ Frequency of Occurrence (O)	1	2	3	4	5	6	7	8	9	10
1	3	3	3	3	3	3	3	3	3	3
2	3	3	3	3	3	3	3	3	3	4
3	3	3	3	3	3	3	3	3	4	4
4	3	3	3	3	3	4	4	4	4	4
5	3	3	3	3	3	4	4	4	4	4
6	3	3	3	4	4	4	4	4	4	4
7	3	3	4	4	4	4	4	4	4	5
8	4	4	4	4	4	4	4	4	5	5
9	4	4	5	5	5	5	5	5	5	5
10	5	5	5	5	5	5	5	5	5	5

(a) Ranking for $D = 3$

Severity(S) \ Frequency of Occurrence (O)	1	2	3	4	5	6	7	8	9	10
1	3	3	3	3	3	3	3	3	3	4
2	3	3	3	3	3	3	3	3	4	4
3	3	3	3	3	3	3	3	4	4	4
4	3	3	3	3	4	4	4	4	4	5
5	3	3	4	4	4	4	4	4	5	5
6	4	4	4	4	4	4	4	5	5	5
7	4	4	4	5	5	5	5	5	5	5
8	4	4	5	5	5	5	5	5	5	5
9	5	5	5	5	5	5	5	5	5	5
10	5	5	5	5	5	5	5	5	5	5

(b) Ranking for $D = 6$

Severity(S) \ Frequency of Occurrence (O)	1	2	3	4	5	6	7	8	9	10
1	3	3	3	3	3	3	3	3	3	4,3
2	3	3	3	3	3	3	3	3	4,3	4
3	3	3	3	3	3	3	3	4,3	4	4
4	3	3	3	3	4,3	4	4	4	4	5,4
5	3	3	4,3	4,3	4,3	4	4	4	5,4	5,4
6	4,3	4,3	4,3	4	4	4	4	5,4	5,4	5,4
7	4,3	4,3	4	5,4	5,4	5,4	5,4	5,4	5,4	5
8	4	4	5,4	5,4	5,4	5,4	5,4	5,4	5	5
9	5,4	5,4	5	5	5	5	5	5	5	5
10	5	5	5	5	5	5	5	5	5	5

(c) Interpolated feasible values for $D = 4, 5$

Severity(S) \ Frequency of Occurrence (O)	1	2	3	4	5	6	7	8	9	10
1	3	3	3	3	3	3	3	3	3	4
2	3	3	3	3	3	3	3	3	4	4
3	3	3	3	3	3	3	3	3	4	4
4	3	3	3	3	4	4	4	4	4	5
5	3	3	4	4	4	4	4	4	5	5
6	4	4	4	4	4	4	4	5	5	5
7	4	4	4	5	5	5	5	5	5	5
8	4	4	5	5	5	5	5	5	5	5
9	5	5	5	5	5	5	5	5	5	5
10	5	5	5	5	5	5	5	5	5	5

(d) Interpolated feasible values for $D = 5$

Severity(S) \ Frequency of Occurrence (O)	1	2	3	4	5	6	7	8	9	10
1	3	3	3	3	3	3	3	3	3	3
2	3	3	3	3	3	3	3	3	3	4
3	3	3	3	3	3	3	3	3	4	4
4	3	3	3	3	3	4	4	4	4	4
5	3	3	3	3	3	4	4	4	4	4
6	3	3	3	4	4	4	4	4	4	4
7	3	3	4	4	4	4	4	4	4	5
8	4	4	4	4	4	4	4	4	5	5
9	4	4	5	5	5	5	5	5	5	5
10	5	5	5	5	5	5	5	5	5	5

(e) Interpolated feasible values for $D = 4$

using

$$\tilde{g}(O, S, d) = g(O, S, d+1) - \left\lfloor \frac{v_d}{\sum_{w=d_L}^{d} v_w} [\max \mathcal{R}(O, S, d_L, d+1) - \min \mathcal{R}(O, S, d_L, d+1)] + 0.5 \right\rfloor, \tag{4}$$

where $\lfloor z \rfloor$ denotes the greatest integer less than or equal to z. The vector $v = \{v_1, v_2, \ldots, v_{10}\}$, where each v_w for $w = 1, \ldots, 10$, allows flexibility in how the importance scores transition from those at $D = d_U$ to those at $D = d_L$ (note that v has the same cardinality as $\mathcal{D}$). Flexibility, as previously discussed, is desirable for greater ranking accuracy. If $v = \{1, \ldots, 1\}$, then interpolated values between $g(O, S, d_U)$ and $g(O, S, d_L)$ decrease from $D = d_U$ to $D = d_L$ in roughly uniform increments over $[d_L + 1, d_U - 1]$ (allowing for the integer nature of importance scores). A power series such as $v_w = 2^{w-1}$ would result in importance scores decreasing by a greater amount from $D = d_U$ to $D = d_U - 1$ than from $D = d_L + 1$ to $D = d_L$. Conversely, $v_w = (\frac{1}{2})^{w-1}$ would cause importance scores to decrease by a lesser amount from $D = d_U$ to $D = d_U - 1$ than from $D = d_L + 1$ to $D = d_L$. Further, any arbitrary set of positive values might be used for v, and it is also possible to define an array $v(O, S) = \{v_1(O, S), \ldots, v_{10}(O, S)\}$ such that v varies with O and S.

In the algorithm below we interpolate using Equation (4) mediated by constraints imposed by Pareto optimality requirements. We leave the investigation of how to specify the most appropriate vector v to follow-on research, and we use $v = \{1, \ldots, 1\}$ to induce a uniform interpolation. Our focus, instead, is on investigating the fundamental effectiveness of interpolated importance scores; moreover, our arbitrary choice of v implies that any favorable results that we might obtain regarding this proposed algorithm could possibly be improved with a more prudent specification of v.

(i) Set $n = 1$.

(ii) Use Theorem 1 to compute the set of allowable importance scores for $D = d_U - n$, $\mathcal{R}(O, S, d_L, d_U - n)$ given $g(O, S, d_U - n + 1)$ and $g(O, S, d_L)$, for each (O, S). Use that result to determine the dominant cells for $D = d_U - n$.

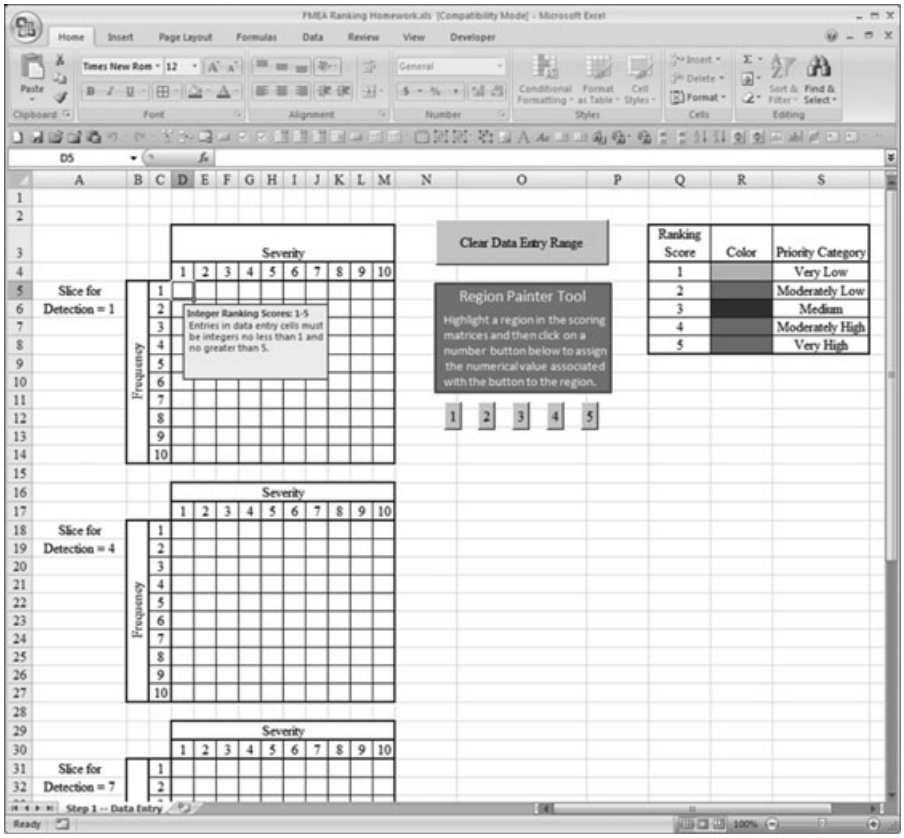
(iii) Set the importance score for each dominant cell $(O, S) \in \mathcal{C}_D(d_U - n)$ to $\widetilde{g}(O, S, d_U - n)$.

(iv) Set the importance score for each subordinate cell $(O, S) \in \mathcal{C}_S(d_U - n)$ to the minimum of $\widetilde{g}(O, S, d_U - n)$ and importance scores of cells that dominate it from $(O, S) \in \mathcal{C}_D(d_U - n)$ (Pareto optimality), which were set in the previous step.

(v) Stop if $n = d_U - d_L - 1$. Otherwise, set $n = n + 1$ and go to Step (ii).

For the example in Figure 2, the algorithm determines the importance scores for $D = d_U - n = 5$ when $n = 1$. Figure 2(d) shows that the interpolation algorithm results in an importance score of 5 for the dominant cell $(O, S) = (9, 7)$ as calculated by Equation (4). The cells $(O, S) \in \{(8, 7), (8, 6), (9, 6)\}$ subordinate to (9, 7) are also 5 because $\widetilde{g}(O, S, 4)$ is equal to 5 for all these cells and the previously determined value for cell (9, 7) does not constrain cells $(O, S) \in \{(8, 7), (8, 6), (9, 6)\}$ to a lesser value. When $n = 2$ in the algorithm, the admissible values for (O, S) and $D = d_U - n = 4$ remain the same as in Figure 2(c). The interpolated value for the dominant cell $(O, S) = (9, 7)$ as calculated in Equation (4) is 4. Subsequently, $\widetilde{g}(O, S, 4) = 4$ for $(O, S) \in \{(8, 7), (8, 6), (9, 6)\}$ which results in importance

Figure 3: Graphical user interface ranking data-elicitation tool.



scores of 4 for all of these cells because the importance score of their dominant cell is also $\widetilde{g}(9, 7, 4) = 4$.

The Spreadsheet Elicitation Tool

The interpolation algorithm was programmed in a spreadsheet, which was also used to elicit importance scores (Figure 3 shows the spreadsheet). The spreadsheet tool simplified the task of specifying importance scores by allowing users to simultaneously assign a common importance score to multiple spreadsheet cells. This was done by selecting a range of cells and then clicking on one of the buttons labeled 1 through 5. The spreadsheet used color coding to indicate the importance score values contained in the cells, which was intended to facilitate visual interpretation of a user's input as shown in Figure 4. The tool prevented users from entering values that violated Pareto optimality, which ensured rational importance scores as required by the interpolation algorithm. Importance scores could be entered and re-entered until a group was satisfied with their response.

Upon completing the specification of a subset of the 1,000 scoring combinations, users click on a button to interpolate for the unknown importance scores.

Figure 4: Example of a group's elicited ranking data.

Frequency \ Severity	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	2	2	3	3	4
2	1	1	1	1	1	2	2	3	3	4
3	1	1	1	1	1	2	2	3	3	4
4	1	1	1	1	1	2	2	3	3	4
5	1	1	1	1	1	2	2	3	3	4
6	2	2	2	2	2	2	2	3	3	4
7	2	2	2	2	2	2	2	3	3	4
8	3	3	3	3	3	3	3	3	3	4
9	3	3	3	3	3	3	3	3	3	4
10	4	4	4	4	4	4	4	4	4	4

(a) $D = 1$

Frequency \ Severity	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	2	2	3	3	4	4
2	1	1	1	1	2	2	3	3	4	4
3	1	1	1	1	2	2	3	3	4	4
4	1	1	1	1	2	2	3	3	4	4
5	2	2	2	2	2	2	3	3	4	4
6	2	2	2	2	2	2	3	3	4	4
7	3	3	3	3	3	3	3	3	4	4
8	3	3	3	3	3	3	3	3	4	4
9	4	4	4	4	4	4	4	4	4	4
10	4	4	4	4	4	4	4	4	4	4

(b) $D = 4$

Frequency \ Severity	1	2	3	4	5	6	7	8	9	10
1	1	1	1	2	2	3	3	4	4	5
2	1	1	1	2	2	3	3	4	4	5
3	1	1	1	2	2	3	3	4	4	5
4	2	2	2	2	2	3	3	4	4	5
5	2	2	2	2	2	3	3	4	4	5
6	3	3	3	3	3	3	3	4	4	5
7	3	3	3	3	3	3	3	4	4	5
8	4	4	4	4	4	4	4	4	4	5
9	4	4	4	4	4	4	4	4	4	5
10	5	5	5	5	5	5	5	5	5	5

(c) $D = 7$

Frequency \ Severity	1	2	3	4	5	6	7	8	9	10
1	1	1	2	2	3	3	4	4	5	5
2	1	1	2	2	3	3	4	4	5	5
3	2	2	2	2	3	3	4	4	5	5
4	2	2	2	2	3	3	4	4	5	5
5	3	3	3	3	3	3	4	4	5	5
6	3	3	3	3	3	3	4	4	5	5
7	4	4	4	4	4	4	4	4	5	5
8	4	4	4	4	4	4	4	4	5	5
9	5	5	5	5	5	5	5	5	5	5
10	5	5	5	5	5	5	5	5	5	5

(d) $D = 10$

That importance score framework, then, allows ranking of failure modes based on their respective occurrence, severity, and detection scores.

Development of Hypotheses

The preceding sections have discussed important characteristics of accurate and easily adopted ranking methods. We have developed a proposed method based on data elicitation and interpolation to circumvent the issues discussed with RPN and Yager's method. A possible key advantage that the proposed method possesses is that its can weight decision criteria flexibly and, thus, perhaps accurately reflect users' preferences. In this section, we synthesize the information on the RPN method, Yager's method, and the proposed method as we construct hypotheses aimed at providing evidence regarding the accuracy of the proposed method and how readily managers might embrace it. Those hypotheses rely on the RPN method as a benchmark because it is so widely used and accepted. We also have one hypothesis regarding the difference between Yager's method and the proposed ranking method.

Accuracy of an FMEA ranking is difficult to assess because the proper benchmark upon which to judge accuracy is a user's true preference, which we cannot know. Indeed, FMEA is aimed at determining a user's preference but, like any

measurement tool, we must assume the presence of measurement error. Moreover, each user's preference might be different than the next, so no one universal ranking benchmark exists for any problem. However, we can measure statistically whether the RPN method and the proposed method yield similar rankings, which can inform our assessment of accuracy. If, for example, both methods' rankings are statistically equivalent, then we can conclude that both methods are equally accurate, although we have not characterized precisely the benchmark of accuracy.

H1: Users' rankings with the RPN method and the proposed method are consistent.

We demonstrated that Yager's method often violates the axiom of "positive association with higher scores." The flexibility of our proposed method permits users to freely specify importance scores reflecting positive association with higher scores. In addition, specifying each importance score independently implies that any arbitrary weighting can be applied to the criteria. Further, Jenab and Dhillon (2005) indicate that specifying weights for criteria in methods such as Yager's can be subjective and difficult. When data is elicited as in the proposed method, numerical weights on criteria are implied, but there is no need to quantify them. Moreover, a practitioner might not comprehend in a formula such as Yager's how the relative weights affect importance scores; if this is true, it might add to the difficulty of setting criteria weights. More generally, one might find a barrier to implementation in Yager's method if decision makers do not find Yager's formula intuitive. To test if the proposed method is substantially different than Yager's method, we test the following:

H2: Users' rankings with Yager's method and the proposed method are consistent.

Our next hypothesis, aimed at evaluating whether managers would find the proposed method easy to use, is suggested by Arrow and Raynaud's argument that the users must feel as though they understand the principles of a decision making method in order for them to willingly adopt it. The following hypothesis is, thus, designed to gauge whether users' find the mechanics of either the RPN ranking or the proposed ranking method easier to understand:

H3: Users find the RPN method and the proposed method to be equivalently intuitive.

We have noted simplicity as an important hallmark of the traditional RPN ranking. Simplicity, however, albeit a very important trait of a ranking method, is not the only consideration; the literature on FMEA ranking using fuzzy logic is focused on accuracy and it raises the issue of a potential trade-off between simplicity and accuracy. Although users might find one method to be more simple or intuitive, it is not a given that they will perceive that same method to also be more accurate: users might ascribe greater accuracy to a method that is more complex and opaque. Thus, where there is a trade-off between simplicity and accuracy, a user might opt for a slightly more difficult method if the gains in accuracy are sufficient. Note that the issue here regarding likelihood of adoption is not a rigorous measurement

of accuracy but, rather, a user's perception of accuracy. We use this test to gauge users' perception of accuracy.

H4: Users perceive the RPN method and the proposed method with elicited data to be of equivalent accuracy.

The key feature of the proposed method that would make it less simple to use than the RPN-based method is specifying importance scores using the spreadsheet tool we developed. Thus, we gauge the difficulty of specifying importance scores in the following hypothesis:

H5: A majority of users perceive the task of eliciting data in the proposed method to be difficult.

EXPERIMENTAL STUDY OF FMEA RANKING

Experimental Design

The experimental design was structured to investigate our hypotheses in Section "Development of Hypotheses." The responses of groups rather than individuals were considered because Guerrero and Bradley (2011) have shown that group FMEA rankings using RPN more closely replicate experts' rankings than do individuals' FMEA rankings. The groups were composed of MBA students who were studying FMEA in a core operations course. This subject pool was selected because the subjects' knowledge and skill sets are representative of individuals who seek Lean Six Sigma certification and, thus, would implement FMEA to improve processes. Many of these students did subsequently enroll in a Lean Six Sigma certification program. The respondents were all first-time users of FMEA so they were given instruction on the traditional RPN method at the outset of the experiment. Before the last stage of the experiment, training was also provided for the spreadsheet tool (Figure 3), which was used in that stage.

The study had multiple stages. First, 96 individuals performed an RPN ranking for a problem posed in a case study by assigning occurrence, severity, and detection scores for eight predefined failure modes. In the second stage, these same individuals were assigned to 21 groups, each of which convened to determine group consensus criteria scores, which determined group-based RPN rankings of the failure modes.

In the final stage of the experiment, a partial set of importance score data was elicited from each group using the spreadsheet tool previously described. Elicitation of importance score data was simplified by requiring groups to specify scores for four levels of detection score, $D = 1, 4, 7, 10$, which constitute 400 of the 1,000 possible combinations of criteria scores. The spreadsheet tool interpolated for the missing importance data using the previously defined interpolation algorithm and then used the full set of importance data to rank the failure modes based on the groups' previously determined consensus criteria scores. We used the set $\mathcal{I} = \{1, 2, 3, 4, 5\}$ for importance scores with the proposed method; while other sets $\mathcal{I}$ with greater cardinality could have been used, we wanted to investigate whether a smaller set of ranking values would simplify the elicitation of data and increase user acceptance. Thus, the groups' consensus criteria scores gave rise to two

rankings, one based on RPN and one based on the proposed method. We compared the two rankings to evaluate H1. We were able to investigate H2 by comparing the ranking with the proposed method to the ranking with Yager's method using criteria weighting information elicited from the groups in the spreadsheet tool.

Both the Kruskal–Wallis test and the Mann–Whitney U test are nonparametric and well suited to comparing ordinal rankings, which is the focus of hypotheses H1 and H2. These tests can be applied to our data by considering each failure mode individually and comparing the distribution of ranks assigned to a failure mode by the groups using one ranking method versus another method. The sampling distribution in the Mann–Whitney is the well-known U distribution, where the U statistic is derived from putting in rank order all the groups' ranks for a particular failure mode for the two respective treatments considered in each of H1 and H2. The Kruskal–Wallis test is more general than Mann–Whitney because it can be used to compare more than two treatments, and it uses the H distribution for its sampling distribution. Our statistical tests for H1 and H2 are based on $n = 20$ rather than $n = 21$ because a portion of one group's response was corrupted.

For H1 we compared the RPN and proposed method rankings, and for H2 we compared Yager's method with the proposed method. For Yager's method in the H2 analysis, we did not ask respondents for specific criteria weights. Rather, we asked the subjects to specify the priority order of the criteria: occurrence, severity, and detection. These data allowed us to do an exhaustive search of all criteria weights to determine which weights yielded the best fit between Yager's method and the proposed method, while adhering to a group's priorities on criteria. The best fit was determined by minimizing the sum of squared differences between the rankings of the proposed method and Yager's method. This analysis was conservative because we allowed ties when searching for the best criteria weights rather then adhering to groups' strict preferences. Subsequently, we applied the Kruskal–Wallis and Mann–Whitney U tests to the resulting Yager ranking. This approach can be viewed as a conservative one because we perform the test on the best possible Yager outcome, which assumes that each group could have identified the precise criteria weights that led to that ranking. We were also able to determine which criteria weights used in Yager's method provided the best fit irrespective of the groups' stated criteria priorities. This allowed us to compare the groups' stated priorities and their implicit criteria weights employing Yager's method.

We asked the following questions to facilitate statistical tests of proportions for H3 through H5, which gauge the proposed method's prospects for adoption by comparing users' perceptions of it with the RPN-based method:

(i) (H3) "Which ranking of the failure modes did your group find most intuitive: (a) the RPN method, (b) the alternative method with elicited and interpolated data, or (c) no difference?"

(ii) (H4) "In general, which approach toward ranking failure modes would you trust most to most accurately determine the most important failure modes, the RPN method or the [elicited data method]?"

(iii) (H5) "True or False: We found the task of assigning priority categories to be relatively easy."

Our sample size ($n = 20$) is sufficient to use the normal distribution as our sampling distribution in testing differences between RPN and the proposed method (in H3 and H4) and in gauging the simplicity of specifying data in the proposed method (H5). For corroboration, we also evaluated H3 through H5 using a chi-square test.

Results

Comparison of RPN-based ranking versus proposed method

The results from the Mann–Whitney and Kruskal–Wallis tests are not materially different for H1 and H2, and so we focus our discussion on the Kruskal–Wallis here in regards to H1 and also later when H2 is discussed. The Kruskal–Wallis test results in Table 2 show that when RPN and the proposed method are compared, four of eight failure modes have statistically significant changes in their rankings at a 0.10 level of significance or greater. Additionally, another failure mode changed at a significance level just slightly greater than 0.10. Of these failure modes with statistically significant changes, failure modes 7 and 8 were ranked more highly with elicited data, whereas failure modes 3, 5, and 6 were ranked lower. We can say, then, that the ranking of failure modes based on elicited importance scores in the proposed method is statistically different than a RPN-based ranking, even when the two rankings are derived from a common set of criteria scores.

From a practical standpoint, perhaps the most relevant measure of difference between two rankings is whether a decision-maker's actions would be materially different with one method versus the other. That is, would the top-ranked failure modes that were addressed first change? We answered that question by computing descriptive statistics regarding the most critical failure modes under both ranking methods, which we defined using the idiomatic 80%–20% rule: we focused on the top two failure modes, which constitute roughly 20% of the failure modes. For 100% of the groups, the top ranked failure mode identified by RPN was also in the top two priorities with elicited data; for 70% of the groups, the second highest priority under RPN was also in the top two with elicited data; for 70% of the

Table 2: Kruskal–Wallis test results for Hypotheses 1 and 2.

	Failure Mode							
	1	2	3	4	5	6	7	8
				Hypothesis 1				
Significance level	0.176	0.372	0.045	0.925	0.053	0.027	0.102	0.001
Direction of change	↑	↑	↓	↓	↓	↓	↑	↑
				Hypothesis 2				
Significance level	0.014	0.797	0.140	0.083	0.008	0.040	0.022	0.000
Direction of change	↓	↑	↑	↑	↓	↑	↑	↓

↓/↑ indicates that a failure mode was judged less/more important relative to other failure modes, respectively, with the proposed method versus the risk priority number (RPN) method (H1) and Yager's Method (H2).

groups the top two rankings were the same for both ranking methods. Thus, there is a large degree of consistency between the two rankings generated by each group using RPN and elicited data with interpolation: for many groups the choice of ranking methodology would not result in a substantially different set of critical failure modes.

Although each group's priorities under the two methods are common to a large degree, a striking difference can be observed in the results of the two ranking methods when we observe which failure modes are identified as critical under the two methods. In both the RPN ranking and the ranking with elicited data, failure modes 7 and 4 were most frequently identified as the top two failure modes. Interestingly, 40% of the groups identified those failure modes as top priorities with RPN, whereas 80% of groups identified those failure modes as most critical with elicited data. Despite many groups' priorities not changing when using elicited data versus RPN, those whose priorities did change did so by promoting the importance of failure mode 7 at the expense of failure modes 3, 5, and 6 (Table 2). Thus, while within groups we found consistency in rankings using RPN and elicited data, among groups the method with elicited data resulted in a significantly greater consensus on which failure modes were most critical.

Comparison of Yager ranking versus proposed method

For H2 we tested for statistically significant differences of failure mode rankings between the proposed method and the best-fit Yager ranking (constrained by the groups' stated priorities among criteria) using the Kruskal–Wallis and Mann–Whitney tests. We found, as reported in Table 2, that five of eight failure modes had statistically significant ranking differences at the $p = .05$ level of significance or greater, one failure mode's ranking difference was significant at greater than $p = .10$, and the difference of another failure mode's ranking was moderately significant at $p = .140$. Thus, moderate or stronger evidence suggests ranking differences in seven of eight failure modes.

Some further descriptive statistics illuminate the significance of this difference. We computed each group's best-fit Yager ranking without constraining the criteria weights to groups' stated preferences. We then computed the ranking of the criteria weights and compared them with the groups' stated preferences. For only 3 of 20 groups did the implicit Yager criteria weights match the stated weights, and for only 5 of 20 groups did the top priority criterion match. The disparity here can be explained either by the inability of the groups to accurately state criteria priorities, the inability of Yager's method to accurately rank failure modes given users' properly stated criteria weights, or the inability of users' to specify importance scores (e.g., in the proposed method).

We observed other notable differences between Yager's method and the proposed method. First, elicited importance scores contradicted property (i) of Yager's method: all groups specified different importance score matrices for $D = 1, 4, 7, 10$. Most notably, groups whose implied Yager weights for the detection criterion were either $D = 1$ or $D = 2$, in contrast, specified unilateral increases in importance scores where $D \leq W'_D$ (specifically, for $D = 1, 4, 7$). Second, and contrary to property (iii) of Yager's method, groups' elicited importance scores did

increase with unilateral increases in occurrence and severity scores over the entire range of criterion scores. Third, contrary to property (iv) of Yager's method, where elicited importance scores increased with increased criteria scores, importance scores most often did not increase one level for every level increase in criterion score. For example, in Figure 4(a) where for $O = 1, \ldots, 5$, an increase in severity score has a differential effect on importance scores depending on whether the severity score is high or low. Twenty of 21 teams exhibited this characteristic.

Finally, another significant disparity between Yager's ranking method and elicited importance scores is visually apparent in comparing one (representative) group's importance scores (Figure 4) with importance scores generated by Yager's method (Figure 1): the regions that share common importance scores (an "isoquant") with elicited data "point to the southeast" whereas the isoquants in Yager's method "point to the northwest." In two cases, respondents also constructed diagonal isoquants, from lower left to upper right, which is not possible with Yager's method. This difference may indeed be a primary cause of the statistical ranking differences that we observed above.

Hypotheses regarding user adoption

We report in this subsection the results of the hypothesis tests H3 through H5 that we proposed earlier regarding factors affecting the likelihood of adoption for the proposed method.

For H3 we accumulated responses for those groups who either preferred RPN or found RPN and the proposed method to be equally intuitive. We tested that proportion against the null hypothesis that the same percentage of groups would prefer each method. This test is balanced in favor of RPN because RPN is given "credit" for those groups who find the methods comparable. Of the 20 groups, 8 found RPN to be more intuitive, 10 found the proposed method to be more intuitive, and 2 found no difference. We obtained consistent results using both a test of proportions using a normal sampling distribution and a chi-square test (run on the number of groups preferring the RPN rather than the proportion of groups). Using either test, we cannot reject the null hypothesis that both methods are equally intuitive with 50% of the groups either favoring RPN in this regard or finding the two methods comparable.

In responding to the question focused on H4, 6 groups perceived the RPN to be more accurate while 14 groups perceived the proposed method to be more accurate. With the null hypothesis that half of the groups would find each method more accurate, we are able to reject that null hypothesis at a significance level of 0.10 with both a Normal test and a chi-square test. Precisely, the significance level is $p = .074$.

For H5, 80% of the respondents agreed with the statement indicating that assigning importance scores was easy, and 20% disagreed. We can, thus, reject the null hypothesis at a significance level of $p = .007$ with both the normal and chi-square tests. We, thus, observe evidence that the proposed method is sufficiently easy such that a majority of groups would not find this factor a barrier in its implementation. We did not have a formal hypothesis about the effectiveness of the spreadsheet tool, but 100% of the respondents agreed with this statement: "The

color coding of the priority categories was helpful in assigning priority categories to scoring combinations." This might be one factor in the perception of a majority of groups that specifying importance scores was easy.

Qualitative responses regarding user adoption

We allowed subjects to respond to open-ended questions explaining their responses to H3 through H5. Although not actionable in a statistical sense, these comments provide depth to the preceding statistics and could also motivate hypotheses in future research.

Following the question on H3, representative comments from those favoring RPN are:

(i) "Our group likes RPN method because it is more user friendly and easier to visualize the effects of severity, occurrence, and detection. [. . .] RPN allows us to do quick calculations and come up with a result."

(ii) "With the RPN method, conceptually it is easier for us to quantify the differences. And it offers a very clear numerical ranking rather than general categories, which [are] too ambiguous."

(iii) "Because [the RPN method] is a one step process it is more direct and we can have better understanding on how the scores are calculated."

The first comment reflects a perceived simplicity in the RPN method. Similar comments, like the second comment above, based this view on the appeal of simple calculations and confidence in numerical importance scores being more accurate than qualitative category labels. The second and third comments refer to a discomfort with performing an additional step (specifying importance scores) and with the abstract notion of specifying a ranking framework for an unspecified problem. Representative comments from teams favoring the method with elicited data are:

(i) "Additionally, we felt this method was better because we can determine which factor should be considered most. The RPN method gives equal weight to the three factors; therefore, we feel that it does not best reflect the order of factors considered."

(ii) "It helps prioritize the characteristics that are most important to you. For example, two failure modes could have the same RPN making it harder to prioritize between the two. The [elicited data] method helps distinguish each failure mode by prioritizing them based on the input."

(iii) "[Failure mode 4] became a critical event under interpolation, but was of less concern under RPN. We felt that even though [failure mode 4] was the most critical failure mode, the simple RPN calculation did not reflect that. We feel that the interpolation method more accurately reflects the most dangerous failure modes."

(iv) "It allowed us to classify each occurrence with a more descriptive ranking scheme, such as very high or moderately low. This would be much more useful when looking at the problems than just raw RPN numbers."

The first three comments are representative of those teams that recognized a capability in the elicited data method to shape preferences more flexibly than in the RPN. Contrary to a comment favoring RPN, the fourth comment reflects an affinity to textual categorical labels rather than numerical scores.

Regarding H4, representative comments from those favoring RPN included:

(i) "As our interpolation method ranking came up either very high or very low, we prefer the RPN method because we think it is more accurate and evenly distributed across levels."

(ii) "RPN because we can force the importance (we make the decision) in a more detailed manner (choosing 1–10); whereas in the [elicited data] method, we choose levels (1,4,7,10) and have less choice on grades (1–5). For example, we ranked "portable case falling on child" RPN = 56 and "tank falling on child" RPN = 45, yet in the [elicited data] method, the RPN's were "moderately low" for both. In effect, you are losing sight of the difference in importance (45 → 56) by using the [elicited data] method."

(iii) "The results were based on the analysis of the data given in the case. The interpolation method was more generalized."

(iv) "In the RPN method we had to rank each failure mode ourselves with our own numbers, while the other method involved some extrapolation of the data. Our scores resulted from us looking at each mode and taking into account all circumstances, which is why we trust it more than a computer model using a couple inputs and then creating a score."

The first comment justifies the RPN-based ranking, interestingly, on a team's intuition about a characteristic that may or may not be associated with ranking accuracy. The second comment, echoing an earlier quote, shows that a team associates greater precision with numerical measures rather than categorical labels; based on the criticisms of RPN, we would argue that this may be a false precision. The remaining comments echo an earlier sentiment that assigning criteria scores is easier when focusing on a particular context than specifying importance scores absent a specific application. One comment also reflects a resistance to take ownership of the importance scores in the method with elicited data because the interpolated scores were not specified directly by the team. This observation supports Arrow and Raynaud's axioms of adoption: if a decision maker does not understand the mechanics of a method (or the method is hidden) they may find the results difficult to accept. Conversely, representative comments from teams who thought elicited data yielded a more accurate result reflected a perceived understanding of the interpolation method:

(i) "I feel that being able to weight the three metrics in this case will result in more accurate/important failure modes. Furthermore, understanding the dynamics behind the [elicited data] method versus the RPN approach, I feel more confident in the [elicited data] method because of its deeper, more detailed analysis."

(ii) "With the [elicited data] method, we are able to prioritize among severity, occurrence, and detection."

Regarding H5, representative comments include the following:

(i) "We thought it was relatively easy to just move "southeast" down the gradient from 1 to 5, and to make the different combinations become a higher priority more quickly as we moved from detection #1 to #10."

(ii) "1. The lecture was informative and helpful 2. Logic textboxes make the scoring self-explanatory"

(iii) "There was some disagreement in the group as to the appropriate numbers to assign to each category. Also there was some debate on which of the categories was the most important and thus should receive more weight. In the end we were able to come to an agreement and I think we benefited from seeing how each person thought on the issues."

(iv) "It is not that intuitive. We had to really think about what was more important and spend some time discussing about it."

(v) "It was relatively easy because reaching consensus on what should have each priority was simple within our group."

(vi) "It was easy because of the capability to assign priority categories to multiple blocks simultaneously."

(vii) "It is very intuitive. Just click, drag, and select a number. It also helps that the program recognizes when your ranking of one square does not agree with the ranking of a similar square in another matrix."

(viii) "Although this was not a daunting task, having to assign priorities to hazardous possibilities challenged us to look at the results from angles that we normally would have overlooked or ignored having done just the RPN method."

Some found the task of specifying preferences easy, while others had more difficulty. It is interesting to note that filling out the matrices did prompt discussion in some groups which many found to be an informative process.

DISCUSSION AND CONCLUSIONS

Motivated by criticisms of the traditional RPN ranking of failure modes in FMEA, as well as methods based on fuzzy logic, we developed a new method based on the idea of eliciting importance scores from users. One goal was to achieve an accurate ranking method, which we intended to achieve vis-á -vis the inherent flexibility of specifying importance scores. Another intent of our design was to create a simple and easy method to motivate its use in practice. Our method was designed to accomplish simplicity through requiring only a partial set of importance score data: our algorithm determines the remaining unknown data using an interpolation method, which we proved maintains Pareto optimality. Further simplifying data elicitation, we developed a spreadsheet tool. After designing the proposed method, we tested it experimentally to evaluate it against these design goals. Most papers on FMEA accomplish only the design task. In addition, we have also tested our design that, to our knowledge, no other paper on the design of FMEA does.

We found a reasonable level of consistency between the RPN-based method and the proposed method: for 70% of the groups the top two failure modes with RPN were the same top two goals using the proposed method. Despite this substantial correspondence borne out in the descriptive statistics. We nonetheless found a statistically significant difference in the rankings of five out of eight failure modes. That significant change in rankings was manifest by five group's changing their top two failure modes, such that fully 80% of the groups agreed on the top two failure modes with the proposed method compared with 40% with RPN. The greater coincidence in groups' rankings might be due in part to the capability of weighting criteria arbitrarily, a ranking process that is less susceptible to error in RPN determination via multiplying ordinal values, or both. Although we have substantial evidence that RPN and the proposed method are different, arguing for one method having greater accuracy is difficult because the subjects' true preferences cannot be known. Nonetheless, the broader agreement in failure mode rankings among subjects with the proposed method would be indicative of greater accuracy assuming that the consensus was on the correct answer and that, indeed, there was one correct answer. Toward the latter requirement, the experiment in this article was controlled by having all teams evaluate the same case, consider the same predetermined failure modes, receive the same training, and use the same tool. Commonality on all of these dimensions, despite the possibility of idiosyncratic group preferences, might suggest a working hypothesis regarding the proposed method's comparative accuracy with respect to the RPN-based method.

A statistical comparison of rankings from Yager's method and the proposed method revealed that the proposed method differed more substantially from Yager's method than from the RPN method; specifically, a greater number of failure modes had statistically significant differences in ranking when the proposed method was compared with Yager's method. Additionally, descriptive statistics that we reported show a discrepancy between users' stated priorities on criteria and the priorities implied in the best-fit Yager rankings. If the reasonable consistency between RPN and the proposed method can be assumed as evidence of the latter method's accuracy (i.e., assuming RPN has reasonable accuracy), then it could be argued that Yager's method might be inappropriate for FMEA problems (but not necessarily for all ranking problems). Other descriptive statistics reported here would support that position.

Regarding the practicality of the proposed method, we found in H3 through H5 that the subjects' perceptions indicated that it fared well in comparison to the RPN-based method. Users found the specification of importance scores to be easy, the visual cues of the spreadsheet tool to be effective, and the algorithm to be intuitive. Furthermore, the users perceived the proposed method's accuracy to be greater than the RPN, which would aid adoption of the proposed method.

Although the proposed method requires the specification of importance scores, it is important to note that once the importance scores have been determined by that tool, they can be deployed on similar problems in the same organization (where ranking preferences are shared). Thus, similar to the RPN calculation, the importance score can be determined by a spreadsheet function. Whereas RPN is calculated as the product of three values, an importance score in the proposed method could be calculated with an Excel VLOOKUP function. Thus, adopting

the proposed method is easier than some of the subjects in our experiment might have perceived. Additionally, we note that while a rigorous mathematical proof was required to ensure Pareto optimality of interpolated importance scores, there is no requirement for users to either understand that proof or even know of its existence: thus, it imposes no barrier to implementation.

Even though the method using elicited data is flexible in how decision criteria are weighted, some teams have demonstrated that the results might not be accepted if the method is not understood. Although improvements in the explanatory lecture might resolve the issues we observed in this regard with a minority of the groups, some teams might always disown the interpolated importance scores because the method that determines them is within a computer, which they could perceive as a "black box." Even though users provided a partial set of importance scores, they might feel that role constituted insufficient control over the ranking process. We also found evidence that some teams viewed numerical measures as being more precise than qualitative labels (even though numbers were associated with these categorical labels in the proposed method), which reflects, perhaps, a false sense of precision in the numerically expressed ordinal measures and the invalid multiplication of three ordinal values in the RPN method. Thus, decision-makers' intuitions and acceptance might not always be aligned with accuracy.

We might infer from Arrow and Raynaud's axioms of adoption that: (i) a ranking process should be sufficiently versatile to allow whatever ranking preferences are appropriate, while (ii) not being so malleable that a decision maker can intentionally or unintentionally act on their biases to promote a particular failure mode's importance without subjecting all the failure modes fairly to the ranking analytics. In the proposed FMEA ranking method, the separation of the determination of the importance score data from the problem at hand, perhaps, might have given the users space to consider their ranking preferences absent their biases about a current problem. Indeed, some groups mentioned useful discussion that ensued during the specification of importance scores. Thus, the proposed ranking method might fulfill these two of Arrow and Raynaud's axioms by giving a decision maker a flexible tool (i.e., any Pareto-optimal set of importance scores can be specified with elicited data), while prohibiting the participants from manipulating the result to fit their *a priori* opinions of a particular case.

Conversely, a minority of groups found difficulty in specifying generalized importance scores absent a particular problem. One possible solution for this would be to elicit data by asking a group to base their input on a tangible scenario. Subsequently, the elicited and interpolated data could be redeployed for other similar problems, using a simple lookup function to make computation of importance scores easy. A risk here would be that the elicited data might be intentionally or unintentionally manipulated to fit a team's preconceived ranking of the failure modes.

Because FMEA is so widely used in product and process design, we believe that more research is warranted to understand its accuracy, while retaining (if not improving) its simplicity. Our observations suggest some potentially fruitful follow-on research, including whether elicitation of importance scores should be determined in the abstract or based on an explicit case. Using our proposed method, which we have introduced in this article, further research might be done in a number

of areas: (i) how greater or lesser cardinality in the set of importance scores would affect reliability and users' impressions of the method's accuracy and intuitiveness; (ii) how a greater or lesser cardinality of criteria scores affects users' acceptance; and (iii) how various alternatives in describing the proposed method might affect decision makers' acceptance of it.

REFERENCES

AIAG (2008). *AIAG FMEA-4: Potential failure mode and effect analysis (FMEA)* (4th ed.). Southfield, MI: The Automotive Division of the American Society for Quality (ASQC) and the Automotive Industry Action Group (AIAG).

Arrow, K. J., & Raynaud, H. (1986). *Social choice and multicriterion decision-making*. Cambridge, MA: MIT Press.

Ben-Daya, M., & Raouf, A. (1996). A revised failure mode and effects analysis model. *International Journal of Quality & Reliability Management*, *13*(1), 43–47.

Besterfield, D. H., Besterfield-Michna, C., Besterfield, G., & Besterfield-Sacre, M. (2002). *Total quality management* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Braglia, M., Frosolini, M., & Montanari, R. (2003). Fuzzy TOPSIS approach for failure mode, effects and criticality analysis. *Quality and Reliability Engineering International*, *19*(5), 425–443.

Chang, C.-L., Wei, C.-C., & Lee, Y.-H. (1999). Failure mode and effects analysis using fuzzy method and grey theory. *Kybernetes*, *28*(9), 1072–1080.

Department of Defense. (1980). *Mil-Std 1629a: Military standard—Procedure for performing a failure mode, effects criticality analysis*. Washington, DC: Department of Defense.

Devadasan, S. R., Muthu, S., Samson, R. N., & Sankaran, R. S. (2003). Design of total failure mode and effects analysis programme. *The International Journal of Quality & Reliability Management*, *20*(5), 551–568.

Franceschini, F., & Galetto, M. (2001). A new approach for evaluation of risk priorities of failure modes in FMEA. *International Journal of Production Research*, *39*(13), 2991–3002.

Gilchrist, W. (1993). Modelling failure modes and effects analysis. *International Journal of Quality & Reliability Management*, *10*(5), 16–23.

Guerrero, H. H., & Bradley, J. R. (2011). Failure modes and effects analysis: An evaluation of group versus individual performance. Working Paper, Mason School of Business, College of William and Mary.

Jenab, K., & Dhillon, B. S. (2005). Group-based failure effects analysis. *International Journal of Reliability, Quality and Safety Engineering*, *12*(4), 291–307.

Pillay, A., & Wang, J. (2003). Modified failure mode and effects analysis using approximate reasoning. *Reliability Engineering and System Safety*, *79*(1), 69–85.

Puente, J., Pino, R., Priore, P., & de la Fuente, D. (2001). A decision support system for applying failure mode and effect analysis. *International Journal of Quality & Reliability Management*, *19*(2), 137–150.

Rhee, S. J., & Ishii, K. (2002). Life cost-based FMEA incorporating data uncertainty. *Proceedings of DETC2002, Montreal, Canada*, *ASME2002 Design Engineering Technical Conferences, DFM-34185*, 1–10.

Sankar, N. R., & Prabhu, B. S. (2001). Modified approach for prioritization of failures in a system failure mode and effects analysis. *International Journal of Quality & Reliability Management*, *18*(3), 324–446.

Sciemetric Instrument, Inc. (2005). *Process Signature Verification for Medical Device Manufacturing*. Accessed February 25, 2010, available at http://sciemetric.envision.s3.amazonaws.com/xmedia/i/Medical2.pdf.

Sellappan, N. & Sivasubramanian, R. (2008). Modified method for evaluation of risk priority number in design FMEA. *The ICFAI Journal of Operations Management*, *7*(1), 43–52.

Sharma, R. K., Kumar, D., & Kumar, P. (2005). Systematic failure mode effect analysis (FMEA) using fuzzy linguistic modelling. *International Journal of Quality & Reliability Management*, *22*(9), 986–1004.

Tay, K. M., & Lim, C. P. (2010). Enhancing the failure mode and effect analysis methodology with fuzzy inference techniques. *Journal of Intelligent and Fuzzy Systems*, *21*(1–2), 135–146.

Wang, Y.-M., Chin, K.-S., Poon, K. K. G., & Yang, J.-B. (2009). Risk evaluation in failure mode and effects analysis using fuzzy weighted geometric mean. *Expert Systems with Applications*, *36*(2), 1195–1207.

Wu, D. D., Kefan, X., Gang, C., & Ping, G. (2010). A risk analysis model in concurrent engineering product development. *Risk Analysis*, *30*(9), 1440–1453.

Yager, R. R. (1981). A new methodology for ordinal multiobjective decisions based on fuzzy sets. *Decision Sciences*, *12*(4), 589–600.

Yang, Z., Bonsall, S., & Wang, J. (2008). Fuzzy rule-based Bayesian Reasoning approach for prioritization of failures in FMEA. *IEEE Transactions on Reliability*, *57*(3), 517–528.

APPENDIX: PROOFS

The observation in the lemma below simplifies the construction of our interpolation method.

Lemma 2: Assuming that the specified importance data for all $\mathcal{D}/d$ constitutes a Pareto-optimal ranking, then $g(O, S, d+1) \leq g(O', S', D)$ for all $O' \geq O, S' \geq S$,

and $D \geq d + 1$, and also $g(O, S, d - 1) \geq g(O', S', D)$ for all $O' \leq O$, $S' \leq S$, and $D \leq d - 1$.

Proof of Theorem 1: The importance score that we assign to $g(O, S, d)$ for each O, S must be dominated by each known data point $g(O', S', D)$ where $O' \geq O$, $S' \geq S$, and $D > d$. The known data is Pareto optimality so that $g(O, S, d + 1)$ is dominated by all the known importance scores $g(O', S', D)$ where $O' \geq O$, $S' \geq S$, and $D \geq d + 1$, which implies that the importance score $g(O, S, d + 1)$ places the tightest upper bound on the members of $\mathcal{R}(O, S, d)$. A similar argument is possible that $g(O, S, d - 1)$ places the tightest lower bound on $\mathcal{R}(O, S, d)$. By Pareto optimality of the known data (i.e., $g(O, S, d + 1) \geq g(O, S, d - 1)$), at least one x exists such that $g(O, S, d + 1) \geq x \geq g(O, S, d - 1)$, thus $\mathcal{R}(O, S, d)$ is a nonempty set.

We must ensure that a Pareto-optimal ranking can be constructed from the nonempty set $\mathcal{R}(O, S, d)$ for each $O \in \mathcal{O}$, $S \in \mathcal{S}$ for each $g(O, S, d)$, which must (weakly) dominate all $g(O', S', d)$ where $O > O'$, $S > S'$, and which must be weakly dominated by each $g(O', S', d)$ for $O < O'$, $S < S'$. If we can show that a Pareto-optimal importance score can be determined for each $g(O, S, d)$ with respect to its immediate neighbors, then, by transitivity, that score will be Pareto optimal with respect to all $g(O, S, d)$ for each O, S. A feasible choice for the upper bounds on $g(O, S, d)$, which are $g(O + 1, S, d)$, $g(O, S + 1, d)$, and $g(O + 1, S + 1, d)$, comes from the known data:

$$g(O + 1, S, d) = g(O + 1, S, d + 1)$$
$$g(O, S + 1, d) = g(O, S + 1, d + 1)$$
$$g(O + 1, S + 1, d) = g(O + 1, S + 1, d + 1),$$

which by Pareto optimality must be at least as great as the upper bound on $g(O, S, d)$:

$$g(O, S, d) \leq g(O, S, d + 1) \leq g(O + 1, S, d + 1) = g(O + 1, S, d)$$
$$g(O, S, d) \leq g(O, S, d + 1) \leq g(O, S + 1, d + 1) = g(O, S + 1, d)$$
$$g(O, S, d) \leq g(O, S, d + 1) \leq g(O + 1, S + 1, d + 1) = g(O + 1, S + 1, d).$$

Thus, the largest elements in $\mathcal{R}(O + 1, S, d)$, $\mathcal{R}(O, S + 1, d)$, and $\mathcal{R}(O + 1, S + 1, d)$ are at least as large as the largest element in $\mathcal{R}(O, S, d)$. A similar argument indicates that the smallest elements in $\mathcal{R}(O - 1, S, d)$, $\mathcal{R}(O, S - 1, d)$, and $\mathcal{R}(O - 1, S - 1, d)$ are at least as small as the smallest element in $\mathcal{R}(O, S, d)$ so that a Pareto-optimal ranking is possible. □

James R. Bradley is a professor in the Mason School of Business at the College of William and Mary in Williamsburg, VA where he teaches manufacturing, operations management, supply chain management, Lean Six Sigma process improvement, and information technology at the graduate and undergraduate levels. He was previously on the faculty at the S. C. Johnson Graduate School of Management at Cornell University. He has published academic research on supply chain management, supply chain risk management, life cycle management, inventory management, lean manufacturing, performance measurement, and the joint

optimization of manufacturing capacity, inventory, and subcontracting policies. His research makes extensive use of applied probability, optimization, and computer simulation. Prior to earning his PhD, he worked for 15 years in manufacturing with General Motors. His consulting clients have included 3M, Digital Equipment Corporation, the Virginia Port Authority, and the Commonwealth of Virginia Employment Commission.

Hector H. Guerrero is a professor in the Mason School of Business at the College of William and Mary in Williamsburg, VA. He teaches in the areas of data analysis; operations management; quantitative methods, modeling, and simulation; and Six Sigma/statistics. Professionally he is active in the areas of operations management, information systems, and engineering management. He has published articles on topics related to logistics, material requirements planning, automated manufacturing, group technology, robotics, forecasting, supply chain management, product design, and demand management. He has taught at the Tuck School of Business (Dartmouth College), University of Notre Dame, and Instituto Latino Americano de Estudios Sociales (Santiago, Chile). Prior to entering academia, he worked as an engineer for Dow Chemical Company and Lockheed Missiles and Space Co. He has been active in executive education and has consulted with a wide variety of clients such as the U.S. Government (TSA, DoD, DoC, DoL, DoJ), Latin American and European firms, as well as many small and large U.S. manufacturing and service firms. He is author of a new book: *Excel Data Analysis, Modeling and Simulation*, published by Springer-Verlag.