Problems with scoring methods and ordinal scales in risk assessment

Article *in* Ibm Journal of Research and Development · May 2010

DOI: 10.1147/JRD.2010.2042914 · Source: DBLP

2 authors:

Douglas Hubbard
Hubbard Decision Reseaarch

Dylan Evans
American University of Beirut

Problems with scoring methods and ordinal scales in risk assessment

D. Hubbard
D. Evans

Risk assessment methods based on scoring methods that rate the severity of each risk factor on an ordinal scale are widely used and frequently perceived by users to have value. We argue that this perceived benefit is probably illusory in most cases. We begin by describing a number of common scoring methods currently used to assess risk in a variety of different domains. We then review the literature on the use of ordinal scales in risk analysis, the use of "verbal scales" for eliciting estimates of risks and probabilities, and the extensive research about peculiar human errors when assessing risks. We also supplement this overview with some data of our own. When these diverse kinds of evidence are combined, the case against scoring methods is difficult to deny. In addition to the evidence against the value of scoring methods, there is also a lack of good evidence in their favor. We conclude our overview by reviewing the reasons why risk assessment approaches should describe risk in terms of mathematical probabilities.

Introduction

Many methods for risk assessment involve the use of scoring methods in which the severity of each risk factor is rated on an ordinal scale. The resulting values are then combined by additive weighting or by multiplication to compute an aggregate measure of overall risk. In this paper, we provide an overview of the existing research and argue that, taken together, these diverse examples of evidence suggest that scoring methods are not useful tools for risk assessment. In arguing for this conclusion, we are not merely claiming that arithmetically combining ordinal scales is logically invalid. This is true regardless of the risk management context. Rather, we claim that the widespread use of scoring methods in risk assessment is flawed for a wide variety of reasons. To a mathematically sophisticated reader, some of our points may appear unoriginal or obvious. This, however, would be to miss our primary point: The widespread use of scoring methods in real-world settings is still a serious problem that needs addressing, and these methods are beset by many flaws aside from the mathematical ones.

We start by describing a number of common scoring methods currently used to assess risk in a variety of different domains. We then identify and discuss four important problems associated with the use of these scoring methods, which we may briefly state as follows. First, they do not usually take into account the findings of psychological research concerning the cognitive biases that impair most people's ability to assess risk. Second, the verbal labels used in ordinal scales are interpreted extremely inconsistently among different users and by the same user. Third, many users treat these scales as if they are ratio scales, with the result that they draw invalid inferences. Fourth, simple scoring methods rarely consider correlations that would change the relative risks. Taken together, these four problems indicate that scoring methods are likely to be poor tools for risk assessment.

The only "evidence" in favor of using scoring methods in risk assessment consists of testimonials from users of such methods, claiming that they perceived great value in these methods. Published case studies even exist in the literature describing the application of such methods. However, subjective perceptions of benefit do not always correlate with objective measures of success. Such objective evidence in favor of scoring methods in risk assessment is lacking. When the lack of evidence in favor of such methods is combined with the powerful arguments against them, one is forced to conclude that scoring methods should not be used for risk assessment. We conclude this paper by arguing that risk

Digital Object Identifier: 10.1147/JRD.2010.2042914

assessment should instead involve the use of mathematical probabilities.

Scoring methods for risk assessment: A brief survey

Many risk management methods promoted by management consultants and international standard organizations involve calculating a numerical value for risk based on simple point scales (defined below) combined in some way. Such scales are subjective and are usually based on some kind of ordinal comparisons or classifications.

On an ordinal scale, factors such as likelihoods are assigned numbers in such a way that the order of the numbers reflects the order of the factors on an underlying attribute scale. Two factors x and y with attribute values $a(x)$ and $a(y)$ are assigned numbers $m(x)$ and $m(y)$ such that $m(x) > m(y)$ if $a(x) > a(y)$.

Two broad categories of subjective ordinal risk assessment methods are widely used. We will refer to them as additive and multiplicative. Additive scores are those in which several independent factors are weighted and added together. For example, when evaluating the credit risk of an international customer, a manufacturer might identify factors for country stability, order size, currency volatility, and so on. Additive scores tend to be used to evaluate overall risks of projects, investments, and policies, for example. Multiplicative scores are those that have just two or three factors that are multiplied together. When two factors are used, they are usually *likelihood* and *impact* (or probability and consequence), and they are generally not individually weighted as in additive schemes. These scores are often used to represent the evaluation of the risk of individual events, such as theft of data by hackers (i.e., malicious users) or power outage in the factory. Some three-factor variations of multiplicative scores, particularly in security-related areas, use *threat*, *vulnerability*, and *impact*.

In each of these approaches, the following characterizations apply. First, some set of factors is identified—possibly many factors for additive methods but usually just two or three for multiplicative methods. Next, each factor is rated according to some kind of point scheme. A typical framework is an ordinal scale of 1–5, but other scales are also used. If project duration is considered a risk factor in an additive scoring method for information technology (IT) project risks, then the duration in months might be converted to discrete values on this five-point scale. Sometimes, simple three-point *high*, *medium*, and *low* scales are used. After rating each factor, the factors are then combined. For additive scoring methods, weights are applied to each factor, and then the weighted values are added. In scoring methods that simply utilize likelihood and impact, the weighted values are usually multiplied. Sometimes, the ordinal scale is converted to yet another scale before the weighted values are combined. For example, on a multiplicative scale, a "low" impact score might be converted to a 1 and a "high" impact converted to a 10. Finally, the resulting value is compared with yet another arbitrary scale (e.g., a score of 10–18 is "medium" risk, and a score of greater than 25 is "extreme risk"). In the case of multiplicative scores, the likelihood and impact might be plotted on a so-called "risk matrix" in which different regions are defined as a high/medium/low risk.

These sorts of weighted scoring methods are relatively easy to create and teach. Consequently, the use of scoring methods tends to spread quickly, and the methods have become popular for several applications. Respected organizations have designed such methods and represent them as "best practice" for thousands of users. For example, the U.S. Army has developed a weighted scoring-based method for evaluating the risks of missions [1]. The U.S. Department of Health and Human Services uses a weighted scoring method to determine vaccine allocations in the event of a pandemic outbreak [2], and the U.S. National Aeronautics and Space Administration (NASA) uses a scoring method for assessments of risks in manned and unmanned space missions [3, 4].

Additionally, a variety of influential standards that are now considered as "best practice" in the information-technology industry include proposed risk assessment methods based on ordinal scores. For example, the National Institute of Standards and Technology (NIST), which is a federal agency of the U.S. tasked with setting measurement standards, has developed the NIST 800-30 standard for assessing information-technology security risks [5]. The Project Management Institute (PMI), which describes itself as the "world's leading not-for-profit association for the project management profession" on its website, has trained and certified thousands of people in its methods that include a scoring approach to evaluating project risk. PMI membership includes many IT professionals as well as non-IT professionals. In its Project Management Body of Knowledge, PMI proposes its own risk scoring approach based on ordinal scales [6]. Finally, the Information Systems Audit and Control Association has certified thousands of professionals in its training programs and has developed the Val IT and Risk IT methods for evaluating the value and risk of IT investments.

Of the cited examples above, all but two are additive scores. The NASA and NIST scores are multiplicative. The NIST score is an example of a scheme that uses high/medium/low categories and then converts those categories to another quantity (0.1, 0.5, and 1 for likelihood; 10, 50, and 100 for impact) before multiplying them together.

Clearly, weighted scoring methods have been adopted by influential organizations for assessing a variety of risky decisions involving not only major investments and projects but also public health and safety. If these methods are flawed, as we argue in the following section, then the consequences could be serious.

Four problems with scoring methods

In this section, we argue that scoring methods are badly flawed and identify four associated problems. Before examining each problem in turn, however, it is valuable to ask why these methods have spread so widely if they are really as problematic as we claim. One main reason may be that these methods tend to be applied to decisions for which the results are not immediately apparent, and the quality of the decision is not obviously measurable. If an organization uses a scoring method to rank the risks of some relatively infrequent event (e.g., flu pandemics, engineering catastrophes, or major IT security breaches), a very long time may be required to determine whether the assessments effectively track outcomes. In addition, this assumes that the organization is systematically tracking the event and statistically analyzing the results with respect to original forecasts, which is not often the case. More often, managers may be interpreting anecdotal outcomes in a way that supports a particular assessment method. As a result, scoring methods are rarely evaluated in a rigorous fashion, and they persist despite their uselessness or even harmfulness.

From our review of the literature, we have identified four important problems, which we mentioned in the introduction, associated with the use of these scoring methods. We now examine each of these four problems in turn.

Partial review of literature on cognitive biases

Several decades of empirical research in psychology have revealed that people are ineffective at assessing risk. Moreover, they depart from normative standards in certain systematic ways. In other words, people are not only irrational, they are also "predictably irrational" [7]. In this section, we provide a brief and partial review of the literature on cognitive bias and discuss some of the implications for the use of scoring methods in risk assessment.

The systematic ways in which people make poor risk assessments are often attributed to a well-documented set of cognitive biases that systematically skew human judgment in certain directions. For example, when estimating the frequency of dangerous events, most people ignore available statistical information about base rates and instead base their estimates on their memories, which are biased toward vivid, unusual, or emotionally charged examples. Researchers refer to this bias as the "availability heuristic," referring to the recall of "available" memories. Most people also tend to assume that individual random events are influenced by previous random events (referred to as the *gambler's fallacy*), and overestimate the probability of good things happening to them compared with other people (referred to as the *optimism bias*). Most people also search for or interpret information in a way that confirms their preconceptions (*confirmation bias*) and claim more responsibility for successes than failures (*self-serving bias*). Many other cognitive biases have been identified by psychologists.

Other phenomena associated with the risk estimation process may affect subjective responses. Anchoring, for example, is the phenomenon of making estimates based on adjustments to a previously conceptualized quantity. This effect appears to occur even when the previous quantity is unrelated to the new estimate. In one experiment, the researchers conducted a survey of college students in which one question asked for the last four digits of their social security number. The very next question asked for an estimate of the number of physicians in Manhattan. The researchers found a correlation of 0.4 between these two answers, indicating that estimates of physician numbers had been influenced by the estimators' social security numbers, although these are clearly unrelated. The same researchers found that questions that are logically identical, yet "framed" differently (such as survival probabilities versus mortality probabilities for medical procedures), can give rise to significantly different responses [8]. These effects are routinely considered in the design of public opinion surveys but are not considered in nearly all of the common risk assessment methodologies.

Perhaps one of the most significant biases related to risk assessments is that of overconfidence [9, 10]. Most people consistently overestimate their certainty about a forecast. In other words, if a manager believes that there is a 90% chance some event will not occur, and this is tracked for a large number of such events, then he will be correct much less than 90% of the time. A review of several experiments shows that when subjects believe that they have identified a range that has a 90% chance of containing the correct value, their track record shows that this range only contains the correct value between 55% and 78% of the time. Since assessing risk must involve assessing the likelihood of some event, overconfidence means that subjective estimates of chance will consistently be biased so that somewhat unlikely events will be underestimated. In other words, if a person believes that there is a 90% chance that a catastrophic event will not happen, then the chance the event will occur may be much higher than they believe. In the technical jargon, subjective probabilities are often *miscalibrated*.

Another important phenomenon is not so much a bias—which tends to be consistent, by definition—but an inconsistency itself. A person asked to evaluate the same item on different occasions tends to give different estimates, although no change in information justifies the change in judgment. Simple linear regressions of human judgments based on various questions as independent variables in a judgment model improve upon unaided human judgment [11, 12]. In other words, when people use such linear models, they are more consistent than when they rely entirely on intuition, and this alone improves performance. One important reason for this is that the linear model removes human inconsistency. Studies of this "intrapersonal reliability" show it to be fairly low [13]. For example, a study of x-ray specialists showed that the same set of cases examined a week apart had a correlation of the prognoses of just 0.6, indicating that specialists may make quite different prognoses when presented with the same cases [14]. The same specialists were giving prognoses that were different for no reason other than random variation.

These biases and related challenges are persistent and are not overcome without specific and extensive training, controls, and adjustments. Some progress has been made in devising mental techniques and tools that people can use to overcome or at least reduce bias and thereby make more rational decisions. These debiasing strategies include meta-cognition, instruction in rational choice theory and probability, decision-support systems, and cognitive forcing [15]. Inconsistency seems to partially offset by repeatedly asking a number of similar questions and statistically assessing the variance. The effects of response item order (due to anchoring) can be offset by asking different persons the same questions in different orders.

There is much more literature on heuristics and biases than would fit within the scope of this paper. However, despite the significance of the research on cognitive biases and its relevance to risk assessment, none of the common scoring methods account for such bias. Because assessing likelihood is so central to risk assessments, *overconfidence* alone would be a debilitating shortcoming for any risk assessment if left unaddressed. The effects of anchoring, framing, and inconsistency would only further compound this challenge.

Existing research regarding the variability of verbal labels

A characteristic of virtually all scoring methods is the use of some verbal labels for eliciting judgments. Likelihood, for example, might be put into categories of "very likely," "likely," "as likely as not," "unlikely," and "very unlikely." Additionally, as in the case of the NIST 800-30 security model, an even simpler "high," "medium," or "low" differentiation may be used. Similarly, such verbal expressions are used to categorize impact (e.g., "moderate" or "extremely severe"). Such kinds of verbal methods are used because it is believed that users of these methods will have better comprehension and, therefore, more reliable use of the method if simple labels are used. It is also assumed that if the labels are further defined in great detail, then the scales will be used consistently. Unfortunately, neither of these assumptions is valid when scrutinized in an empirical fashion. In this section, we review some of the existing evidence that undermines these assumptions.

For example, the Intergovernment Panel on Climate Change report made use of verbal expressions of likelihood instead of quantitatively expressed probabilities. A table was provided to define what each of the labels meant in quantitative terms. The table would define "very likely" as meaning "greater than 90%" and "unlikely" as meaning "less than 33%." Research showed that when given specific uses in context (e.g., "It is very likely that hot extremes, heat waves, and heavy precipitation events will continue to become more frequent") and even when provided with detailed definitions of the probability labels, subjects who read the report interpreted the labels to mean a wide variety of possible values even so far as to violate the defined rule [16]. In other words, even when users were given tables that explained what range of values were valid for specific labels, users still interpreted the labels differently—sometimes even outside of the specified ranges. It was even possible for the label "unlikely" to be interpreted as meaning "as much as a 66% probability."

In the same manner, the NIST 800-30 guideline for assessing IT security risks defines high, medium, and low for both likelihood and impact. The definition of "low likelihood" states that "the threat-source lacks motivation or capability, or controls are in place to prevent, or at least significantly impede, the vulnerability from being exercised" and goes on to state that this will be converted to a likelihood of 0.1.

When different individuals interpret the same labels to mean very different things, there can arise an "illusion of communication" [16]. Subjects may describe the probability of a given event with the same verbal label and conclude on this basis that they agree; however, since they may implicitly attach different probability ranges to the verbal label, their agreement may be illusory. To further complicate matters, the same individual may attach a different probability range to the same label in different contexts. The additional variance caused by the inconsistent interpretation of these labels—even when care is taken to provide detailed definitions—is another source of error in most scoring models.

Invalid inferences

Arbitrary features of the scoring scheme itself often have a larger impact on results than the users might be aware of. In the first place, people who use them treat the values arrived at by means of these methods as if they were values on a ratio scale, when in fact they are, of course, values on an ordinal scale. This lack of care leads users to draw invalid inferences that are always misleading and sometimes harmful.

In 1946, Stevens introduced the idea of different "scales" of measurement [17]. The scales describe the nature of information contained within the numbers assigned to attributes. Among the various levels of measurement identified by Stevens were ordinal and ratio scales. Note that in this paper we limit our discussion to the choice between ordinal and ratio scales. Additionally, we do not believe that the nominal and interval scales that Stevens also identified are relevant to the debate on methods for risk assessment. Unlike ordinal scales (described earlier), ratio scales assign numbers to attributes so that differences and ratios between

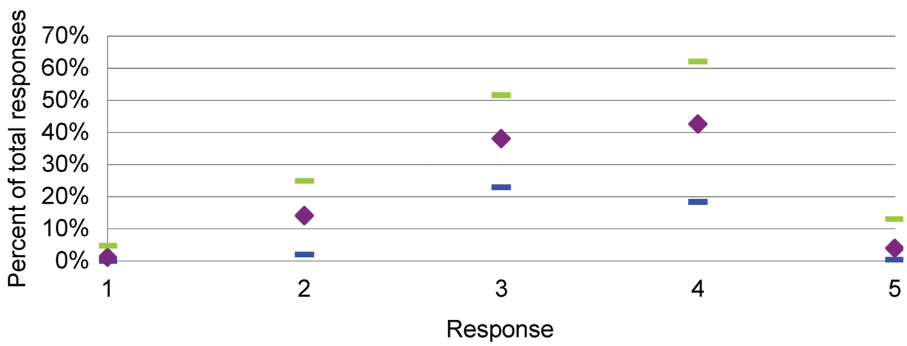

Figure 1

Distribution of item responses.

the numbers are meaningful; that is, such a scale has a natural zero value and is unique up to the choice of the scale unit. In other words, the scale is nonarbitrary.

We believe that many people tend to treat measurements as if they were made on a ratio scale unless explicitly warned to do otherwise. A certain kind of mental discipline and mathematical sophistication is needed to remember that one is using an ordinal scale, and we suspect that this discipline and sophistication is often lacking in those who use scoring methods in risk assessments. This can lead users to draw invalid inferences that may sometimes be harmful.

For example, likelihood may originally be assessed on a five-point verbal scale and later converted into a numerical representation for ease of processing. If "not very likely" is converted into a value of 2, and "very likely" is converted into a value of 4, then someone who treats this ordinal scale as if it were a ratio scale might infer that the likelihood of the latter category is exactly double the likelihood of the former. This inference may well be invalid, since ordinal scales involve an indeterminate distance metric.

"Range compression" is another problem that has been pointed out regarding the use of such scales [18]. Since ordinal scores tend to use scales with a small number of discrete values, there can be a significant loss of resolution in the assessment. This loss of resolution is more apparent when, as is done in some scoring models, known quantities are converted to an ordinal scale. For example, an additive scoring model that assesses the risk of IT projects may convert "planned duration" in months to a five-point ordinal scale such that less than three months is a "1," three to six months is a "2," and so on. One scoring method that has been proposed for ranking IT project portfolios converts return on investment to an ordinal scale so that each scale increment is an extremely wide range of values (e.g., 1%–299% is converted to a score of "1") [19].

If the original quantity were uniformly distributed among the discrete score values, then the additional error added by this rounding off (i.e., conversion of wide ranges of values to a few scores) would be half of the range of values in one ordinal increment. Note that this rounding error is exacerbated by two factors. First, many of the scales defined are not linear (a "1" for project duration may mean one to three months, while a "4" may mean one to two years). Second, responses in the purely subjective scales are often highly clustered. One of the authors collected the data for scoring responses from the scoring schemes used in five different organizations. All were used for IT project risk assessment, and all were based on five-point scales. In each case, 3–12 individuals had to assess up to 40 factors for each member in a list of separate IT risks or IT projects for a total of more than 2,000 individual item responses. The distribution of responses for all five methods is shown in **Figure 1** (the bars indicate the minimum, average, and maximum among the seven methods). As may be observed, just two of the values on the scale account for the majority of choices. In effect, the scale is used more like a two- or three-point scale, further reducing the resolution and increasing the rounding error. This is a cursory examination of a few real-world examples, but it does make the nature of this problem clear.

Another assumption is that functions based on simple additive or multiplicative operations are an adequate approximation of a function that might realistically represent risks. This assumption may have an analogy with the fanciful example of aircraft engineers attempting to accurately compute the range of an aircraft by making use of weighting and adding factors associated with wingspan and fuel tank size instead of using the relevant differential equations. Actuaries and risk analysts who use quantitative simulations routinely need to use higher-order nonlinear equations that are either derived from fundamental axioms of probability theory or empirically validated.

Invisible correlations

More elaborate quantitative models of risks must generally account for a variety of possible correlations among the modeled variables. A portfolio that has 50 investments is not well diversified if there are high correlations among all of them (e.g., if they all involve parts for oil rigs, and their business is a function of oil prices). Excluding such correlations from the model would cause a systematic underestimation of risks, and it would be considered a significant oversight for such models to exclude correlations. However, such issues are almost never considered in any of the ordinal scoring methods.

Furthermore, it is common in quantitative models of the risks of systems (e.g., nuclear power plants) to model "cascade failures" and "common mode failures." These are single failures that simultaneously cause several other failures and failures that cause a chain reaction sequence of other failures, respectively. As with correlations, leaving such relationships out of a model would be considered a significant shortcoming that, again, results in the systematic and dramatic underestimation of risks.

A multiplicative model, for example, will assess the likelihood and impact of several different risk events as if they were entirely independent. When the scores are aggregated, three different events may be considered to be "medium" or "low" risks. However, if the events are correlated and have an impact on one another, then three low-to-medium risks can together produce one very high risk.

Describing risk in terms of probabilities

We may summarize the four problems we identified in the previous section in terms of four common fallacies that have arisen about scoring methods. The first fallacy is the idea that without further aid or adjustment, subjective assessments of experts are an adequate assessment of risks. This is not true; the research in human judgment and decision making cited in the section "A partial review of literature on cognitive biases" shows that various known errors and biases are highly significant in the subjective assessment of risks, even for—in fact particularly for—experts.

The second fallacy is the idea that if a scale is very "rigorously" defined, then it will be used in a predictable consistent manner. This is not true either; the research cited in the section "Existing research regarding the variability of verbal labels" shows that the verbal scales are used in a highly inconsistent manner, even when users are given a great deal of descriptive detail about the definitions of the values.

The third fallacy is the idea that the exact scales and scoring calculations chosen are arbitrary and will not ultimately have as much bearing on the outcomes as the input of the experts. This is patently false; the arbitrary choice of whether a scale is a three- or five-point scale, together with the choice of how the scales are to be combined, can affect the results or produce nonsensical results. Furthermore, the arbitrary application of operations such as addition and multiplication to these ordinal scales adds even more error.

The fourth and final fallacy is the idea that correlations and other interactions among events are not necessary to model a rough approximation. This is false: In quantitative risk models, it is known that excluding correlations and relationships such as cascade failures or common mode failures can cause a model to greatly underestimate risks. There is no reason to believe this issue is somehow alleviated by using ordinal scales.

All of the previously identified scoring methods are associated with these fallacies. However, this is also true for the more "rigorous" approaches, such as the analytic hierarchy process (AHP), that mathematically attempt to analyze verbal preferences. Since the AHP can be used to estimate the coefficients of ordinal scales, proponents of AHP point out that such scales will then directly yield a cardinal product that is potentially accurate and relevant. However, this does not imply that AHP thereby avoids the problems we have identified with the use of ordinal scales in risk assessment, since these mathematical adjustments are only valid for converting ordinal utility measures to cardinal utility values. However, in risk assessment, subjective utility is only part of the concern; without objectively valid (i.e., well-calibrated) probability assessments and cost estimates, even the most finely tuned subjective utility values will be useless. The theory behind converting ordinal utility into cardinal utility (as developed by von Neumann and Morgenstern [20]) is quite a different topic than the degree to which the predictive power of subjectively chosen scores is empirically validated.

Likewise, the theoretical soundness of AHP (which has been disputed [21–27]) is a separate issue from the empirical description of whether the potential accuracy of AHP is ever realized in practice. Regardless of how much personal satisfaction that people may derive from using AHP, its objective value as a risk assessment tool depends on showing that experts using AHP outperform experts using their own unaided intuition in a controlled forecasting experiment. There is, however, no such evidence. Instead, the research concerning the validity of AHP, as with other scoring models, has focused on other measures, such as how well the method predicts the *stated preferences* of users, and not how well it forecasts real-world events, such as costs, project failures, and industrial accidents [28]. Furthermore, nothing in AHP solves the problems described in the section "Existing research regarding the variability of verbal labels." The "garbage in/garbage out" problem (i.e., erroneous input produces erroneous output) exists for AHP if eliciting probabilities and risks by scales (in this case, pairwise comparisons) is flawed.

Again, care must be taken not to confuse the perception of a benefit with actual improvements in the accuracy of assessments. Studies have already shown that gathering more information and interacting with other individuals before making a decision can improve subjective confidence in the decision without improving its objective quality; indeed, it can even decrease the objective quality [29, 30]. One empirical study of decision quality does, in fact, show that the use of a well-known AHP tool improved confidence without an improvement—or perhaps even a degradation—in decision quality [31].

To avoid all of these problems, risk assessment should use methods that meet the following three criteria. First and foremost, useful risk assessment methods must use explicit probabilities and magnitudes of losses expressed quantitatively instead of using surrogate verbal or ordinal scales. In other words, instead of stating that likelihood and impact are "high" and "medium," we would state that there is a "10% chance of a loss of inventory worth $2 million to $4.5 million." One objection to this approach is that the data may not be available. However, the previously cited research shows that experts can be trained to subjectively estimate the values. Note that ordinal scales, which also depend on subjective assessments, do not in any way alleviate the problem associated with a lack of data; on the contrary, they merely hide it. With quantitative probability estimates, however, uncertainty is explicitly incorporated. Furthermore, these estimates can be refined (given the existence of the extensive research about human performance in assessing subjective probabilities) and tracked against actual outcomes.

The second criterion that good risk assessment methods should satisfy is that they should use Monte Carlo simulations to explicitly model the relationships of components in a system and their correlations. In other words, instead of just adding or multiplying abstract factors, realistically descriptive functions are defined. If the risks of a major engineering project missing a deadline were being assessed, then the function may involve the rules of a detailed work plan with a range of durations defined for inclement weather or accidents. This exhibits no more of a "garbage in/garbage out" problem than any of the ordinal scoring methods, but it does allow for explicit known mathematical relationships (such as correlations that may exist among the components of a chemical processing plant, or the effect of demographic trends on a new mall) for which ordinal scales hardly provide even a gross approximation.

Finally, the third criterion for good risk assessment methods is that they should allow for research in human judgment and decision making to be applied in a corrective fashion. Many of the probabilities required for such quantitative models may still rely on the subjective judgments of experts, but methods can be employed to correct for biases, such as overconfidence. For example, the U.K. government explicitly acknowledges that optimism bias is a problem in planning and budgeting and has developed measures for how to deal with optimism bias in the government. The U.K. Department for Transport requires project planners to use so-called "optimism bias uplifts" for large transport projects to arrive at accurate budgets for planned ventures [32].

The first two criteria involve the use of methods already widely used in actuarial science, probability theory, statistics, and decision sciences, where risk is thought of as the probability of a loss of some magnitude. Industries associated with nuclear power, oil exploration, weather models, and actuarial science already make extensive use of such methods. One key advantage of such methods is that they produce results that can be compared with observation. While it may not be possible to validate whether an assessment of a "medium" probability was realistic (since it is ambiguous), numerical probabilities can at least be compared with observed frequencies. Furthermore, if probability functions are known for input variables, probability theory and simulations can be used to assess the risks of events that have not yet occurred (e.g., the 1-in-500-year events routinely assessed in studies of nuclear power safety).

We should make note of one clarification. In these quantitative solutions to risk assessments, sometimes probability and magnitude are multiplied together. This may, at first, seem like a multiplicative version of the ordinal scales, but in this case, neither dimension is ordinal. Probabilities are measured quantitatively as real number values between 0 and 1, inclusive, which is a ratio scale. Magnitude of loss, likewise, is a ratio scale, not an ordinal scale. Loss could be expressed as a monetary value or, as in the case of public health and safety, numbers of injuries or deaths.

Validated methods for eliciting subjective probabilities

It is a common belief among users of ordinal scoring methods that their situations are uniquely complex or lack the large volumes of data they imagine are at the disposal of analysts in other environments. They apparently conclude, then, that the use of subjective ordinal scores is their only recourse. However, methods exist that make the subjective assessments of probabilities practical while controlling for many of the sources of error mentioned previously. Three such methods appear to significantly improve the ability of experts to provide subjective estimates.

First, regular feedback can be provided to experts with incentives for improved performance. Weather forecasters using incentive systems based on the *Brier score* have developed a highly calibrated sense of subjective probabilities [33]. That is, when they say there is a 95% chance of precipitation, it rains 95% of the time. While weather forecasters have the advantage of empirical data and detailed weather models, the forecasts often require subjective assessments (and the original research in this area, in the late 1970s, predates the more advanced weather models used today).

Second, experts can be provided with so-called *calibration training*. There is evidence of some success in using various training techniques to calibrate experts to aid in assessing subjective odds. One of the authors has found that a half-day of training significantly reduces both overconfidence and underconfidence [34]. Other researchers have found that similar simple techniques have a significant effect on the calibration of probabilities [35]. Training methods typically involve repeated tests with trivia questions and feedback on results. Specific techniques involve treating the probability assessment in such a way that it is compared with a corresponding bet involving known chances. One such method is sometimes called the *equivalent urn* method [36]. If an expert believes an event has a 20% chance of occurrence, then a bet on that event should be considered equivalent to a bet based on whether a red marble will be randomly selected from an urn filled with marbles of which 20% are red and 80% are green.

Finally, experts can participate in prediction markets. When a number of people trade contingent assets whose value is tied to a future event such as "Barack Obama will win a second term of office," the current price at which these assets change hands provides a good indication of the probability of the relevant event. This approach seems to be reasonably well calibrated even if the market involves only "play" money and a scenario in which players compete for points [37].

As the previous citations illustrate, these methods of eliciting subjective probabilities are well-researched alternatives to ordinal scales. Such methods can and have been used in any environment where historical data are not considered sufficient for quantitative analysis and subjective estimates are required. These methods avoid the ambiguity error introduced by using verbal labels and can directly be used in probabilistic simulations.

However, even users of more sophisticated simulations might not be aware of the benefits of these methods. One of the authors has conducted a survey of users of common personal computer Excel** spreadsheet-based Monte Carlo simulation tools (such as *@Risk* and *Crystal Ball*) and finds that while subjective probability estimates are common in Monte Carlo simulations, calibration training is rarely used. From a total of 72 models created by 34 individual analysts, most involved a significant number of subjective probability estimates, but none used calibration training.

Currently, one of the authors has provided probability calibration training to a total of 190 individuals, all of whom completed training in which they were asked to provide a probability they will be correct in the answers for a series of true/false trivia questions. Of those, 166 also completed training in which subjects were asked to provide subjective 90% confidence intervals for a variety of trivia questions. This is similar to the methods used in previously cited research on overconfidence. When these individuals were initially tested, on average, they correctly answered only 47% of the answers within their stated 90% confidence intervals. After training and three to five rounds of additional tests, they on average correctly answered 80% of the questions within their 90% confidence intervals, and 66% of the subjects were within the statistically allowable error for perfect calibration. The calibration of binary probabilities showed that subjects who assessed probabilities produced an expected percentage correct to within 5% of their actual percentage correct (i.e., they would be correct about 80% of the time when they said they were 80% confident). Initially, they were overconfident by an average of 14% (i.e., when they said they were 90% confident they were correct only 76% of the time). These results are more encouraging than some of the attempts at calibration in the previous literature. Differences may be due to the fact that this training employed a combination of several calibration techniques, whereas in the previous literature subjects were typically tested one method at a time. In addition, unlike in some of the previous calibration research, the group of subjects for this training consisted entirely of professionals who are routinely confronted with real-world estimation problems.

Evidence of improvements on ordinal scores

There is little detailed analysis on the performance of ordinal scales since the results are rarely tracked (and this may very well explain why ordinal scales persist). Even most users of detailed Monte Carlo simulations rarely keep track of forecasting performance. However, we do know of at least one interesting example for which a fairly large sample of situations have been assessed with both a multiplicative ordinal "risk map" and quantitative models with Monte Carlo simulations, and which therefore permits a comparison of the two types of approaches to risk assessment. This example concerns NASA, which requires engineers and mission managers to develop a "5 $\times$ 5 risk matrix" for each new mission or space vehicle to identify key risks that may result in partial or complete mission failure [4]. At the same time, the NASA Headquarters Cost Analysis Division (CAD) has been conducting a Monte Carlo-based analysis of the same missions. The Monte Carlo simulations address cost and schedule risks as well as the risk of various mission failures. All of these missions have also been part of a historical analysis in which "best-fit" regression models compare indicators of mission complexity, budget constraints, and time constraints to actual mission failures.

The Monte Carlo simulation produced by the CAD uses data that are based, in part, on a logistic regression of previous mission failure rates with respect to a "complexity index" consisting of 37 factors [38]. Some subjective estimates of probabilities were also used when historical data were insufficient. The simulations outperformed the

5×5 matrices in predicting partial and complete mission failures. While several observed causes of mission failures were identified by the group that conducted the Monte Carlo simulations, not one actual mission failure was ever anticipated by the groups that created 5×5 matrices.

It is worth noting that the Monte Carlo models also consistently outperformed traditional deterministic estimates of costs and duration. Just over half of the simulations produced cost and schedule results that were within 10% of actual observations, whereas the deterministic methods consistently underestimated costs and schedule delays by 25%–55%.

The analysts of the CAD also note an important reaction by mission managers and engineers regarding the use of models that depend in part on quantitative historical data. There is a tendency to think of each mission as unique so that historical data cannot be used in a meaningful way to assess the risks of the new vehicle or mission. In our experience, we have also found this opinion to be not uncommon among managers in information technology, business operations, and a variety of other fields. However, the data clearly show that, in the case of NASA, the quantitative models that were based on historical data analysis and quantitative subjective estimates outperform the qualitative judgments of experts.

While previously published research comparing the performance of various risk analysis methods involving real-world problems is scant, what does exist seems to support the experience at NASA. One study in the oil exploration industry showed that the sophistication of the quantitative risk assessments correlated with various measures of financial performance of oil exploration companies [39]. Furthermore, the effect on financial performance was observed to occur *just after* the more advanced methods were adopted [40]. These researchers equated "more advanced" with the use of quantitative risk estimates, decision trees, Monte Carlo simulations, options theory, and portfolio theory.

Conclusion

In this overview, we have argued that scoring methods are poor tools for risk assessment. We first described a number of common scoring methods currently used to assess risk in a variety of different domains. We then identified and discussed four important problems associated with the use of these scoring methods. We concluded by arguing that risk assessment should describe risk in terms of mathematical probabilities.

The problem discussed in this paper is serious. The fact that simple scoring methods are easy to use, combined with the difficulty and time delay in tracking results with respect to reality, means that the proliferation of such methods may well be due entirely to their perceived benefits and yet have no objective value.

A practical alternative to scoring methods exists, namely, the use of quantitative probability estimates. Methods involving such probability estimates have been validated by observation and, at a minimum, control for many of the sources of error. We, therefore, conclude that caution should be exercised before using scoring methods to make critical risk assessments. The size and scope of many investment decisions and public health and safety decisions easily justify a more rigorous analysis than is permitted by the use of scoring methods involving ordinal scales. At a minimum, rigorous empirical tests of the predictive validity of a scoring method should be required before it is applied to such critical risk analysis problems.

References

1. J. D. Van Vactor, "Risk mitigation through a composite risk management process: The U.S. army risk assessment," *Org. Dev. J.*, vol. 27, no. 4, pp. 84–97, 2009.
2. United States Department of Health and Human Services, Draft Guidance on Allocating and Targeting Pandemic Influenza Vaccine, 2007. [Online]. Available: http://www.pandemicflu.gov/vaccine/prioritization.html
3. J. V. Turner, "Risk management on the space shuttle upgrades development program," in *Proc. Joint ESA-NASA Space-Flight Safety Conf.*, B. Battrick and C. Preyssi, Eds., 2002, European Space Agency, ESA SP-486.
4. D. Hubbard, *The Failure of Risk Management: Why It's Broken and How to Fix It*. Hoboken, NJ: Wiley, 2009, pp. 237–238.
5. *Risk Management Guide for Information Technology Systems, NIST 800-30*, Nat. Inst. Stand. Technol., Gaithersburg, MD, 2002. [Online]. Available: http://csrc.nist.gov/publications/nistpubs/800-30/sp800-30.pdf
6. *A Guide to the Project Management Body of Knowledge*, Project Manage. Inst., Newtown Square, PA, 2008.
7. D. Ariely, *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. London, U.K.: Harper Collins, 2008.
8. D. Kahneman, P. Slovic, and A. Tversky, *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge, U.K.: Cambridge Univ. Press, 1982.
9. D. Kahneman and A. Tversky, "Subjective probability: A judgement of representativeness," *Cogn. Psychol.*, vol. 3, no. 3, pp. 430–454, Jul. 1972.
10. D. Kahneman and A. Tversky, "On the psychology of prediction," *Psychol. Rev.*, vol. 80, no. 4, pp. 237–251, Jul. 1973.
11. E. Brunswick, "Representative design and probabilistic theory in a functional psychology," *Psychol. Rev.*, vol. 62, no. 3, pp. 193–217, May 1955.
12. N. Karelaia and R. M. Hogarth, "Determinants of linear judgement: A meta-analysis of lens studies," *Psychol. Bull.*, vol. 134, no. 3, pp. 404–426, May 2008.
13. D. G. Ullman and M. E. Doherty, "Two determinants of the diagnosis of hyperactivity: The child and the clinician," *Adv. Dev. Behav. Pediatr.*, vol. 5, pp. 201–211, 1984.
14. P. Hoffman, P. Slovic, and L. Rorer, "An analysis of variance model for an assignment of configural cue utilization in clinical judgment," *Psychol. Bull.*, vol. 69, no. 5, pp. 338–349, May 1968.
15. C. P. Bradley, "Can we avoid bias?" *Brit. Med. J.*, vol. 330, no. 7494, p. 784, Apr. 2005.
16. D. V. Budescu, S. Broomell, and H.-H. Por, "Improving communication of uncertainty in the reports of the intergovernmental panel on climate change," *Psychol. Sci.*, vol. 20, no. 3, pp. 299–308, Mar. 2009.

17. S. S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, no. 2684, pp. 677–680, Jun. 1946.
18. L. A. Cox, Jr., "What's wrong with risk matrices?" *Risk Anal.*, vol. 28, no. 2, pp. 497–512, Apr. 2008.
19. M. Parker, R. Benson, and H. Trainor, *Information Economics: Linking Business Performance to Information Technology*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
20. J. V. Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton Univ. Press, 1944.
21. J. S. Dyer, "Remarks on the analytic hierarchy process," *Manage. Sci.*, vol. 36, no. 3, pp. 249–258, Mar. 1990.
22. J. S. Dyer, "A clarification of 'Remarks on the analytic hierarchy process'," *Manage. Sci.*, vol. 36, no. 3, pp. 274–275, Mar. 1990.
23. R. D. Holder, "Some comment on the analytic hierarchy process," *J. Oper. Res. Soc.*, vol. 41, no. 11, pp. 1073–1076, Nov. 1990.
24. R. D. Holder, "Response to holder's comments on the analytic hierarchy process: Response to the response," *J. Oper. Res. Soc.*, vol. 42, no. 10, pp. 914–918, Oct. 1991.
25. S. Schenkerman, "Inducement of nonexistent order by the analytic hierarchy process," *Decis. Sci.*, vol. 28, no. 2, pp. 475–482, Apr. 1997.
26. J. Pérez, "Some comments on Saaty's AHP," *Manage. Sci.*, vol. 41, no. 6, pp. 1091–1095, Jun. 1995.
27. J. Pérez, J. Jimeno, and E. Mokotoff, "Another potential shortcoming of AHP," *TOP*, vol. 14, no. 1, pp. 99–111, Jun. 2006.
28. P. J. H. Schoemaker and C. C. Waid, "An experimental comparison of different approaches to determining weights in additive utility models," *Manage. Sci.*, vol. 28, no. 2, pp. 182–196, Feb. 1982.
29. C. Tsai, J. Klayman, and R. Hastie, "Effects of amount of information on judgment accuracy and confidence," *Org. Behavior Human Decis. Process.*, vol. 107, no. 2, pp. 97–105, Nov. 2008.
30. C. Heath and R. Gonzalez, "Interaction with others increases decision confidence but not decision quality: Evidence against information collection views of interactive decision making," *Org. Behavior Human Decis. Process.*, vol. 61, no. 3, pp. 305–326, Mar. 1995.
31. M. L. Williams, A. R. Dennis, A. Stam, and J. E. Aronson, "The impact of DSS use and information load on errors and decision quality," *Eur. J. Oper. Res.*, vol. 176, no. 1, pp. 468–481, Jan. 2007.
32. B. Flyvbjerg, *Procedures for Dealing With Optimism Bias in Transport Planning*. London, U.K.: Brit. Dept. Transp., 2004.
33. A. H. Murphy and R. L. Winkler, "Can weather forecasters formulate reliable forecasts of precipitation and temperature?" *Nat. Weather Dig.*, vol. 2, no. 2, pp. 2–9, 1977.
34. D. Hubbard, *How to Measure Anything: Finding the Value of Intangibles in Business*. Hoboken, NJ: Wiley, 2007.
35. S. Lichtenstein, B. Fischhoff, and L. D. Phillips, "Calibration of probabilities: The state of the art to 1980," in *Judgement Under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic, and A. Tversky, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1982, pp. 306–334.
36. J. M. Hampton, P. G. Moore, and H. Thomas, "Subjective probability and its measurement," *J. R. Stat. Soc., Ser. A (General)*, vol. 136, no. 1, pp. 21–42, 1973.
37. E. Servan-Schreiber, J. Wolfers, D. M. Pennock, and B. Galebach, "Prediction markets: Does money matter?" *Electron. Markets*, vol. 14, no. 3, pp. 243–251, Sep. 2004.
38. D. Bearden, C. Freaner, R. Bitten, and D. Emmons, "An assessment of the inherent optimism in early conceptual designs and its effect on cost and schedule growth," in *Proc. SSCAG/SCAF/EACE Joint Int. Conf.*, Noordwijk, The Netherlands, 2008.
39. G. S. Simpson, F. E. Lamb, J. H. Finch, and N. C. Dinnie, "The application of probabilistic and qualitative methods to asset management decision making," presented at the Proc. SPA Asia Pacific Conf. Integr. Modeling Asset Manage., Yokohama, Japan, 2000, Paper 59455.
40. W. Bailey, B. Couet, F. Lamb, G. Simpson, and P. Rose, "Taking a calculated risk," *Oilfield Rev.*, vol. 12, no. 3, pp. 20–35, 2000.

Received March 9, 2009; accepted for publication May 15, 2009

Douglas Hubbard *Hubbard Decision Research, Glen Ellyn, IL 60137 USA (dwhubbard@hubbardresearch.com).* Mr. Hubbard has an M.B.A. degree in management information systems from the University of South Dakota. He is the author of *How to Measure Anything: Finding the Value of Intangibles in Business* (Wiley, 2007) and *The Failure of Risk Management: Why It's Broken and How to Fix It* (Wiley, 2009). He has published articles in *Information Week, CIO Enterprise, Architecture Boston*, and *DBMS Magazine*. He was formerly with Coopers & Lybrand and has more than 20 years of experience in IT management consulting, including 15 years of experience specifically in teaching organizations to use his AIE (Applied Information Economics) method. His other professional experience includes managing large IT strategy and software projects in insurance, manufacturing, nuclear power, banking, and pharmaceuticals.

Dylan Evans *University College Cork, School of Medicine, Brookfield Health Sciences Campus, College Cork, Cork, Ireland (d.evans@ucc.ie).* Dr. Evans is Lecturer in Behavioral Science in the School of Medicine at University College Cork, Ireland. After receiving his Ph.D. degree in philosophy from the London School of Economics, he conducted postdoctoral research in philosophy at King's College London and in robotics at the University of Bath before moving to the University of the West of England where he was Senior Lecturer in Intelligent Autonomous Systems. In January 2008, he moved to University College Cork to become a Senior Research Scientist at the Department of Computer Science, where he conducted research on decision theory and risk management, before taking his current post.