Metadata of the chapter that will be visualized online

Chapter Title	Risk Intelligence	
Copyright Year	2011	
Copyright Holder	Springer Science+Business Media B.V.	
Corresponding Author	Family Name	**Evans**
	Particle	
	Given Name	**Dylan**
	Suffix	
	Division/Department	School of Medicine
	Organization/University	University College Cork
	Street	Brookfield Health Sciences Complex, College Road
	City	Cork
	Country	Ireland
	Phone	353(0)21420-5303
	Email	d.evans@ucc.ie
	Email	evansd66@googlemail.com

Abstract

Risk intelligence is the ability to estimate probabilities accurately. In this context, accuracy does not imply the existence of objective probabilities; on the contrary, risk intelligence presupposes a subjective interpretation of probability. Risk intelligence can be measured by calibration testing. This involves collecting many probability estimates of statements whose correct answer is known or will shortly be known to the experimenter, and plotting the proportion of correct answers against the subjective estimates. Between 1960 and 1980, psychologists measured the calibration of many specific groups, such as medics and weather forecasters, but did not gather extensive data on the calibration of the general public. This chapter presents new data from calibration tests of over 6,000 people of all ages and from a wide variety of countries. High risk intelligence is rare. Fifty years of research in the psychology of judgment and decision-making shows that most people are not very good at thinking clearly about risky choices. They often disregard probability entirely, and even when they do take probability into account, they make many errors when estimating it. However, there are some groups of people with unusually high levels of risk intelligence. Lessons can be drawn from these groups to develop new tools to enhance risk intelligence in others. First, such tools should accustom users to specifying probability estimates in numerical terms. Second, they should focus on a relatively narrow area of expertise, if possible. Thirdly, these tools should provide the user with prompt and well-defined feedback. Regular calibration testing might fulfill all three of these requirements, though training assessors by giving them feedback about their calibration has shown mixed results. More research is needed before we can reach a definitive verdict on the value of this method.

20 Risk Intelligence

Dylan Evans
University College Cork, Cork, Ireland

S. Roeser (ed.), *Handbook of Risk Theory*, DOI 10.1007/978-94-007-1433-5_20,

Abstract: Risk intelligence is the ability to estimate probabilities accurately. In this context, accuracy does not imply the existence of objective probabilities; on the contrary, risk intelligence presupposes a subjective interpretation of probability. Risk intelligence can be measured by calibration testing. This involves collecting many probability estimates of statements whose correct answer is known or will shortly be known to the experimenter, and plotting the proportion of correct answers against the subjective estimates. Between 1960 and 1980, psychologists measured the calibration of many specific groups, such as medics and weather forecasters, but did not gather extensive data on the calibration of the general public. This chapter presents new data from calibration tests of over 6,000 people of all ages and from a wide variety of countries. High risk intelligence is rare. Fifty years of research in the psychology of judgment and decision-making shows that most people are not very good at thinking clearly about risky choices. They often disregard probability entirely, and even when they do take probability into account, they make many errors when estimating it. However, there are some groups of people with unusually high levels of risk intelligence. Lessons can be drawn from these groups to develop new tools to enhance risk intelligence in others. First, such tools should accustom users to specifying probability estimates in numerical terms. Second, they should focus on a relatively narrow area of expertise, if possible. Thirdly, these tools should provide the user with prompt and well-defined feedback. Regular calibration testing might fulfill all three of these requirements, though training assessors by giving them feedback about their calibration has shown mixed results. More research is needed before we can reach a definitive verdict on the value of this method.

Introduction

Although the term "risk intelligence" has been gaining currency during the past few years, there is still no consensus as to what it means. According to David Apgar, the term denotes "an individual's or an organization's ability to weigh risks effectively," and involves "classifying, characterizing, calculating threats; perceiving relationships; learning quickly; storing, retrieving, and acting upon relevant information; communicating effectively; and adjusting to new circumstances" (Apgar 2006). According to Frederick Funston, coauthor of *Surviving and Thriving in Uncertainty: Creating the Risk Intelligent Enterprise* (Funston and Wagner 2010), risk intelligence is "the ability to effectively distinguish between two types of risks: the risks that must be avoided to survive by preventing loss or harm; and, the risks that must be taken to thrive by gaining competitive advantage," and involves the ability to "translate these insights into superior judgment and practical action to improve resilience to adversity and improve agility to seize opportunity" (Krell 2010). Funston is a principal at Deloitte & Touche LLP, and Deloitte seems keen for people to associate the phrase "risk intelligence" with the its brand, to judge by the series of research papers they have published on this topic and their release of an iPhone app which purports to let users create "risk intelligence map."

The trouble with both of these definitions is that they are rather vague, and encompass so many abilities as to be practically useless, and certainly immune to any kind of scientific measurement. I prefer a much more restricted definition: risk intelligence is the ability to estimate probabilities accurately (Evans, in press). Not only is this concept simpler to grasp and more precise than those suggested by Apgar and Funston, it is also more susceptible to measurement. Before explaining how to measure risk intelligence as I define it, however, I will first address some common objections I have encountered when explaining my definition. Au1

Objections to My Definition of Risk Intelligence

The most common objection to my proposed definition seems to be that it makes no reference to notions of harm, threat or danger, which some people consider to be central to the concept of risk. The observation is correct, but I regard this feature as a virtue of my approach rather than a fault. While it is true that the term "risk" is intimately associated in the vernacular with undesirable possibilities, those who study the subject largely agree that this restriction is somewhat arbitrary. Risk experts never tire of pointing out that an exclusive focus on "downside risk" tends to encourage a risk-averse attitude and to discourage an awareness of the potential rewards from exploiting risky opportunities. Indeed, Funston goes so far as to make this point central to his definition of risk intelligence, as we have already seen. My definition of risk intelligence avoids value judgments altogether by referring simply to "probabilities," regardless of whether such probabilities refer to outcomes that we find pleasant or not.

Another objection to my definition comes from those who are fond of the distinction between "risk" and "uncertainty" proposed by the American economist Frank Knight in 1921. In that year Knight published his influential book, *Risk, Uncertainty, and Profit*, in which he argued that:

> ► Uncertainty must be taken in a sense radically distinct from the familiar notion of Risk, from which it has never been properly separated. The term "risk," as loosely used in everyday speech and in economic discussion, really covers two things which, functionally at least, in their causal relations to the phenomena of economic organization, are categorically different. . . . The essential fact is that "risk" means in some cases a quantity susceptible of measurement, while at other times it is something distinctly not of this character; and there are far-reaching and crucial differences in the bearings of the phenomenon depending on which of the two is really present and operating. . . . It will appear that a measurable uncertainty, or "risk" proper, as we shall use the term, is so far different from an unmeasurable one that it is not in effect an uncertainty at all. "We . . . accordingly restrict the term 'uncertainty' to cases of the non-quantitive type." (Knight 1921)

This distinction has become so influential that economists now talk about "Knightian uncertainty" when referring to risks that are immeasurable or impossible to calculate, in contrast to risks that can be quantified. This way of explicating the distinction is misleading, however, since both risk and Knightian uncertainty can be measured and quantified. The putative distinction is really about the way in which the probabilities are calculated in each case.

The prototypical case of risk, in Knight's sense, is a casino game like roulette or blackjack. To work out what the odds of a given bet should be in these games, all you need to know is the rules of the game itself. You do not need to collect any data or observe how the game is actually played. You can just read the rule book in the comfort of your armchair and work it out with pen and paper (though a laptop would often help considerably).

This may be contrasted with, say, working out what the odds of a given bet should be in a horse race. In this case, it will not help much if you simply read the rule book. In addition to this, you also need to gather lots of data about the horses in the race, and the jockeys, and the racetrack, and the likely weather on the day of the race, and who knows what else. You can get this data in all sorts of ways – reading the "form" of the horses as published in newspapers, looking carefully at the horses with your own eyes, talking to tipsters, listening to the weather forecast, and so on. And then you need to crunch all this data and come up with an estimate of how likely it is that this horse will win this race.

One way to crunch the data is to use a computer. Another is to use your brain; first absorb the data by reading, watching, and listening, and then mull it over in your own head and come up with an estimate of how likely it is that the horse will win. Doing that well is what I call risk intelligence.

I'm sticking to my guns, and will continue to refer to this ability as *risk* intelligence (rather than, say, *uncertainty* intelligence) because I do not think the distinction proposed by Frank Knight holds water. For one thing, the mere fact that the odds in a racetrack world cannot be worked out from first principles does not mean they cannot be measured or quantified. On the contrary, when gamblers or bookmakers estimate the chances of a horse winning a race or a team winning a basketball match, what else are they doing if not quantifying uncertainty?

More importantly, pure casino worlds do not exist – except in the pages of economics textbooks. In this respect, I agree with maverick trader Nassim Nicholas Taleb, who wrote in his bestselling book, *The Black Swan*, that:

> ► In real life you do not know the odds; you need to discover them, and the sources of uncertainty are not defined. Economists, who do not consider what was discovered by noneconomists worthwhile, draw an artificial distinction between Knightian risks (which you can compute) and Knightian uncertainty (which you cannot compute), after one Frank Knight, who rediscovered the notion of unknown uncertainty and did a lot of thinking but perhaps never took risks, or perhaps lived in the vicinity of a casino. Had he taken financial or economic risk he would have realized that these "computable" risks are largely absent from real life! They are laboratory contraptions!" (Taleb 2007)

When Taleb states that computable risks are "laboratory contraptions," he means that casino worlds are artificial entities which have to be deliberately manufactured under sterile conditions, like an unstable element that only exists for a few brief moments in a physics lab. It took thousands of years for the irregular-shaped knucklebones used in ancient Rome, and the Vibhīdaka nuts used in ancient India, to evolve into the precision dice used in modern casinos, with their pips drilled and then filled flush with a paint of the same density as the acetate, such that the six numbers are equally probable. It takes even greater engineering prowess to produce a fair roulette wheel. The manufacturers of roulette wheels perform elaborate tests to ensure that the numbers generated are truly random, and even then the wheels still have flaws, allowing some cunning players to make a fortune with a "biased wheel attack." In 1873, for example, a mechanic from Lancashire called Joseph Jagger identified a biased wheel in Monte Carlo and won the equivalent of $70,000 in 1 day.

Taleb tells a lovely story to illustrate the unreality of casino worlds. A casino in Las Vegas thought it had all the bases covered in risk management. It was sufficiently diversified across the various tables to not have to worry about taking a hit from lucky gamblers. It had a state-of-the-art surveillance system to catch cheats. But the four largest risks faced by the casino in the past few years lay completely outside its risk management framework. For example, it lost around $100 million when an irreplaceable performer in the main show was maimed by his tiger. In other words, even casinos are not pure casino worlds.

A third objection to my proposed definition is that it makes no reference to the concept of risk appetite. Again, I think this is an advantage rather than a defect. Risk intelligence is a cognitive capacity, a purely intellectual ability to estimate probabilities accurately. It involves gauging the extent of one's knowledge on a given topic, and can be objectively assessed to distinguish between those with high risk intelligence and those with low RQ. Risk *appetite*, on the other hand, is an emotional trait. It has to do with preferences. Some people enjoy taking on risk, and other people avoid risk like the plague; some people are willing to expose themselves to danger, while others prefer to shield themselves from as many losses as possible. Unlike risk intelligence, there's no right or wrong about risk appetite; it's just a matter of taste.

Measuring Risk Intelligence

How do we judge the accuracy of probability estimates? One way is to compare subjective probability estimates to objective statistics. For example, one can ask people to estimate the probability of death from various causes for some particular demographic group, and compare these estimates to the mortality data. This method is restricted, of course, to subject areas for which data are readily available.

Another way to measure a person's ability to provide accurate probability estimates is calibration testing (Lichtenstein et al. 1982). This involves collecting many probability estimates about statements whose correct answer is known or will shortly be known to the experimenter, and plotting the proportion of correct answers against the subjective estimates. For example, suppose that every day you estimate the probability that it will rain in your neighborhood the following day, and then you note whether or not it did, in fact, rain on each day. To simplify things a little, let us assume that you can only choose from a discrete set of probability values, such as 0, 0.1, 0.2, etc. Over the course of a year, you collect 365 estimates, for each of which you have also indicated whether it did, in fact, rain or not. Suppose that you estimated the chance of rain as 0 on 15 days. If you are well calibrated, it should have rained on none of those days. Again, if there were 20 days which you assigned a 0.1 probability of rainfall, it will have rained on 2 of those days if you are well calibrated. Perfect calibration, in other words, would be shown by all points falling on the identity line.

The nice thing about this method is that, unlike comparing subjective estimates to objective statistics, it does not take objective probabilities to be conceptually prior. Throughout this chapter, I use numerical probabilities to express degrees of belief – or, to put it another way, to quantify subjective uncertainty. For the sake of fairness, however, I should point out that this is a minority view. A rival school of thought holds that numerical probabilities refer to objective facts about the world, namely, long run frequencies. According to this view, the statement that "there is a 50% chance of this coin landing on heads" does not have anything to do with anyone's beliefs; rather, it means that, in the long run, the coin will land heads up on half of all the times it is tossed.

These two schools of thought are happy to use the same mathematical tools; the probability calculus is uncontroversial. They differ only in their interpretation of what the math means. The subjectivist school is the older one; when Jacob Bernoulli first showed how any probability could be represented as a number, it was degrees of belief that he had in mind. During the twentieth century, however, the frequentist approach became more popular, and this is now the dominant view. For reasons I do not have time to go into here, I think the frequentist view is fundamentally flawed.

According to the subjectivist view, there is no such thing as a "true" probability, in the sense of some objective fact existing out there in the world; probabilities are just numerical expressions of our subjective degree of belief. I suppose a subjectivist could say that a probability estimate is "true" when it accurately expresses the strength of one's conviction – indeed, that is what risk intelligence is all about – but that is a far cry from the frequentist view of probabilities as facts.

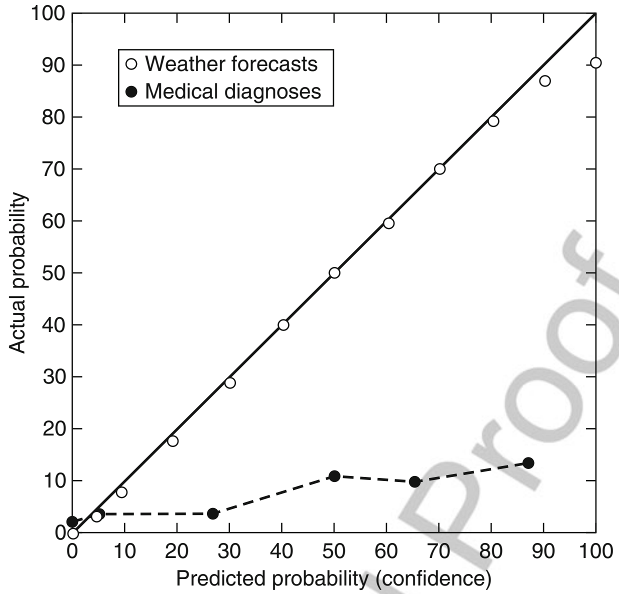

Fig. 20.1
Calibration curves for US weather forecasters and for doctors. (Data is from Murphy and Winkler 1977 and Christensen-Szalanski and Bushyhead 1981)

Calibration testing might be seen as incorporating something of the frequentist approach, in the sense that they plot probability estimates (subjective probabilities) against the proportion of correct predictions (an objective measure). But the proportion of correct predictions is not the same thing an "objective probability." So while I sometimes talk loosely of "making accurate probability estimates," strictly speaking this phrase is incoherent. The accuracy of an estimate can only be measured by comparing it to some objective fact, and such facts do not exist in the case of probabilities. This is why experts who study risk intelligence usually prefer to speak of "well-calibrated" probability estimates rather than of accurate ones.

Nobody is perfectly calibrated, but as you can see from *Fig. 20.1*, US weather forecasters are pretty close. However, as the same figure also shows, doctors are very badly calibrated.

The doctors whose data is shown in *Fig. 20.1* were asked estimate the probability that real patients had pneumonia after taking a medical history and completing a physical examination. When these doctors estimated that there was about a 5% that a patient had pneumonia, they were about right. But only about 15% of the patients to whom these doctors assigned a 90% chance of pneumonia turned out to have the disease. In other words, these doctors had much more faith in the accuracy of their diagnoses than was justified by the evidence.

The Distribution of Risk Intelligence in the General Population

Between 1960 and 1980, psychologists measured the calibration of many specific groups, such as medics (Christensen-Szalanski and Bushyhead 1981) and weather forecasters (Murphy and

Winkler 1977), but did not gather extensive data on the calibration of the general public (the data from these two groups is shown in ➲ *Fig. 20.1*, above). In their survey of the research up to 1980, Lichtenstein and colleagues report studies of hospital patients, psychologists, military personnel, engineering students, and other groups, but no large cross-sectional studies of the general population (Lichtenstein et al. 1982). One reason for this was no doubt because the testing was done with pen and paper, which made data collection and processing a time-consuming process. It appears that interest in calibration testing began to decline after 1980, and has not progressed much since then. This area of research is ripe for revival, especially now that the Internet allows testing and data collection to be automated.

In this section I present data from calibration tests of over 6,000 people of all ages and from a wide variety of countries. I was able to collect such a large amount of data by using an online calibration test rather than a pencil-and-paper version.

In December 2009 my co-investigator Benjamin Jakobus and I created an online calibration test (http://www.projectionpoint.com), and promoted the site through press releases, media interviews, blogs, and Internet discussion forums. The test consisted of 50 statements (see Appendix), below each of which were 11 buttons indicating percentage values ranging from 0 to 100 in increments of ten. Visitors to the site were instructed to indicate how likely they thought it was that each statement was true according to the following rules:

- If you are absolutely sure that a statement is true, you should click on the button marked 100%.
- If you are completely convinced that a statement is false, you should click on the button marked 0%.
- If you have no idea at all whether it is true or false, you should click on the button marked 50%.
- If you are fairly sure that it is true, but you are not completely sure, you should click on 60%, or 70%, or 80%, or 90%, depending on how sure you are.
- If you are fairly sure that it is false, but you are not completely sure, you should click on 40%, or 30%, or 20%, or 10%, depending on how sure you are.

After participants had answered all 50 questions in the test, they were asked if they would like to take part in our study. If they declined, they were given their test results, and then their data were deleted from the server. If they agreed, they were asked to specify the following demographic details: gender, nationality, age, highest level of academic education, and profession. They were then given their test results.

The test results were calculated as follows. First, all the times that the participant assigned a likelihood of 0% to a statement were counted, and then we counted how many of those statements were actually true. We proceeded in the same way for each of the other likelihoods and plotted each data point on a graph with the probability estimates on the x-axis and the proportion of correct answers in each category on the y-axis. We always plotted the points for the categories of 0% and 100% (if a participant never used the 100% category, we plotted that as 100% correct), but we only plotted points for the other categories if the participant had used them at least three times. We then connected the points by a continuous line. This line is henceforth referred to as the participant's calibration curve.

As already noted, a perfect calibration curve would lie on the identity line $x = y$. The further away from that diagonal line the curve lies, the poorer the calibration is. We created a simple index of a participant's calibration by calculating the area between their calibration curve and the identity line and scaling the result to a number between 0 and 100, where 100 = perfect calibration (i.e., this number is inversely proportional to the size of the area between the calibration curve and the identity line). Henceforth we refer to this number as the "RQ score."

We chose to use this measure, rather than using the better-known Brier score (Brier 1950) for three reasons. Firstly, our approach is much easier to understand for a lay audience. Secondly, we find some of the statistical properties of the Brier score to be unsatisfactory. Suppose a person chooses p_i for n_i questions, of which k_i are correct. The Brier score would yield $(p_i - k_i/n_i)^2$, which implies that it is irrelevant whether $k_i = 1$ and $n_i = 2$ or $k_i = 1{,}000$ and $n_i = 2{,}000$. We object to this because we believe that the scores should be weighted by the number of answers. We could remedy this by adjusting the Brier score as follows: $n_i{}^{*}(p_i - k_i/n_i)^2$ (Aaron Brown, personal communication), but again we prefer a simpler approach. Finally, the Brier score is a composite measure of calibration, resolution, and knowledge (Murphy 1973), whereas we wish to measure only calibration.

Some people realize fairly quickly that there is an easy way to game this test. If a participant always selects the 50% category – and if the test contains equal numbers of true and false statements – they will obtain an RQ score of 100. The high score generated by this strategy does not reflect good calibration, however, so to remedy this we created a second indicator which we call the "K factor" (for Keynes and Keats). To calculate the K factor, each time that a participant uses the categories 10%, 20%, 30%, 40%, 60%, 70%, 80%, or 90%, they get one point. When they use 0%, 50%, or 100% they get zero. The maximum K factor is therefore 50 for a fifty-question test. The K factor gives an indication of how reliable a participant's RQ score is as an indicator of their calibration.

Between December 10, 2009, and February 6, 2010, a total of 21,910 people visited the web site, of whom 10,187 took the online calibration test and gave us permission to use their data in our research. We removed all participants with an RQ score of 0 ($N = 30$) on the grounds that this was probably due to error. We then removed all those with a K score of less than 10 ($N = 3{,}295$) on the grounds that in such cases the RQ score would not be a good indicator of calibration. We then removed all those who did not specify their gender ($N = 154$) or their educational achievement ($N = 10$). After these adjustments, a total of 6,698 participants remained in our sample. ➲ *Table 20.1* shows the composition of this sample by gender and educational achievement.

The participants' ages ranged from under 10 to over 80, though most participants were aged either 21–30 ($N = 2{,}242$) or 31–40 ($N = 1{,}852$). Participants came from every continent,

Table 20.1
Composition of the sample by gender and educational achievement

Education	Men	Women	Total
Primary or less	37	10	47
Secondary	974	223	1,197
First degree	2,617	675	3,292
Masters	1,113	338	1,451
Ph.D.	567	144	711
Total	5,308	1,390	6,698

with over 20 nationalities, with the most well-represented countries being (in order): the USA ($N = 2{,}633$), the UK ($N = 1{,}118$), Ireland ($N = 1{,}023$), Canada ($N = 402$), Australia ($N = 343$), and Germany ($N = 188$).

The mean RRQ score for the sample of 6,698 was 65.02. *Figure 20.2* shows the distribution of RQ scores in this sample.

As shown in *Table 20.2*, the mean RQ score of the men in our sample was significantly higher than the mean RQ score of the women (two-tailed t-test for independent samples, not assuming equal variance, $p < .0001$; similar results were found with a Mann–Whitney rank sum test). When analyzed according to educational achievement, however, the difference between the mean RQ scores of the men and women in our sample is only significant in those whose highest level of educational achievement is a first degree or masters. As the graph in *Fig. 20.3* makes clearer, education seems to make little or no difference to the calibration of women until they achieve a Ph.D., while every increase in educational achievement seems to improve calibration in men. It is this "education effect" that explains the higher mean RQ score of men in our sample, since it contains a high proportion of people whose highest level of educational achievement is a first degree or masters (70.7%). It is only at this level of achievement that the education effect produces a clear difference in calibration between men and women. Something about university education seems to boost calibration in men to levels significantly higher than those in women, and this gap only closes when people have attained the highest level of educational achievement – the Ph.D.

Previous studies have not found differences in calibration between men and women (Lichtenstein and Fischhoff 1981) or between people with different levels of education (Lichtenstein and Fischhoff 1977). However, this may be due to the fact that most studies have typically involved fewer than 200 participants, and participants have generally been required to provide a smaller number of probability estimates – both of which have severely

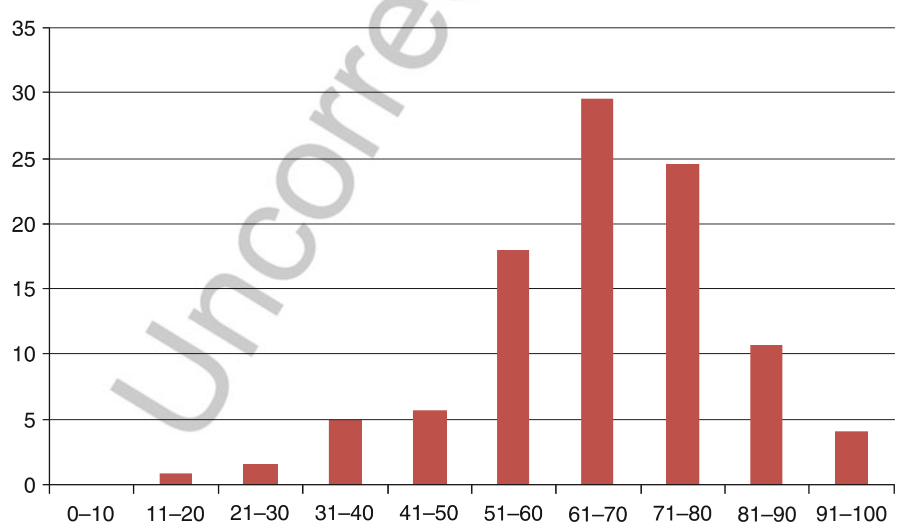

Fig. 20.2

The distribution of RQ scores in the sample of 6,698 people reported here. The *x*-axis shows the RQ score, and the *y*-axis the percentage of the sample in each category

Comp. by: DMuthuKumar Stage: Proof Chapter No.: 20 Title Name: HbRT
Page Number: 0 Date:2/6/11 Time:14:15:58

Table 20.2
Mean RQ scores by gender and educational achievement (*=significant at $p < .05$)

	Men		Women			95% CI of the difference		
Education	Mean RQ	(Stdev)	Mean RQ	(Stdev)	Standard error difference	Lower	Upper	*p*
Primary or less	56	18.55	60	13.71	5.30	−15.12	7.07	0.427
Secondary	63	16.25	62	16.58	1.21	−1.26	3.49	0.365
First degree	65	15.20	62	16.03	0.66	1.87	4.47	<.0001*
Masters	67	13.50	62	16.67	0.99	3.32	7.22	<.0001*
Ph.D.	69	14.21	66	15.30	1.41	0.05	5.60	0.046*
ALL	66	15.11	63	16.22	0.48	2.21	4.10	<.0001*

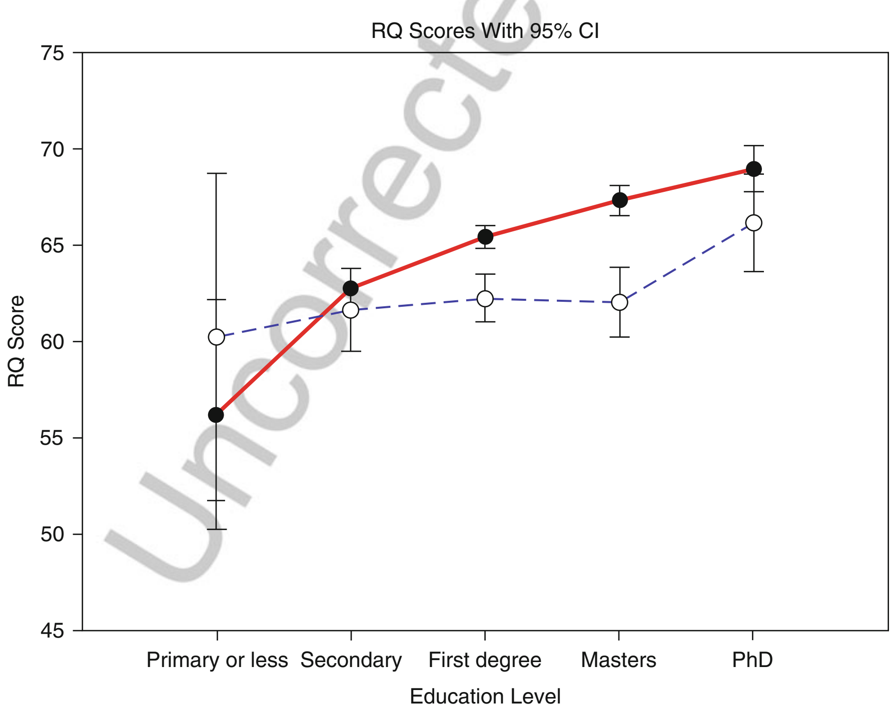

Fig. 20.3
Effect of education on calibration in men and women (*dashed line*=women)

limited the capacity of previous research to detect individual differences with a high enough degree of statistical significance. Our sample of 6,698 people, in which each participant provided 50 probability estimates, provides a dataset which is an order of magnitude larger than any previous study, and this may have permitted us to detect patterns which were previously invisible.

The Dunning–Kruger Effect

Calibration tests measure the extent to which one is able to gauge how much one knows, not knowledge per se. The extent of one's knowledge should, therefore, make no difference to one's score on a calibration test, and empirical research seems to bear this out (Lichtenstein and Fischhoff 1977). If education improves calibration in men, as it seems to, this effect must therefore be due to something other than the greater knowledge that greater education typically bestows. A possible candidate for this other factor is enhanced metacognition. Metacognition refers to "the ability to know how well one is performing, when one is likely to be accurate in judgment, and when one is likely to be in error" (Kruger and Dunning 1999). Kruger and Dunning (1999) have argued that the skills that engender competence in a domain are often the very same skills necessary to *evaluate* competence – including one's own competence – in that domain. Education may therefore make people not only more knowledgeable, but also more aware of the limits of their knowledge. Contrary to what Lichtenstein and Fischhoff (1977) concluded, then, it does appear that "those who know more also know more about how much they know," but as ➋ *Fig. 20.3* shows, this dictum seems to apply more to men than to women. Conversely, as Charles Darwin noted, "ignorance more frequently begets confidence than does knowledge" (Darwin 1871).

If education has this effect in men, why does it not also have this effect in women? I confess that I am at a loss to explain this apparent sex difference. In the rest of this section, I will limit myself to some general remarks about the Dunning–Kruger effect and risk intelligence in both sexes.

Like the Roman god Janus, risk intelligence has two faces looking in opposite directions. One looks outward at the external world, and attempts to gather objective data that will throw light on the matter at hand. The other looks inward and attempts to assess how much relevant knowledge one really has. Good probability estimates require that both faces see clearly – in other words, risk intelligence requires both objective and subjective knowledge (cognition and metacognition).

To change the metaphor, picture your mind as a light bulb shining in an otherwise darkened room. Some nearby objects are fully illuminated; you can see them in every detail, present and identifiable. These are the things you know very well – the names of your friends, what you had for breakfast this morning, how many sides a triangle has, and so on. The objects on the other side of the room are completely shrouded in darkness. These are the things about which you know nothing – the five thousandth digit of pi, the composition of dark matter, or King Nebuchadnezzar's favorite color. Between the light and the darkness, however, lies a gray area in which the level of illumination gradually shades away. In this twilight zone, the objects are not fully illuminated, but neither are they completely invisible. You know something about these things, but your knowledge is patchy and incomplete – the law of the land (unless you are a lawyer), the evidence for climate change (unless you are a climatologist), and the causes of the credit crunch (even economists are still arguing about this). Risk intelligence involves gauging exactly how illuminated the objects in this twilight zone really are.

Nothing in our education system or our culture prepares us to operate in the twilight zone. If we are cautious, we relegate everything beyond the zone of complete illumination to complete obscurity, not daring to venture an opinion on things of which we do, in fact, have some inkling. If we are overconfident, we do the opposite, expressing views about things in the twilight zone with more conviction than is justified. It is hard to steer between these two extremes; daring to speculate, but with prudence. Yet that is what risk intelligence is all about.

This is why calibration tests which require users to estimate the likelihood of general knowledge statements is a perfectly acceptable way to measure risk intelligence. When putting a probability value on these statements, one is required to weigh up all the relevant evidence that one possesses and gauge one's true level of uncertainty on the matter. There is no reason to restrict the content of these statements to threats, dangers, and other concepts generally associated with the vernacular view of risk.

Nor should it matter if a calibration test is dominated by questions that refer to a particular area of knowledge or a particular part of the world. A number of people in the study reported in section "➔ The Distribution of Risk Intelligence in the General Population" complained that they thought the test had a "US bias." However, since the test measures how well one is able to gauge how much one knows, rather than knowledge per se, any such bias is irrelevant.

Of course, if a calibration test contained too many questions about which a user knew absolutely nothing, and which the user should therefore use the 50% category many times, the result would not be an accurate measure of the person's level of risk intelligence. To provide a good measure of risk intelligence, it is necessary to gather probability estimates across the whole range of possible values from 0% to 100%. This is what the K-factor described in section "➔ The Distribution of Risk Intelligence in the General Population" was intended to capture.

Why Is Risk Intelligence Important?

The doctors whose data is graphed in ➔ *Fig. 20.1* were extremely overconfident in their diagnoses. When a patient had a 15% chance of having pneumonia, they would give them a 90% chance. That meant they were likely to recommend more tests than were strictly necessary, prescribe more treatments than were warranted, and cause their patients needless worry.

It is not always true to say "better safe than sorry," either. Some tests are invasive and painful, and there are many cases where the treatment prescribed for an ailment that is not present can be harmful to the patient. It is always better to make a choice on the basis of accurate information than the basis of error, no matter what the context. And the data clearly shows that these diagnoses are full of error.

Doctors are not the only professionals who require good risk intelligence. Finance professionals are also required to estimate probabilities as a regular part of their work. Yet as the recent financial crisis has shown us, bankers and those who work at credit rating agencies can make basic errors when assessing the likelihood of certain outcomes. What would happen if trading desks implemented some kind of regular calibration testing, or if a requirement for calibration training were incorporated into new financial regulations? a? The Basel Committee on Banking Supervision, which formulates broad supervisory standards for financial institutions around the world, is currently working on a new update to the Basel Accords. Basel II required numerical assessments of both the probability of default and the expected loss given default, and it specifically forbade relying on rating agency assessments to estimate these, but these estimates were typically produced by computer models. What if future rounds of the Basel Accords were to include some provision for testing risk intelligence? The head of a prop desk in a bank, for example, could ask each trader to estimate the probability for each trade that it will make a profit and keep track of each trader's calibration.

The commitment, detection, and investigation of crime also offer many opportunities for the exercise of risk intelligence. For example, when police officers question suspects, they must judge whether the responses they receive are truthful or not. It is rarely the case that such judgments are clear cut; more typically, the officer has some index of suspicion which lies somewhere in between complete confidence and absolute distrust. In other words, the question of whether a suspect is lying usually demands a probability estimate rather than a simple yes or no. As with all probability estimates, their accuracy may be measured by means of a calibration test when the true answers are known. Psychologists have carried out many such tests, and the results all point to the same conclusion; police officers and other professional investigators are all massively overconfident about the ability to discern lies. Although they are convinced they can spot deception, their real ability to sift fact from fiction is scarcely better than flipping a coin. One reason for this is that most police officers look at the wrong signals; shifty eyes, for example, are not a good signal of deception, and yet it is one that many investigators rely on. This has serious consequences for the criminal justice system, and law enforcement officials should therefore be trained in risk intelligence if we are to reduce miscarriages of justice.

High levels of risk intelligence will also be required among the general population if we are to deal effectively with any of the big challenges that humanity faces in the twenty-first century. Climate change is a case in point. Nobody knows precisely how increasing levels of greenhouse gasses in the atmosphere will affect the climate in different regions around the globe. The Intergovernmental Panel on Climate Change (IPCC) does not make definite predictions; instead, it sets out a variety of possible scenarios and attaches different probabilities to them to indicate the level of uncertainty associated with each one. Knowing how to make sense of this information is crucial if we are to allocate resources sensibly to the various alternative solutions, from carbon trading schemes to the development of alternative energy sources, or even planetary-scale geo-engineering. How can citizens make informed decisions about such matters if they are not equipped to think clearly about risk and uncertainty? Too often, the public figures who take opposite views about climate change make exaggerated claims which convey greater certainty than is warranted by the evidence. Critics dismiss the claims of the IPCC out of hand, while believers in climate change proselytize with equal dogmatism. Both kinds of exaggeration seriously hamper informed debate; the latter kind also terrifies kids. One survey of 500 American preteens found that one in three children between the ages of 6 and 11 feared that the earth would not exist when they reached adulthood because of global warming and other environmental threats (Björn 2009). We see the same pattern in the UK, where a survey showed that half of young children aged between seven and eleven are anxious about the effects of global warming, often losing sleep because of their concern. Without the tools to understand the uncertainty surrounding the future of our climate, we are left with a choice between two equally stupid alternatives – ignorant bliss or fearful paralysis.

Methods for Increasing Risk Intelligence

Leading medical schools around the world are beginning to wake up to the problem of low risk intelligence among doctors. Something called "confidence-based assessment" is increasingly being used in these schools. In this form of assessment, students must not only give the right answer, but also assess the confidence with which they give each answer. If a student gives the wrong answer confidently, that receives the worst possible grade; if they give the wrong answer but are not confident, then they get a better grade; giving the right answer but without confidence is OK, but not ideal, as in reality it could end up with them wasting time having to consult others; and the best answer is that which is correct and made with confidence. This form of assessment is intended to help students know when to consult others (or text books etc) and when to act independently.

All well and good, but how do you prepare students for this kind of test? Nothing in our current repertoire of educational tools and methods seems well designed to equip someone with the skills to be confident when justified but doubtful when necessary. In medical schools in particular, doubt has often been perceived as a sign of weakness.

The same could be said of the financial sector, where similar macho attitudes played no small part in stoking the bubble that burst in late 2007. What if the bankers who were making all those dubious loans in the preceding decade had undergone regular calibration testing, of the sort described here? It is an interesting thought.

The fact that weather forecasters are so much better calibrated than doctors suggests that one's level of risk intelligence is not relatively fixed like IQ, but susceptible to improvement given the right conditions. Sarah Lichtenstein, an expert in the field of calibration testing, speculates that several of these conditions favor the weather forecasters (Lichtenstein et al. 1982). First, they have been expressing their forecasts in terms of numerical probability estimates for many years; since 1965, US National Weather forecasters have been required to say not just whether or not it will rain the next day, but how likely they think this is in actual percentage terms. They have got used to putting numbers on such things, and as a result are better at it. Doctors, on the other hand, are under no such obligations. They remain free to be as vague as they like.

Second, the task for weather forecasters is repetitive. The question to be answered ("Will it rain?") is always the same. Doctors, however, must consider all sorts of different questions every day: "Does he have a broken rib?" "Is this growth malignant?" "How will she respond to a different type of antidepressant?"

Finally, the feedback for weather forecasters is well defined and promptly received. This is not always true for doctors. Patients may not come back, or may be referred elsewhere. Diagnoses may remain uncertain. Most theories of learning emphasize the need for rapid feedback; the longer the delay between an action (or, in this case, a prediction) and a corrective signal, the lower the chance that the later information will enable the recipient to profit from it.

These speculations could assist the development of tools to enhance risk intelligence. First, such tools should accustom users to specifying probability estimates in numerical terms. Second, they should focus on a relatively narrow area of expertise, if possible. Thirdly, these tools should provide the user with prompt and well-defined feedback. Regular calibration testing might fulfill all three of these requirements, though training assessors by giving them feedback about their calibration has shown mixed results. It should be pointed out however, that only a few studies have been carried out in this area, and they are now several decades old. More research is needed before we can reach a definitive verdict on the value of this method.

Another approach to improving calibration involves requiring people to think of reasons why they might be wrong. In one study, subjects took two calibration tests similar to the one described in section "The Distribution of Risk Intelligence in the General Population." In the second test, one group was asked to write down a reason supporting each of their answers, another group was asked to write down a reason contradicting each answer, and a third group wrote down two reasons, one supporting and one contradicting. Only the group asked to write down contradicting reasons showed improved calibration in the second test. This suggests that one partial remedy for overconfidence is to search for reasons why one might be wrong (Koriat et al. 1980).

Further Research

As already noted, it appears that interest in calibration testing began to decline after 1980, and research in this area has not progressed much since then. I believe this area is ripe for revival, especially now that the Internet allows testing and data collection to be automated. The study reported in section "The Distribution of Risk Intelligence in the General Population" is ongoing, and a further 25,000 people have taken the online calibration test between the date when the dataset reported there was collected and the present time of writing (April 2011), bringing the total sample size to over 35,000 participants so far. Analysis of this growing dataset is ongoing, and further features are being added to the calibration test, including the measurement of time taken to complete the task.

In a study of horse handicappers, Steven Ceci and Jeffrey Liker found that handicapping expertise had zero correlation with IQ (Ceci and Liker 1986). IQ is the best single measure of intelligence that psychologists have, because it correlates with so many cognitive capacities. Indeed, it is this very correlation that underpins the concept of "general intelligence." The discovery that expertise in handicapping does not correlate at all with IQ means that, whatever cognitive capacities are involved in estimating the odds of a horse winning a race may be, they are not a part of general intelligence. Or, to put it the other way round, IQ is unrelated to some real-world forms of cognitive complexity that are clear-cut cases of intelligence.

Of course, not everyone is comfortable with blanket terms like "general intelligence." One of its most high-profile critics, the psychologist Howard Gardner, prefers to conceive of intelligence as consisting of multiple, special-purpose skill sets (Gardner 1983). It was Ceci and Liker's paper that first led me to wonder if the ability to estimate probabilities accurately, and make wise decisions under uncertainty, might constitute a special kind of intelligence to be added to Gardner's list.

Gardner identifies eight different kinds of intelligence: bodily-kinesthetic, interpersonal, verbal-linguistic, logical-mathematical, naturalistic, intrapersonal, visual-spatial, and musical. None of these involves an ability to estimate probabilities accurately, and yet the study of Ceci and Liker shows that this is a complex cognitive skill that some people are very good at, and this suggests that it could constitute a ninth kind of intelligence that Gardner had not considered.

To clinch this argument, I would have to show that risk intelligence is typically implemented in the brain by a specific set of neural pathways, for Gardner's framework restricts intelligences to cognitive capacities that can be localized neurologically. Linguistic intelligence, for example, is rooted in certain structures in the brain's left hemisphere. I believe that recent neuroscience imaging studies would also support a distinct neural architecture for risk assessment, but further research is needed to confirm this.

High risk intelligence is rare. Fifty years of research in the psychology of judgment and decision-making shows that most people are not very good at thinking clearly about risky choices. They often disregard probability entirely, and even when they do take probability into account, they make many errors when estimating it. Like most psychologists, I had assumed that these patterns of bias were universal – until I read the paper by Ceci and Liker. These two young psychologists seemed to have stumbled on a rare breed of individuals who had somehow escaped the influence of the well-known cognitive biases that affect most peoples' ability to judge risk – expert gamblers.

As has already been noted, it appears that US weather forecasters form another group with unusually high levels of risk intelligence (Murphy and Winkler 1977). Further research might identify yet more such groups. For example, finance professionals are probably too diverse and heterogeneous to constitute such a group, but there may be particular kinds of finance professional who display higher than average risk intelligence. I suspect that hedge fund principals may be one such subgroup.

Risk intelligence differs utterly from what we normally consider intelligence to be – which is why, when we get it wrong – when banks fail, doctors misdiagnose, and weapons of mass destruction turn out not to exist in a country we have invaded – we are in such a bad position to understand the reasons. Despite what those headlines said in the wake of the financial crisis, we need gamblers and weather forecasters – because we can learn an enormous amount from them, not just about money and rainfall, but about the way we make decisions in all aspects of our lives.

Appendix

The 50 true/false statements in the online calibration test were as follows:

ta.1	A one followed by 100 zeros is a Googol	T
ta.2	Africa is the largest continent	F
ta.3	Alzheimer's accounts for under half the cases of dementia in the USA	F
ta.4	An improper fraction is always less than 1	F
ta.5	Armenia shares a common border with Russia	F
ta.6	There have been over 40 US Presidents	T
ta.7	In 1994, Bill Clinton was accused of sexual harassment by a woman called Paula Jones	T
ta.8	Canberra is the capital of Australia	T
ta.9	Cats are not mentioned in the Bible	T
ta.10	Christianity became the official religion of the Roman empire in the third century AD	F
ta.11	Commodore Matthew Perry compelled the opening of Japan to the West with the Convention of Kanagawa in 1870	F
ta.12	El Salvador does not have a coastline on the Caribbean	T
ta.13	Gout is known as "the royal disease"	F
ta.14	Harry Potter and the Goblet of Fire tells the story of Harry Potter's third year at Hogwarts	F

	Humphrey Bogart had two wives before Lauren Bacall	F
ta.15	In 2008 the population of Beijing was over 20 million people	F
ta.16	In the Old Testament, Jezebel’s husband was Ahab, King of Israel	T
ta.17	Iron accounts for over 30% of the Earth’s composition	T
ta.18	It is possible to lead a cow upstairs but not downstairs, because a cow’s knees cannot bend properly to walk back down	T
ta.19	Lehman Brothers went bankrupt in September 2008	T
ta.20	LL Cool J got his name from the observation “Ladies Love Cool James”	T
ta.21	Male gymnasts refer to the pommel horse as “the pig”	T
ta.22	Mao Zedong declared the founding of the People’s Republic of China in 1949	T
ta.23	More than 10 American states let citizens smoke marijuana for medical reasons	T
ta.24	More than 8 out 10 victims infected by the Ebola virus will die in 2 days	T
ta.25	Most of the terrorists who carried out the attacks on 9/11 were from Saudi Arabia	T
ta.26	Mozart composed over 1,000 works	F
ta.27	Natural gas has an odor	F
ta.28	Of all Arab nations, Lebanon has the highest percentage of Christians	T
ta.29	Over 40% of all deaths from natural disasters from 1945 to 1986 were caused by earthquakes	T
ta.30	Over 50% of Nigeria’s population lives on less than $1 per day	T
ta.31	Stalagmites grow down, and stalactites grow up	F
ta.32	The Italian musical term adagio means that the music should be played quickly	F
ta.33	The Euphrates river runs through Baghdad	F
ta.34	The face on a $100,000 bill is that of Woodrow Wilson	T
ta.35	The Islamic Resistance Movement is better known to Palestinians as Hizbollah	F
ta.36	The Japanese were largely responsible for building most of the early railways in the US West	F
ta.37	The last Inca emperor was Montezuma	F
ta.38	The most frequently diagnosed cancer in men is prostate cancer	T
ta.39	The only stringed symphonic instrument that has a pedestal and a crown is a double bass	F
ta.40	The president of Russia is Vladimir Putin	F
ta.41	The San Andreas Fault forms the tectonic boundary between the Pacific Plate and the North American Plate.	T
ta.42	The US civil war broke out the same year the federal government first printed paper money	T
ta.43	The US Declaration of Independence begins: “We the People of the United States. . .”	F
ta.44	The word “robot” was coined by the American science fiction writer, Isaac Asimov	F
ta.45	The world’s highest island mountain is Mauna Kea	T
ta.46	The Taj Mahal was built by Emperor Shah Jahan in memory of his favorite wife	T
ta.47	There are more people in the world than chickens	F
ta.48	There are no diamond fields in South America	F
ta.49	Wikipedia was launched in 1999 by Jimmy Wales and Larry Sanger	F
ta.50	Number of True Statements	25

References

Apgar D (2006) Risk intelligence: learning to manage what we don't know. Harvard Business School Press, Cambridge, MA

Björn L (2009) Scared silly over climate change. Guardian online. http://www.guardian.co.uk/commentisfree/cif-green/2009/jun/15/climate-change-children. Accessed 3 March 2011

Brier GW (1950) Verification of forecasts expressed in terms of probability. Mon Weather Rev 78(1):1–3

Ceci SJ, Liker JK (1986) A day at the races: s study of IQ, expertise, and cognitive complexity. J Exp Psychol Gen 115:255–266

Christensen-Szalanski JJJ, Bushyhead JB (1981) Physicians' use of probabilistic information in a real clinical setting. J Exp Psychol Hum Percept Perform 7:928–935

Darwin C (1871) The descent of man. John Murray, London

Evans D (in press) Risk intelligence: how to live with uncertainty. New York, Free Press

Funston F, Wagner S (2010) Surviving and thriving in uncertainty: creating the risk intelligent enterprise. Wiley, Hoboken

Gardner H (1983) Frames of mind: the theory of multiple intelligences. Basic Books, New York

Knight FH (1921) Risk, uncertainty and profit. University of Michigan Library, Michigan, 2009 edition

Koriat A, Lichtenstein S et al (1980) Reasons for confidence. J Exp Psychol Hum Learn Mem 6:107–118

Krell E (2010) RiskChat: what is risk intelligence? http://businessfinancemag.com/article/riskchat-what-risk-intelligence-0621. Accessed 7 July 2010

Kruger J, Dunning D (1999) Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psychol 77(6):1121–1134

Lichtenstein S, Fischhoff B (1977) Do those who know more also know more about how much they know? Organ Behav Hum Perform 20:159–183

Lichtenstein S, Fischhoff B (1981) The effects of gender and instructions on calibration. Decision research report. Decision Research, Eugene, OR, 81

Lichtenstein S, Fischhoff B et al (1982) Calibration of probabilities: the state of the art to 1980. In: Kahneman D, Slovic P, Tversky A (eds) Judgement under uncertainty: heuristics and biases. Cambridge University Press, Cambridge, pp 306–334

Murphy AH (1973) A new vector partition of the probability score. J Appl Meteorol 12:595–600

Murphy AH, Winkler RL (1977) Reliability of subjective probability forecasts of precipitation and temperature. J R Stat Soc C Appl Stat 26(1):41–47

Taleb NN (2007) The black swan: the impact of the highly improbable. Allen Lane, London

Author Query Form

Handbook of Risk Theory
Chapter No.: 20

Query Refs.	Details Required	Author's response
AU1	Please update reference Evans (in press) in text as well as in the reference list.	