The Art of Judgment

Andy Nolan – Chief of Project Estimating

Presented for the International Cost Estimating & Analysis Association - www.iceaaonline.com

Introduction

Abstract

Many of us rely on “expert judgement” when estimating, but the evidence is that it can be unreliable. Based on a study of 3760 guesses, we noticed that 70% of people tended to underestimate. A new study in 2020 of 7640 guesses, showed that 66% of people tended to quote a narrow min-max range when guessing a 3-point estimate. When using judgement, most of us are too low and too precise, we are precisely wrong! This research showed that within 30 minutes, 60% of people could improve their judgement accuracy. It seems most people can quickly calibrate themselves, but not everyone. In 2018 we developed a quick and simple calibration test to determine if someone could reliably estimate. The trials spanned 578 people and several companies. It showed that we could reliably predict who, and when, we can trust a guess. The paper summarises four years of research, offers training methods and methods for how to enhance judgement accuracy. It also suggests that for some, they will need to rely on data rather than their judgement.

Key Messages

- 84% of people reported that they relied on judgement at some point when estimating.
- When using judgement, 70% of people tended to under estimate, especially when unsure.
- When using judgement, 66% of people quote a narrow Min – Max range i.e. they are too precise.
- The Wisdom of the Crowd method is an effective way to improve judgement accuracy
  - Judgement Accuracy = Crowd Size * Confidence
- Where it is not possible to form a group, the Calibration Test will determine an individuals judgement style
  - Judgement Accuracy = Calibration * Confidence

Presented for the International Cost Estimating & Analysis Association - www.iceaaonline.com

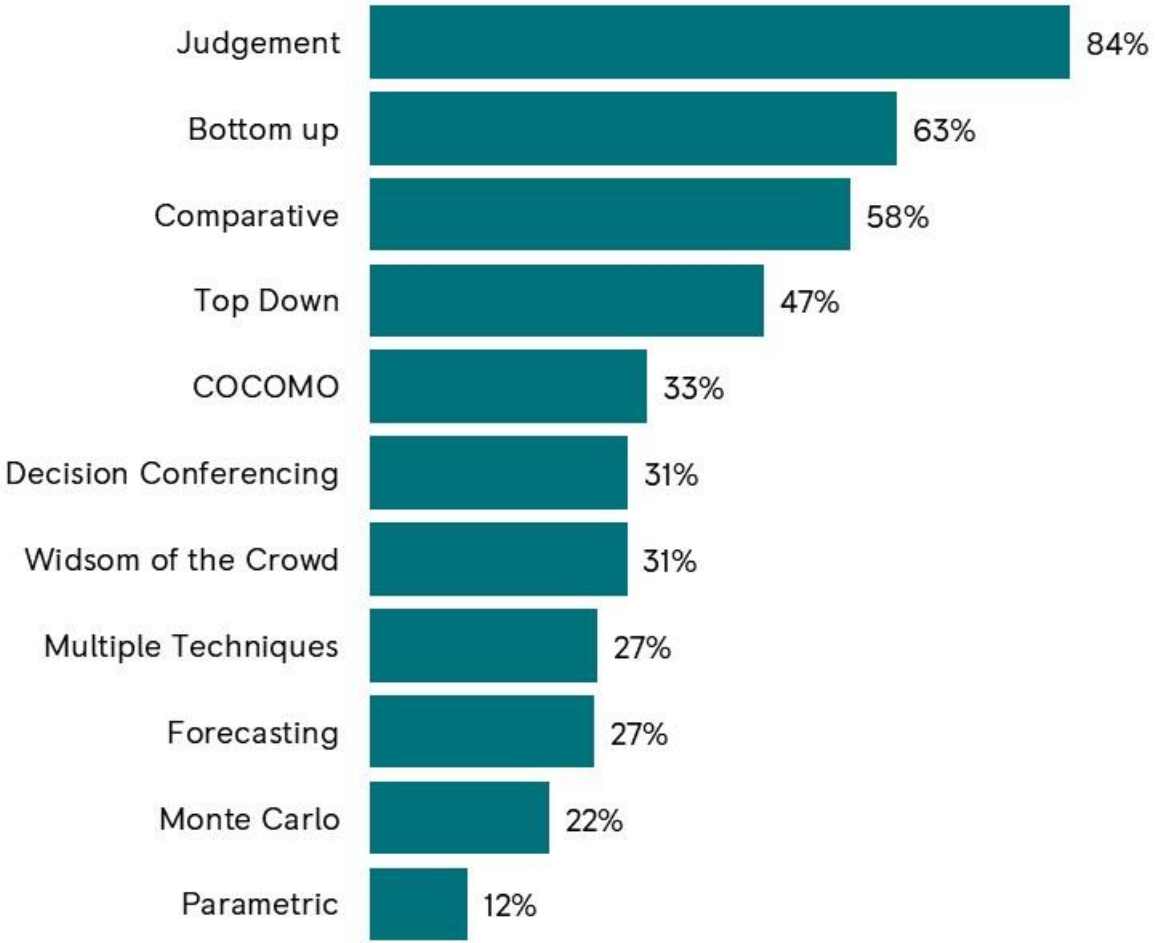
2014 Research

84% of people reported that they relied on judgement at some point when estimating.

What estimating techniques do we use?

According to a 2014 study of 158 estimates, it showed that judgement was one of the most commonly used techniques. We have to use judgement sometimes, even if it is an expression of our experience.

If we have to use judgement we need to determine who and when we can rely on it.



2017 Research

When using judgement, 70% of people tended to under estimate, especially when unsure.

The research

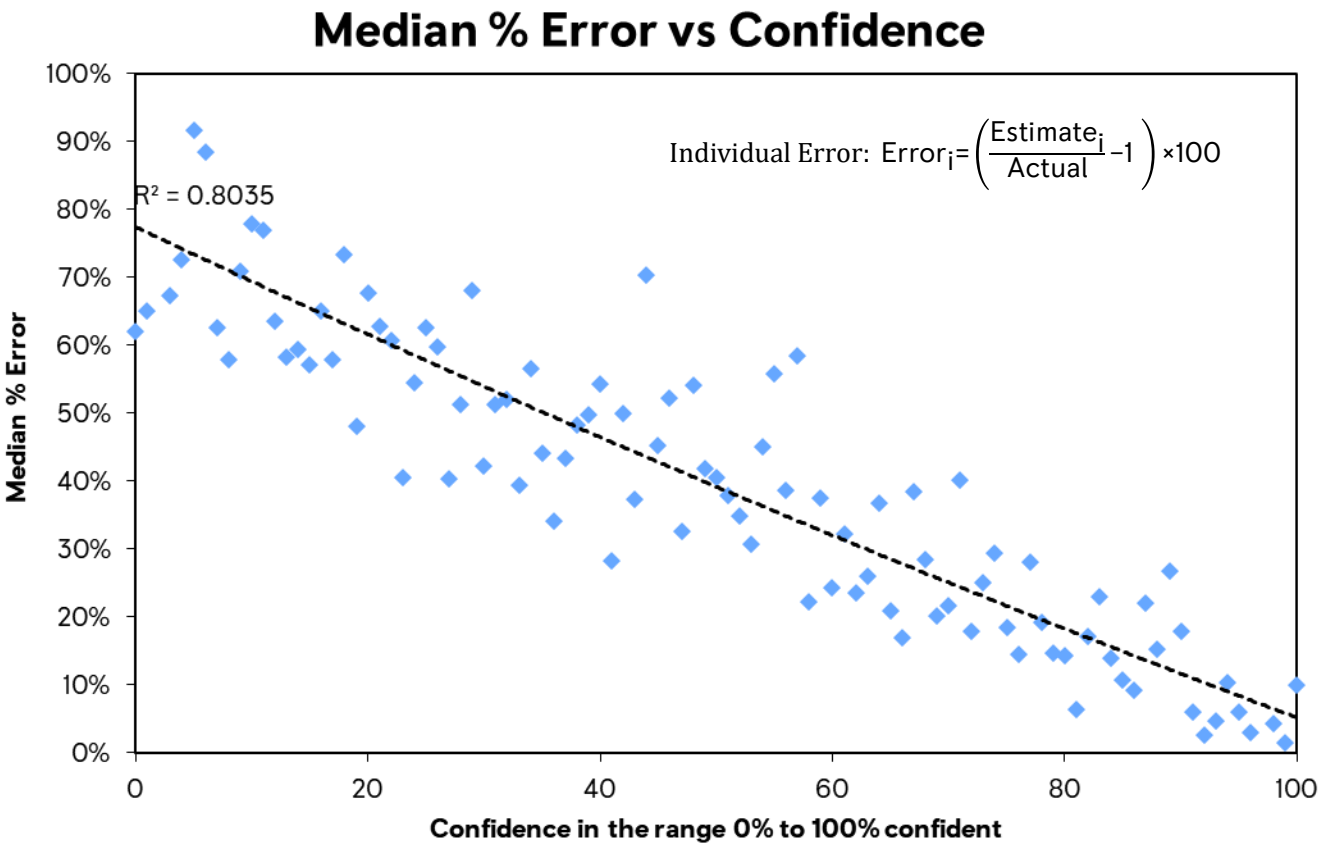
The research was to see if there was a relationship between peoples Confidence in their guess and the guess accuracy

1. People were given a range of general knowledge questions where they had to guess a numerical value like the height of buildings, populations of countries and so on
2. For each guess, the participants were asked to score their Confidence in their answer from 0 (no confidence) to 100 (certainty)

Accuracy correlates with Confidence

Based on 3760 answers, there is a 0.8 R^2 correlation between a "groups" Confidence and the accuracy of their median guess.

We can probably assume that Confidence is related to a groups knowledge or experience of a topic

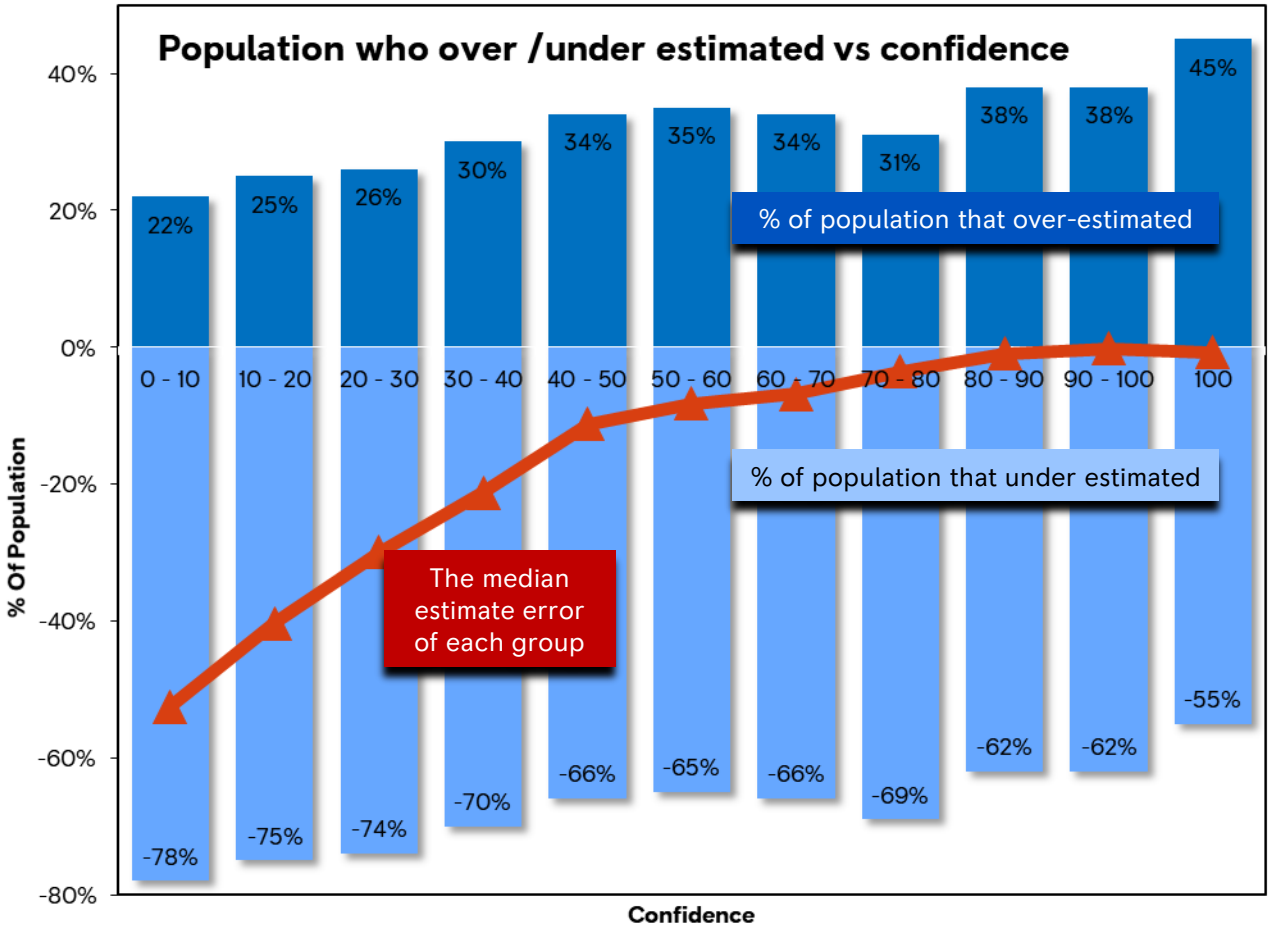


70% of us under estimate!

The results of 3760 guesses showed that 70% of people tended to under estimate, especially when unsure.

The red line on the chart is the median %error plotted against Confidence (0 = no confidence) to 100 (certain).

The level of under estimating increased as people became less Confident.



2020 Research

When using judgement, 66% of people quote a narrow Min – Max range i.e. they are too precise.

Can we learn to be unsure?

In 2020, we performed research to see if people can express an adequate Min – Max range for their estimates. *A guess was considered to be accurate when the real answer lay inside the Min to Max range*.

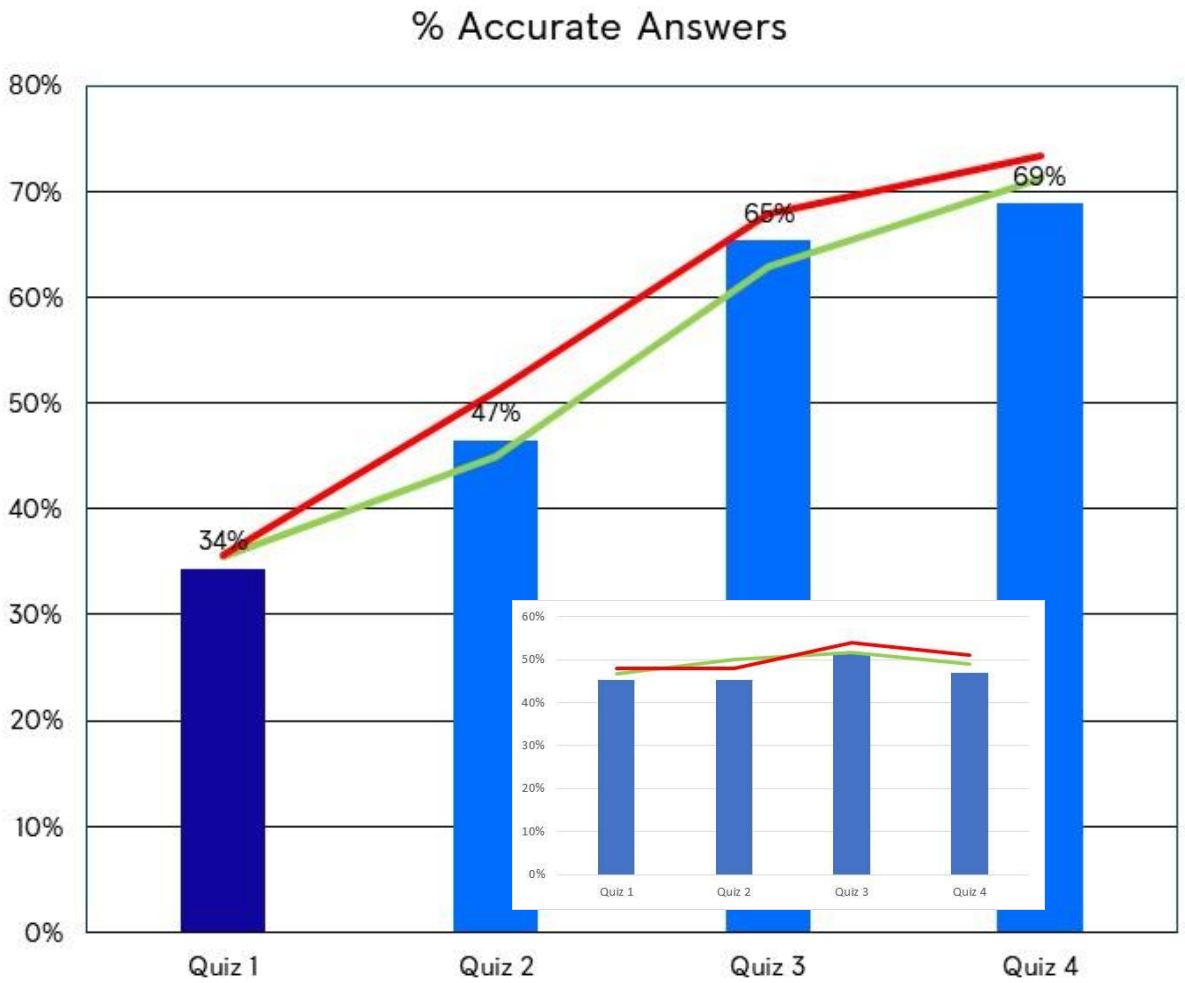
1. They were given 4 quizzes, each quiz having 10 general knowledge questions. For each question, the participant had to quote a Min and Max that they believed would span the right answer.
2. The estimators were asked to complete quiz 1, learn from their results, then see if they can improve their accuracy on subsequent quizzes. Quiz 1 was therefor used as a “datum” to assess for relative improvement.
3. Participants were also split into groups. The control group were given no instructions, the other groups were given different types of incentive. The theory was that incentivised groups would perform better than the control group.

Most people can learn to be unsure

We got 191 completed quizzes containing 7640 guesses. Despite having immediate feedback, 42% of people showed no signs of improvement from Quiz 1 to Quiz 4. Some people prefer to be precise than accurate!

Of the 58% that showed improvement (see graph), we see that for Quiz 1, they were on average 34% accurate, showing a general difficulty in expressing an adequate Min to Max range. However, this group of people went on to double their accuracy by Quiz 4.

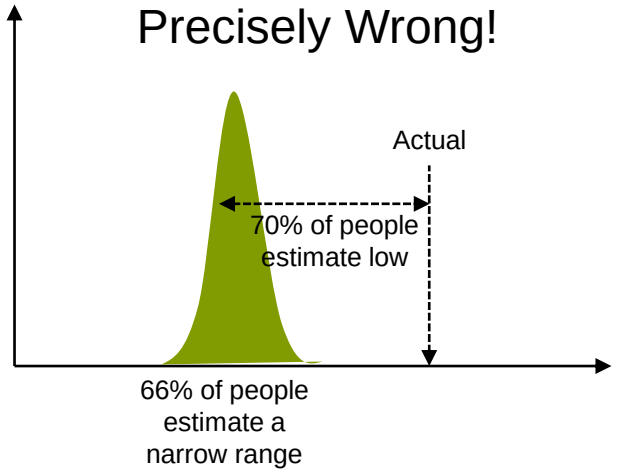
The green line represents a sub-group that were incentivized with a reward. They showed little improvement over the "control" group. The red line were people who would be "penalised" for a wrong answer. They showed a slight improvement over the control group



In summary

The 2017 research showed that we tend to under estimate. The 2020 research shows that we also under estimate the Min – Max range.

It seems that when using judgement, many of us are too precise and inaccurate – we are "precisely wrong".

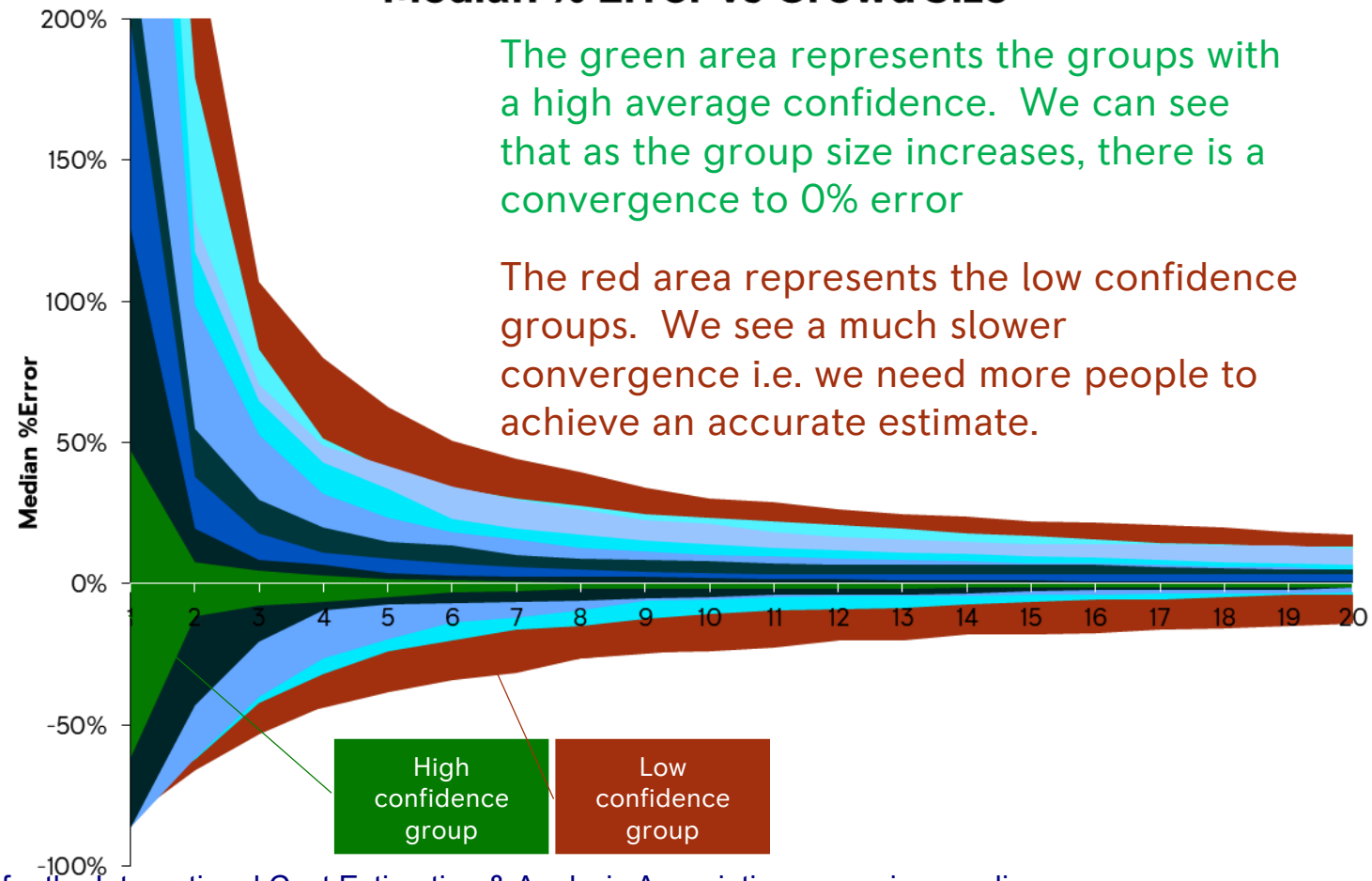


The Wisdom of the Crowd

Judgement Accuracy = Crowd Size * Confidence

The Wisdom of the Crowd

Based on the 3760 guesses, we randomly formed groups. For each group, we calculated the median error and the groups average Confidence. This chart is based on 20 million simulations .



Judgement Accuracy = Crowd Size * Confidence

The table shows the number of people needed to achieve an estimate that meets a desired +/- precision. As confidence drops, the crowd size needs to increase. The +/- values shown along the title axis of the table is used to express a moor appropriate Min – Max range, with a confidence interval of 95%,

Median		Required Estimate Precision								
		-93% - 1280%	-87% - 640%	-77% - 320%	-62% - 160%	-45% - 80%	-29% - 40%	-17% - 20%	-9% - 10%	-5% - 5%
Estimator Confidence	100	1	1	1	1	2	3	4	5	5
	90 - 100	1	1	1	2	3	3	4	5	12
	80 - 90	1	1	2	3	5	5	7	11	41
	70 - 80	1	2	2	4	6	10	17	40	185
	60 - 70	1	2	2	4	8	16	41	117	X
	50 - 60	1	2	3	6	14	30	101	X	X
	40 - 50	1	2	2	6	16	44	X	X	X
	30 - 40	2	2	6	12	39	162	X	X	X
	20 - 30	2	4	8	21	115	X	X	X	X
	10 - 20	3	6	12	47	X	X	X	X	X
	0 - 10	4	8	22	168	X	X	X	X	X

Calibration

Judgement Accuracy = Calibration * Confidence

Do we know ourselves?

In the previous research, we asked people to score their Confidence that they were accurate. We saw a correlation between accuracy and Confidence. But can we trust peoples self assessment of their Confidence?

The research was to see if we could develop a simple way to test for self-awareness and then to see how self-awareness related to judgment accuracy.

1. We asked each person to complete a 20 question test.
2. Each question needed a yes/no response
3. Each person had to score their Confidence that they got the right answer.

Accuracy = number of correct answers

Average Confidence

#	Question	Yes / No	Conf
1	The Eiffel Tower is taller than the Empire State Building?	No	60
2	The Amazon River is considered to be the longest river in the world	Yes	80
3	A Tennis ball is bigger in diameter than a Cricket ball	No	50
4	More people live in China than India	Yes	10
5	There are 52 states in the USA	No	50
6	A sextillion is bigger than a septillion	Yes	70
7	On average dogs lives longer than cats	No	50
8	Titanium has a higher melting point then stainless steel	Yes	90
9	The United States is bigger in area than Canada	No	100
10	The Amazon rainforest is bigger in area than the Great Barrier Reef	Yes	80
11	An Indian elephant can live longer than an African Elephant	No	90
12	Kilimanjaro is smaller than K2	Yes	10
13	More people live in Germany than Japan	No	80
14	The Atlantic Ocean is bigger than The Pacific Ocean	Yes	50
15	The Sahara Desert is bigger than Europe	No	100
16	A table tennis ball is bigger (diameter) than a golf ball	Yes	60
17	The planet Mercury has a bigger diameter than Mars	No	80
18	-25 degrees C is warmer than -25 degrees F.	Yes	20
19	Venezuela is north of the equator	No	60
20	Newton discovered the equation E=mc2	Yes	100

Calibration-Factor

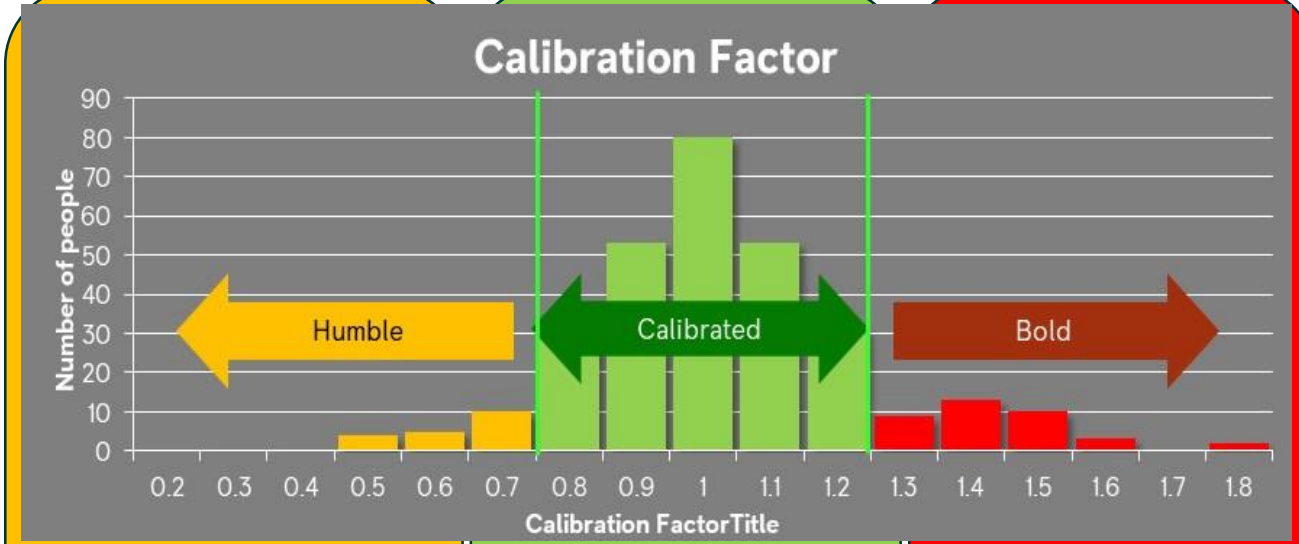
1. Average Confidence: We take the average confidence for the 20 questions. In this case the average confidence is 65%
2. Accuracy: We calculate the % of questions they answered correctly. In this case 13 questions were accurate = 65%
3. *A persons Calibration Factor = Average Confidence / Accuracy*
   - *65% / 65% = 1*

Calibrated, Humble & Bold

The ideal calibration factor is 1. We set the upper and lower bands at 1 sigma.

Someone is considered Humble if their Calibration Factor < 0.8

Someone is considered Bold if their Calibration Factor > 1.2



Humble

Calibration Factor < 0.8. A person whose Confidence is lower than their Accuracy and this means a person got more answers right then they believed they should. For example, if they had a Confidence of 40% but got 80% of the questions right, they would have a Calibration Factor of 0.5.

Calibrated

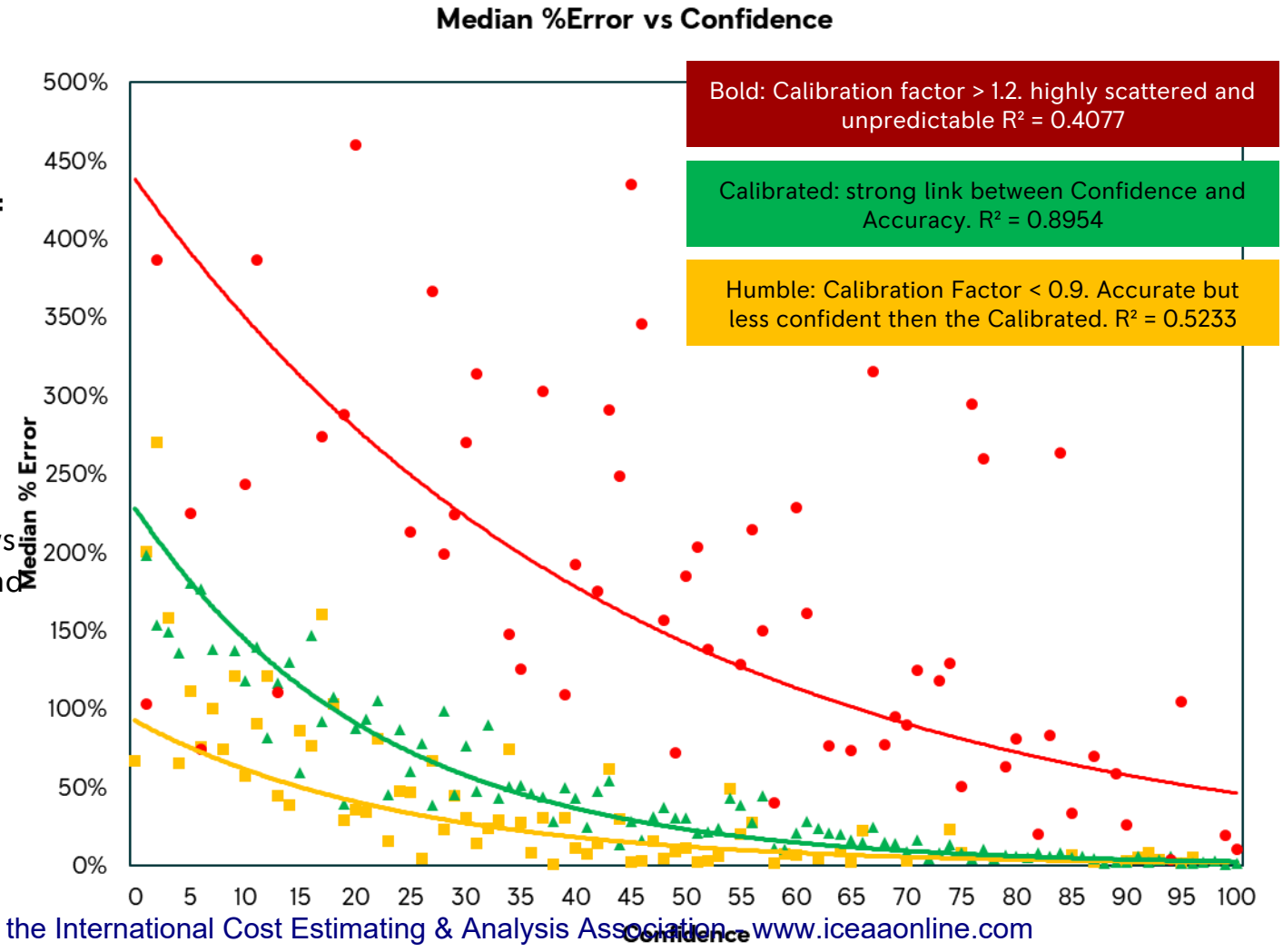
Calibration Factor between 0.8 and 1.2 . A person whose Confidence matches their Accuracy. If a person had an average Confidence of 50% and they got 50% of the answers right, they would have a Calibration Factor of 1. So would a person who was 30% Confident and 30% Accurate. Also for a person who was 80% Confidence and 80% Accurate.

Bold

Calibration Factor > 1.2. A person whose Confidence is higher than their Accuracy and this means a person got less answers right then they believed they should. For example, if they had a Confidence of 80% but got 40% of the questions right, they would have a Calibration Factor of 2.

Judgement Accuracy = Calibration * Confidence

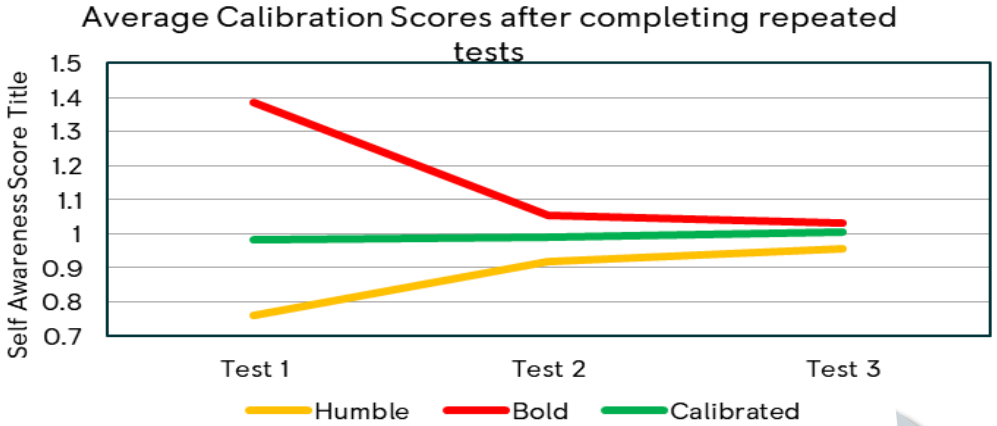
The chart shows the guesses from 578 people, split into the 3 groups of Bold, Humble and Calibrated. The chart shows horizontally Confidence and vertically the absolute % median error of the group (people with the same Confidence)

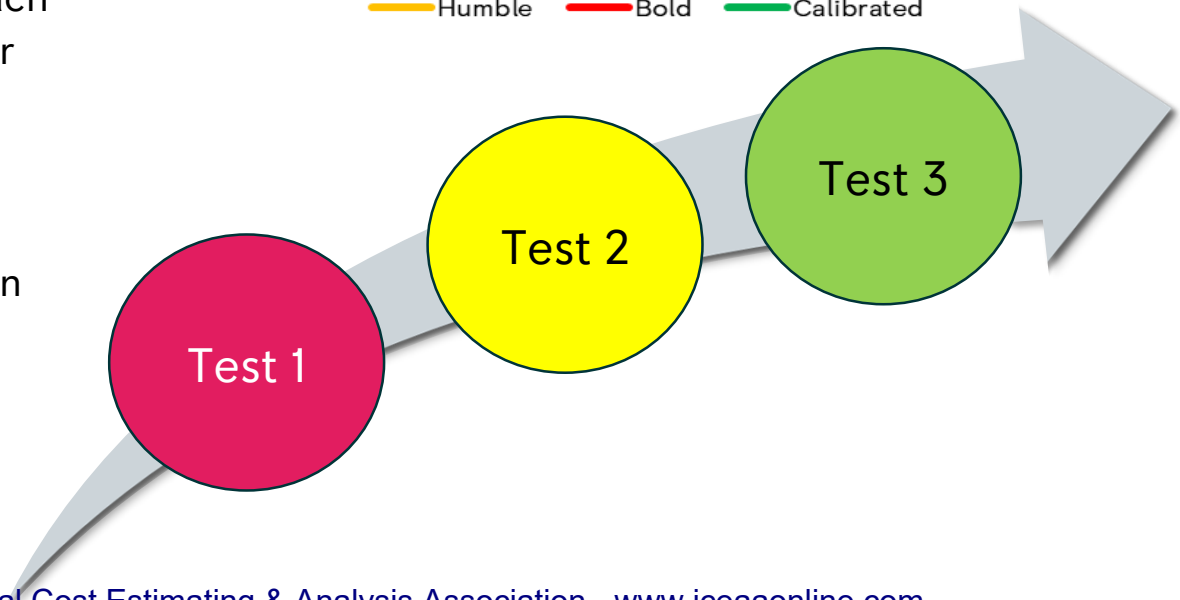


Most people can calibrate themselves

Like an archer that learns from each shot, it is possible to improve your Calibration by learning from your results and taking another test.

Of the 44 volunteers, we saw an overall improvement in Calibration scores from repeated testing.

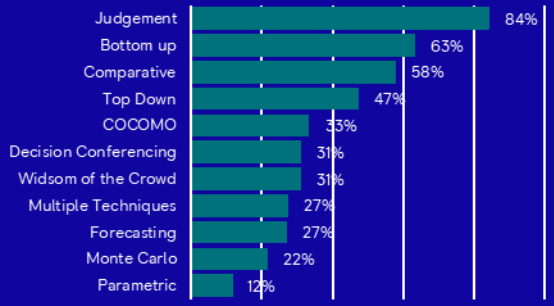




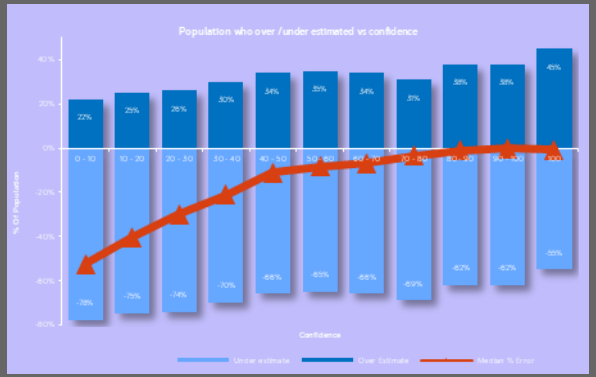
Conclusions

Conclusions

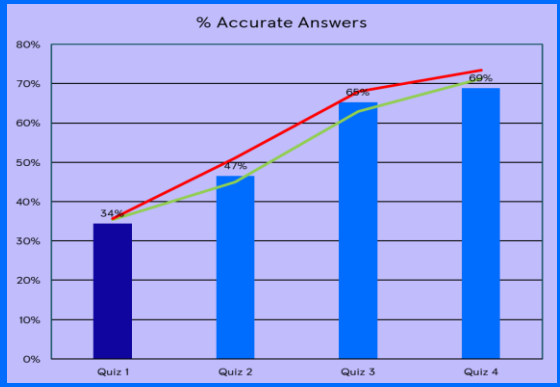
84% of us rely on Judgement when estimating



70% of us will under estimate, especially when unsure



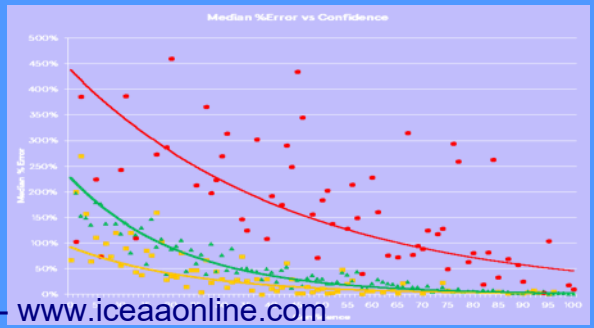
66% of us will be too precise in our 3-point ranges



The Wisdom of the Crowd Accuracy = Crowd Size * Confidence

Median	Required Estimate Precision								
Estimator Confidence	-93% - 1280%	-87% - 640%	-77% - 320%	-62% - 160%	-45% - 80%	-29% - 40%	-17% - 20%	-9% - 10%	-5% - 5%
100	1	1	1	1	2	3	4	5	5
90 - 100	1	1	1	2	3	3	4	5	12
80 - 90	1	1	2	3	5	5	7	11	41
70 - 80	1	2	2	4	6	10	17	40	185
60 - 70	1	2	2	4	8	16	41	117	X
50 - 60	1	2	3	6	14	30	101	X	X
40 - 50	1	2	2	6	16	44	X	X	X
30 - 40	2	2	6	12	39	162	X	X	X
20 - 30	2	4	8	21	115	X	X	X	X
10 - 20	3	6	12	47	X	X	X	X	X
0 - 10	4	8	22	158	X	X	X	X	X

The Calibration Test Accuracy = Calibration * Confidence



Links to associated papers

- How Many Estimators Does It Take To Change A Lightbulb
  - The wisdom of the crowd
- The Good, The Bad & The Ugly
  - Calibration test