WORC 2024

World Overflight Risk Conference

2 - 4 July 2024 | Warsaw, Poland

Educational Keynote: The Failure of Risk Management and How to Fix It

Douglas Hubbard, President, Hubbard Decision Research

#WORC2024

PRESENTED BY

OSPREY FLIGHT SOLUTIONS

EASA European Union Aviation Safety Agency

Introduction

The Biggest Risk

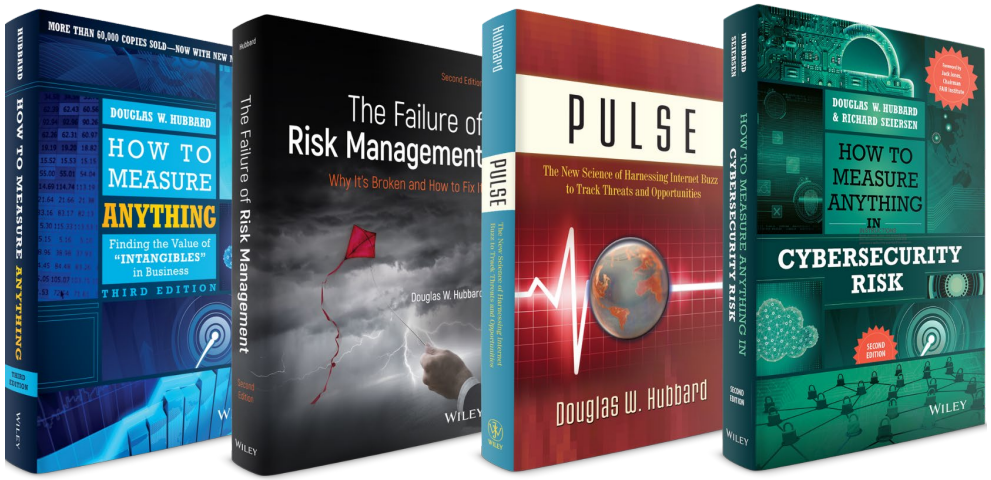
Question: What is your single biggest risk?

Answer: How you measure risk.

How do we know what works and what doesn’t?

Hubbard Decision Research Background

In 200+ major analysis projects, HDR has been able to show that no matter how difficult the measurement and monetization problem appears to be, we find a way to evaluate it and communicate the results.



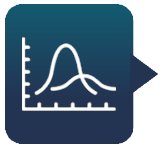
- The benefits and risks of **dams on the Mekong River**
- The relative value of ***R&D portfolios*** in aerospace, biotech, and pharma
- ***Logistics forecasts for the battlefield*** and the effectiveness of training for the US Military
- ***IT Project Portfolio*** *and* ***Cybersecurity Risk Assessments*** *in several industries*
- *Risks and benefits of* ***Environmental policy*** *for US farmers and public health*
- *The benefits of* ***Educational assistance*** *in inner city schools*
- *The benefits of roads, schools and hospitals in Haiti and how to prioritize them for the* ***United Nations***

Key Points

According to many large and diverse studies, experience alone isn't a "best practice."

The most popular and seemingly "structured" methods improve confidence in risk management while making judgements worse.

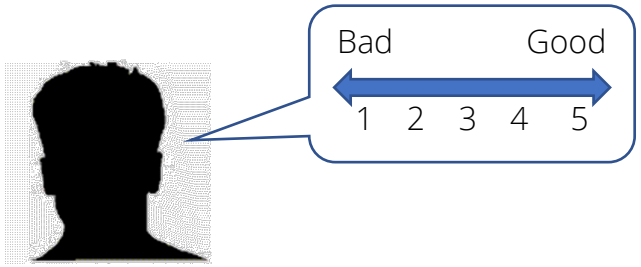
However, some subjective methods *objectively* outperform other subjective methods.

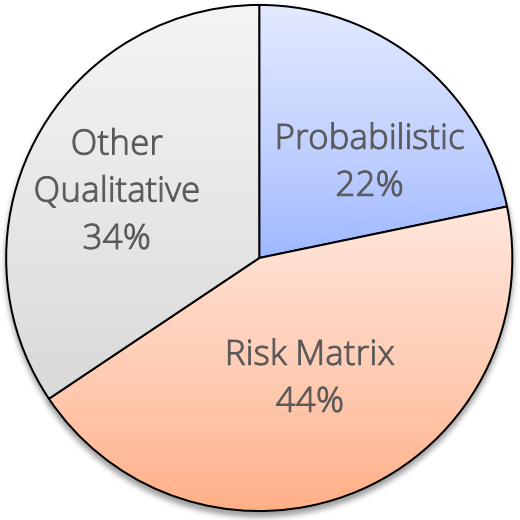


Nothing is immeasurable. If it matters at all, it has observable consequence.

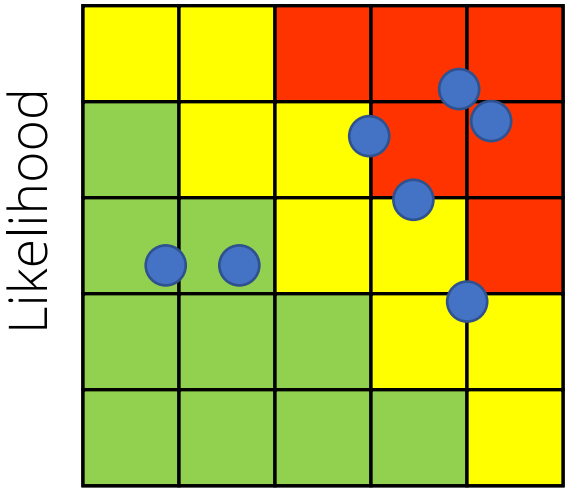
Do “Scores” and “Scales” Work?

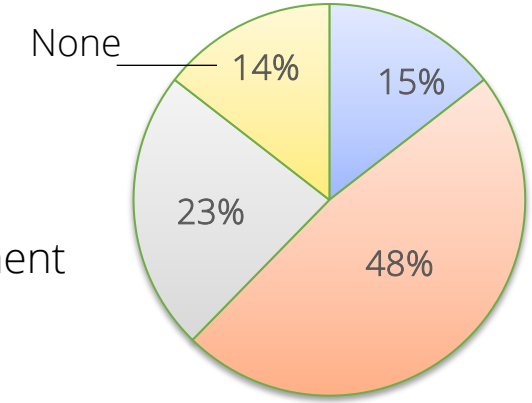
The Current Most Popular Methods in Risk Assessment



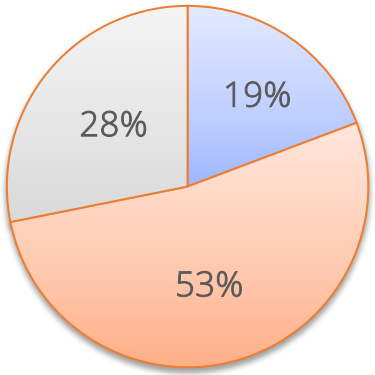


Enterprise Risk Management





Project Management



Cybersecurity

The Analysis Placebo

Confidence in Decision Making Methods is Detached From Performance

Organizational Behavior and Human Decision Processes 107, no. 2 (2008): 97– 105.

Journal of Behavioral Decision Making 3, no. 3 (July/ September 1990): 153– 174.

Law and Human Behavior 23 (1999): 499– 516.

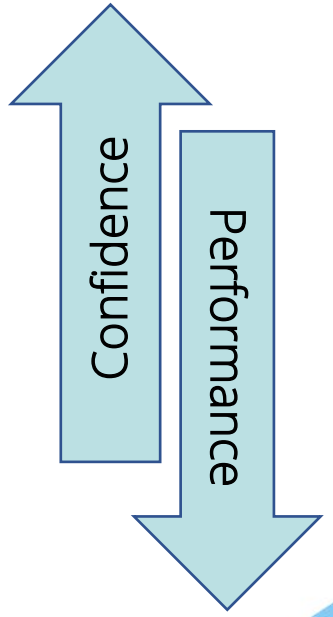
Organizational Behavior and Human Decision Processes 61, no. 3 (1995): 305– 326.

Interaction with Others Increases Decision Confidence but Not Decision Quality: Evidence against Information Collection Views of Interactive Decision Making

Heath and Gonzalez

Abstract

We present three studies of *interactive decision making*, where decision makers interact with others before making a final decision alone. Because the theories of lay observers and social psychologists emphasize the role of information collection in interaction, we



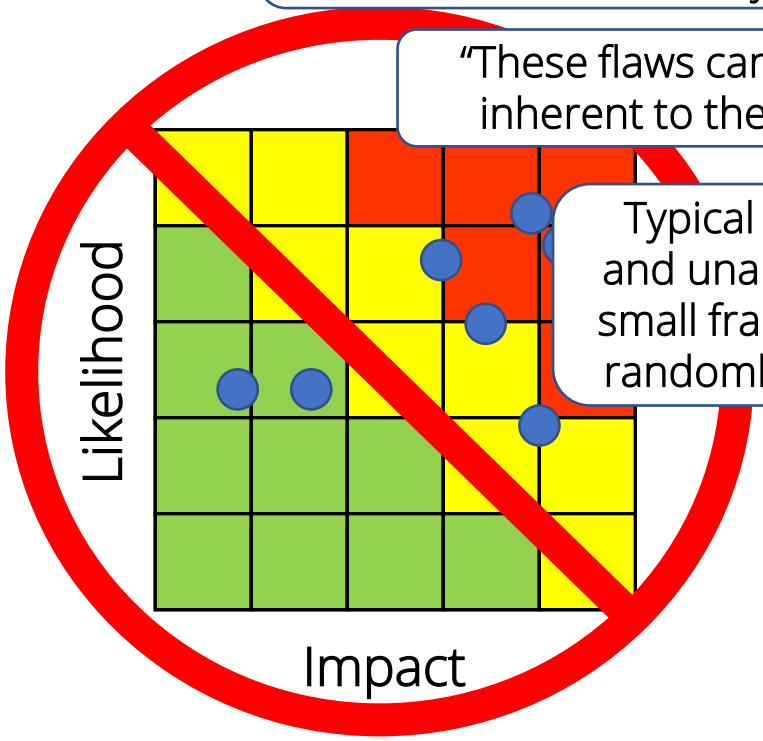
So How Do We Know What Works?

Research shows that learning from experience requires consistent, fast, unambiguous feedback in an environment that isn't entirely random. (Kahneman, Klein)

- We don't get feedback like that in most risk management fields.
- Lacking that, we can look at hundreds of studies over decades with tens of thousands of data points which measured the performance of various approaches.

Do "Scores" and "Scales" Work?

The Ubiquitous Risk Matrix



"The ranking produced by RMs was shown to be unduly influenced by their design, which is ultimately arbitrary."

"These flaws cannot be corrected and are inherent to the design and use of RMs."

Typical risk matrices can correctly and unambiguously compare only a small fraction (e.g., less than 10%) of randomly selected pairs of hazards.

"Effective allocation of resources to risk-reducing countermeasures cannot be based on the categories provided by risk matrices."

"For risks with negatively correlated frequencies and severities, they can be 'worse than useless,' leading to worse-than-random decisions."

"Risk Matrices should not be used for decisions of any consequence"

ty of Petroleum Engineers Ecor
): 56–66.

The Risk of Using Risk Matrices

P. Thomas, R. Bratvold, and J. E. Bickel

Abstract

a widely espoused approach to assess and analyze risks in the

28, no. 2 (2008).

What's Wrong with Risk Matrices?

L. A. Cox, Jr.

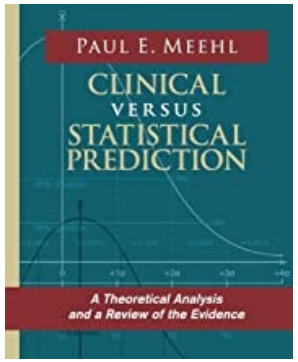
erity" r
rism ri
esponding risk priority
way construction
nt,
climate ch
and ente
Standar
many or

Experts vs. Algorithms

What the Research Says About Statistical Methods vs. Subject Matter Experts

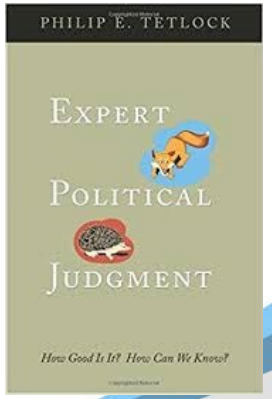
Paul Meehl assessed 150 studies comparing experts to statistical models in many fields (sports, prognosis of liver disease, etc.).

"There is no controversy in social science which shows such a large body of qualitatively diverse studies coming out so uniformly in the same direction as this one."

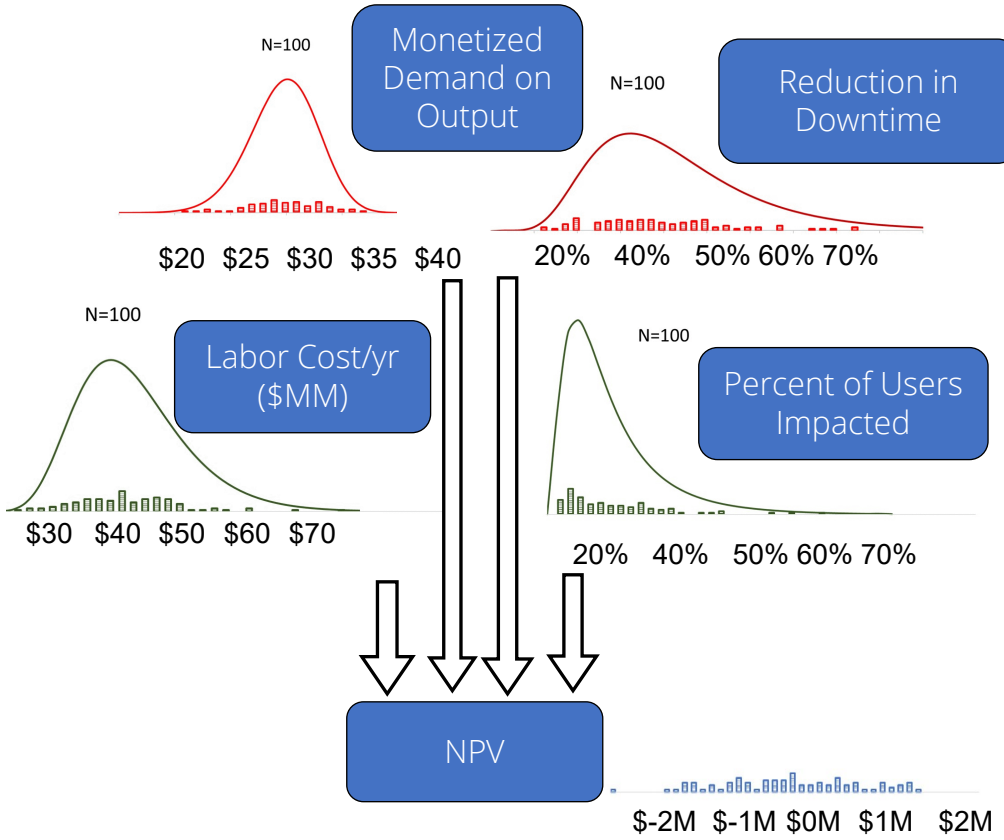


Philip Tetlock tracked a total of over 82,000 forecasts from 284 experts in a 20-year study covering politics, economics, war, technology trends and more.

"It is impossible to find any domain in which humans clearly outperformed crude extrapolation algorithms, less still sophisticated statistical ones."



Doing the Math with Monte Carlo



Published in *International Journal of Forecasting*, 10 (1994), 495-906

Judgmental Decomposition: When Does It Work?

Donald G. MacGregor

Society of Petroleum Engineers (2000)

The Application of Probabilistic and Qualitative Methods to Asset Management Decision Making

SSCAG/SCAF/EACE Joint International Conference (2008)

An Assessment of the Inherent Optimism in Early Conceptual Designs and Its Effect on Cost and Schedule Growth

D. Bearden, C. Freaner, R. Bitten, and D. Emmons

Abstract

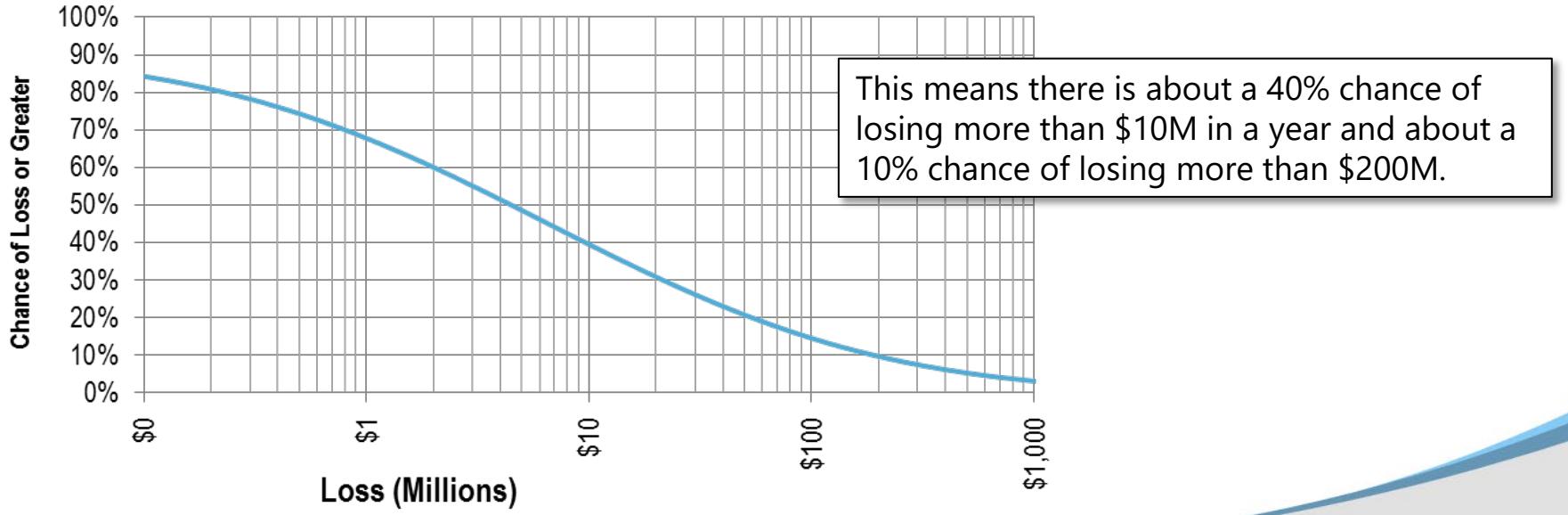
When missions experience cost growth, cost estimators are often criticized for underestimating the cost of missions in the early conceptual design stage. The final spacecraft and instrument payload configuration at launch, however, can be significantly different as the project evolves, thereby leading to cost "growth" as compared to these lower initial estimates. In order to make a more robust initial

What Measuring Risk Looks Like

What if we could measure risk more like an actuary? For example, "The probability of losing more than $10 million due to security incidents next year is 16%."

What if we could prioritize RM investments based on a "Return on Mitigation"?

RM Strategy	Cost (Present Value)	Annualized Return on Mitigation
Supply Chain Diversification A	$11MM to $30MM	160%
Hardened Data Initiative B	$5MM to $8MM	85%
Backup Facility C	$22MM to $45MM	20%
Etc.		



This means there is about a 40% chance of losing more than $10M in a year and about a 10% chance of losing more than $200M.

A Version of Risk Tolerance

The Loss Exceedance Curve

Unambiguous risk lets us have unambiguous risk tolerance.

Neuron Vol. 47, (2005): 763–770

The Neural Basis of Financial Risk Taking

Camelia M. Kuhnen and Brian Knutson

Abstract

Investors systematically deviate from rationality when making financial decisions, yet

Journal of Personality and Social Psychology
2001, Vol. 81, No. 1, 146–159

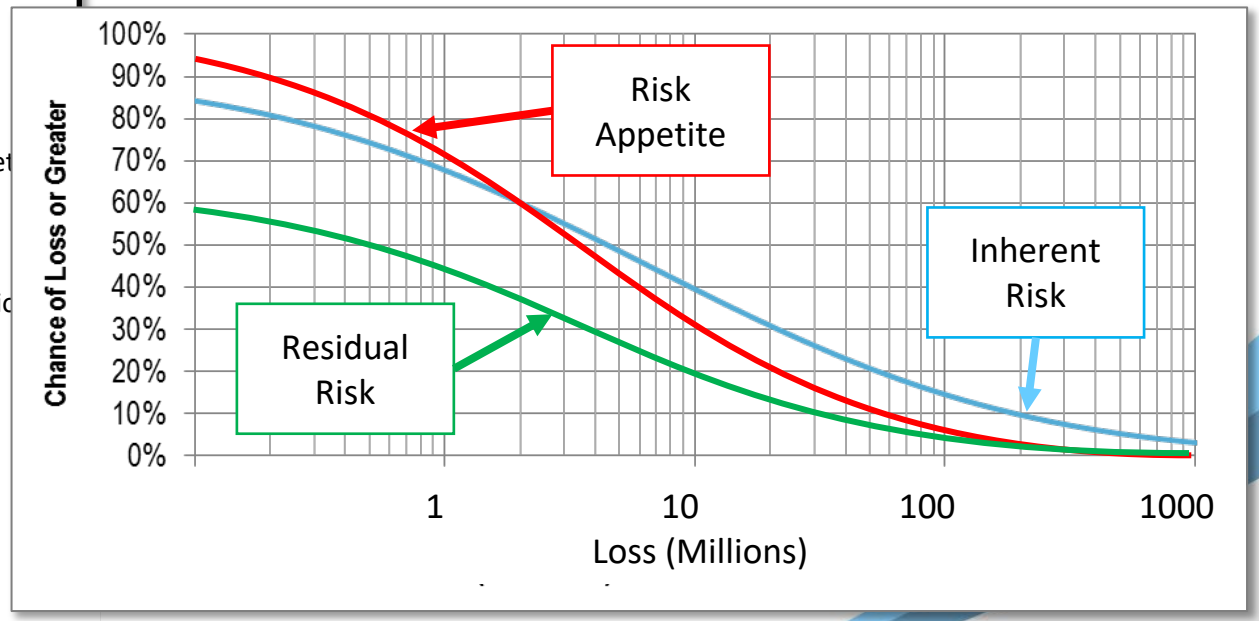
Copyright 2001 by the American Psychological Association, Inc.
0022-3514/01/$5.00 DOI: 10.1037//0022-3514.81.1.146

Fear, Anger, and Risk

Jennifer S. Lerner
Carnegie Mellon University

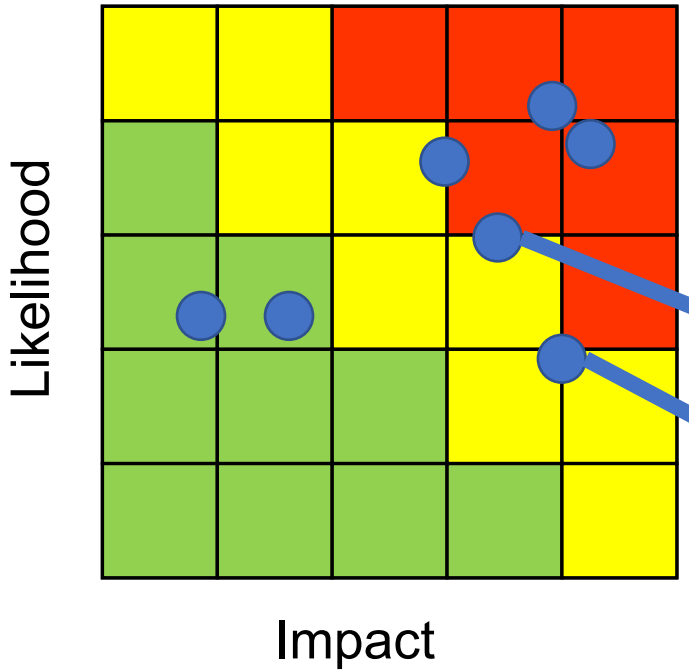
Dacher Keltner
University of California, Berkeley

Drawing on an appraisal-tendency framework (J. S. Lerner & D. Keltner, 2000), the authors predicted and found that fear and anger have opposite effects on risk perception. Whereas fearful people expressed pessimistic risk estimates and risk-averse choices, angry people expressed optimistic risk estimates and risk-seeking choices. These opposing patterns emerged for naturally occurring and experimentally induced fear and anger. Moreover, estimates of angry people more closely resembled those of happy people than those of fearful people. Consistent with predictions, appraisal tendencies accounted for these



Converting From the Risk Matrix

Each of these examples can be found on
https://www.howtomeasureanything.com/riskmanagement/



One-For-One Substitution Model ← Uncollapse for HDR PRNG

Hubbard Decision Research

Contact HDR to develop custom quantitative methods for your firm.

www.hubbardresearch.com

info@hubbardresearch.com

Introduction: The One-For-One Substitution Model shows how we can replace the basic heat map using purely quantitative methods. Like the heat map, this approach relies entirely on the subjective estimates of likelihood and impact from experts. Unlike the heat map, it uses mathematically unambiguous probabilistic expressions of likelihood and impact which can be used in simulations where risks can be properly "added up" to compute risks for a larger portfolio. The first tab requires user input. Simply describe the risk or event that is being considered in column B under "Risk Name". Then, estimate the

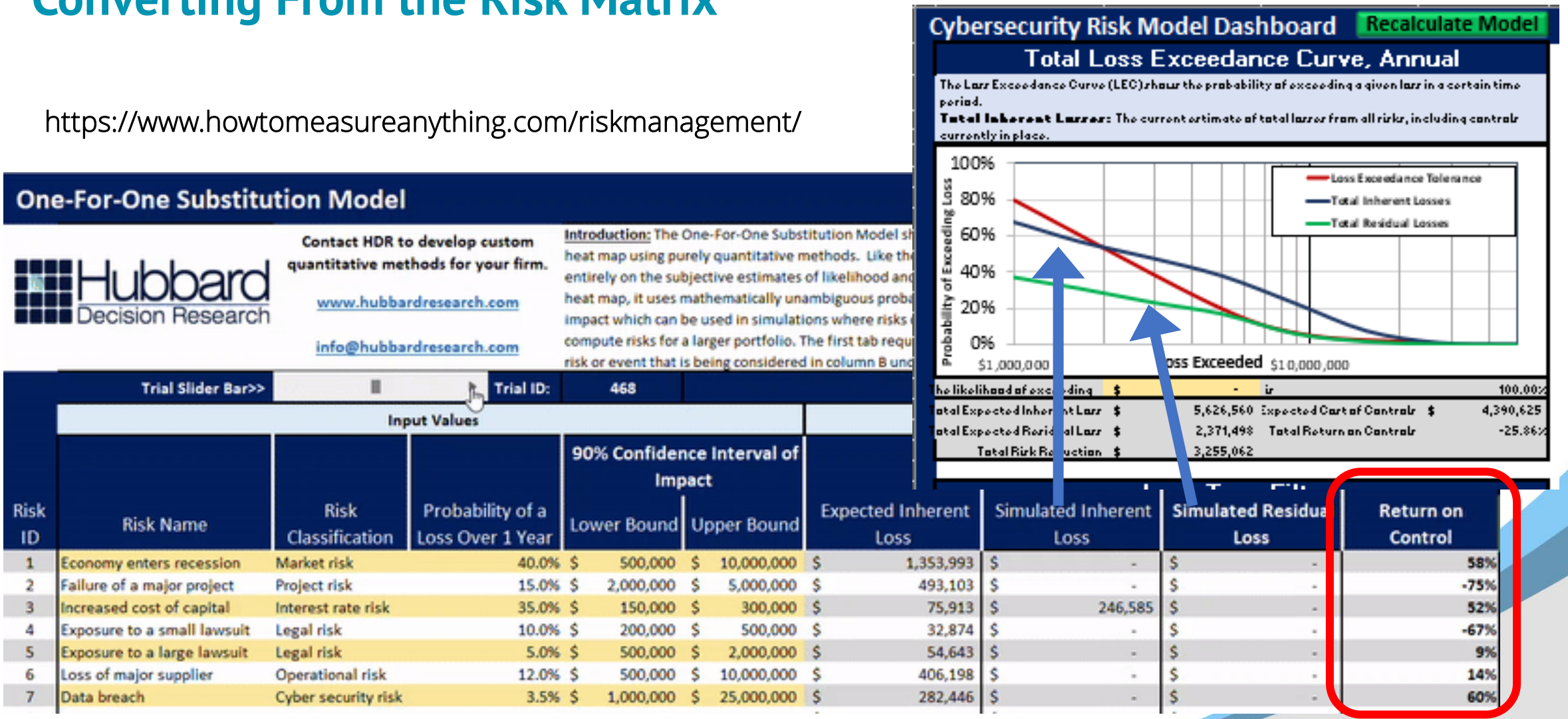
Trial Slider Bar>> Trial ID: 320

Input Values

Risk	Risk Name	Risk Classification	Probability of a Loss Over 1 Year	90% Confidence Interval of Impact		Expected Inherent Loss	Simulated Inherent Loss
				Lower Bound	Upper Bound		
1	Economy enters recession	Market risk	40.0%	$ 500,000	$ 10,000,000	$ 1,353,993	$ 1,212,085
2	Failure of a major project	Project risk	15.0%	$ 2,000,000	$ 5,000,000	$ 493,103	$ -
3	Increased cost of capital	Interest rate risk	35.0%	$ 150,000	$ 300,000	$ 75,913	$ -
4	Exposure to a small lawsuit	Legal risk	10.0%	$ 200,000	$ 500,000	$ 32,874	$ -
5	Exposure to a large lawsuit	Legal risk	5.0%	$ 500,000	$ 2,000,000	$ 54,643	$ -
6	Loss of major supplier	Operational risk	12.0%	$ 500,000	$ 10,000,000	$ 406,198	$ -
7	Data breach	Cyber security risk	3.5%	$ 1,000,000	$ 25,000,000	$ 282,446	$ -

Converting From the Risk Matrix

https://www.howtomeasureanything.com/riskmanagement/



One-For-One Substitution Model

Hubbard Decision Research

Contact HDR to develop custom quantitative methods for your firm.

www.hubbardresearch.com

info@hubbardresearch.com

Introduction: The One-For-One Substitution Model sh
heat map using purely quantitative methods. Like the
entirely on the subjective estimates of likelihood and
heat map, it uses mathematically unambiguous proba
impact which can be used in simulations where risks
compute risks for a larger portfolio. The first tab requ
risk or event that is being considered in column B und

Trial Slider Bar>> | Trial ID: 468

Risk ID	Risk Name	Risk Classification	Probability of a Loss Over 1 Year	90% Confidence Interval of Impact: Lower Bound	90% Confidence Interval of Impact: Upper Bound	Expected Inherent Loss	Simulated Inherent Loss	Simulated Residual Loss	Return on Control
	Input Values								
1	Economy enters recession	Market risk	40.0%	$ 500,000	$ 10,000,000	$ 1,353,993	$ -	$ -	58%
2	Failure of a major project	Project risk	15.0%	$ 2,000,000	$ 5,000,000	$ 493,103	$ -	$ -	-75%
3	Increased cost of capital	Interest rate risk	35.0%	$ 150,000	$ 300,000	$ 75,913	$ 246,585	$ -	52%
4	Exposure to a small lawsuit	Legal risk	10.0%	$ 200,000	$ 500,000	$ 32,874	$ -	$ -	-67%
5	Exposure to a large lawsuit	Legal risk	5.0%	$ 500,000	$ 2,000,000	$ 54,643	$ -	$ -	9%
6	Loss of major supplier	Operational risk	12.0%	$ 500,000	$ 10,000,000	$ 406,198	$ -	$ -	14%
7	Data breach	Cyber security risk	3.5%	$ 1,000,000	$ 25,000,000	$ 282,446	$ -	$ -	60%

Cybersecurity Risk Model Dashboard

Recalculate Model

Total Loss Exceedance Curve, Annual

The Loss Exceedance Curve (LEC) shows the probability of exceeding a given loss in a certain time period.

Total Inherent Losses: The current estimate of total losses from all risks, including controls currently in place.

The likelihood of exceeding	$ -	ir	100.00%		
Total Expected Inher t Loss	$ 5,626,560	Expected Cost of Controls	$ 4,390,625		
Total Expected Resi l Loss	$ 2,371,498	Total Return on Controls	-25.86%		
Total Risk Re uction	$ 3,255,062				

So Why Don't We Use More Quantitative Methods?

Commonly Stated Reasons For Not Using Quantitative Methods

Have you heard (or said) any of these?

"We don't have sufficient data."

"Risk management is too complex to model."

"Each situation is too unique and complex to apply scientific analysis of historical data."

"How do you know you have all the variables?"

The implied (and unjustified) conclusion from each of these is….

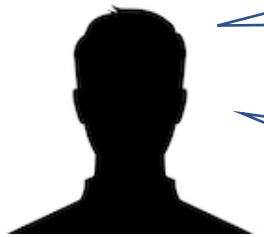
"Therefore, we are better off relying on our experience."

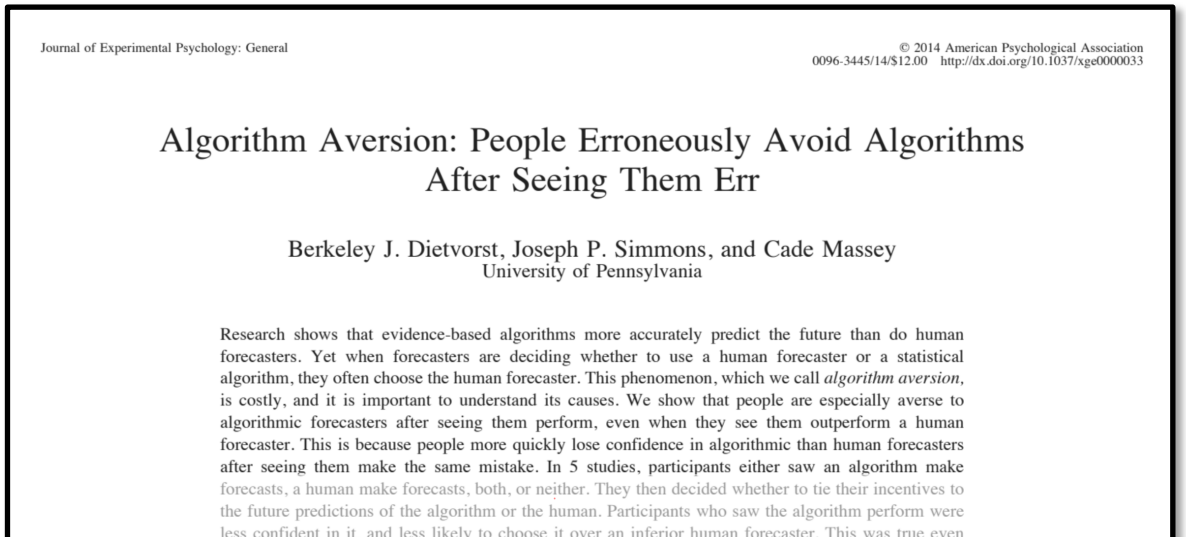
Evaluation of Decision Making for the Meta-Decision

All models are "wrong" in the sense that, by definition, they are abstractions of reality.

Intuition and non-quantitative methods are also models.

The only question is which model is *less wrong?*





Journal of Experimental Psychology: General

© 2014 American Psychological Association
0096-3445/14/$12.00 http://dx.doi.org/10.1037/xge0000033

Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err

Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey
University of Pennsylvania

Research shows that evidence-based algorithms more accurately predict the future than do human forecasters. Yet when forecasters are deciding whether to use a human forecaster or a statistical algorithm, they often choose the human forecaster. This phenomenon, which we call *algorithm aversion*, is costly, and it is important to understand its causes. We show that people are especially averse to algorithmic forecasters after seeing them perform, even when they see them outperform a human forecaster. This is because people more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake. In 5 studies, participants either saw an algorithm make forecasts, a human make forecasts, both, or neither. They then decided whether to tie their incentives to the future predictions of the algorithm or the human. Participants who saw the algorithm perform were less confident in it, and less likely to choose it over an inferior human forecaster. This was true even

Question:
What challenges of quantitative methods are *alleviated* by unaided intuition, non-quantitative or pseudo-quantitative methods?

Measurement Misconceptions

CONCEPT of Measurement	The definition of measurement itself is widely misunderstood.
OBJECT of Measurement	The thing being measured is not well defined.
METHOD of Measurement	Many procedures of empirical observation are misunderstood.

Misconceptions About Statistical Inference

There are widely held misconceptions about probabilities and statistics – especially if they vaguely remember some college stats.

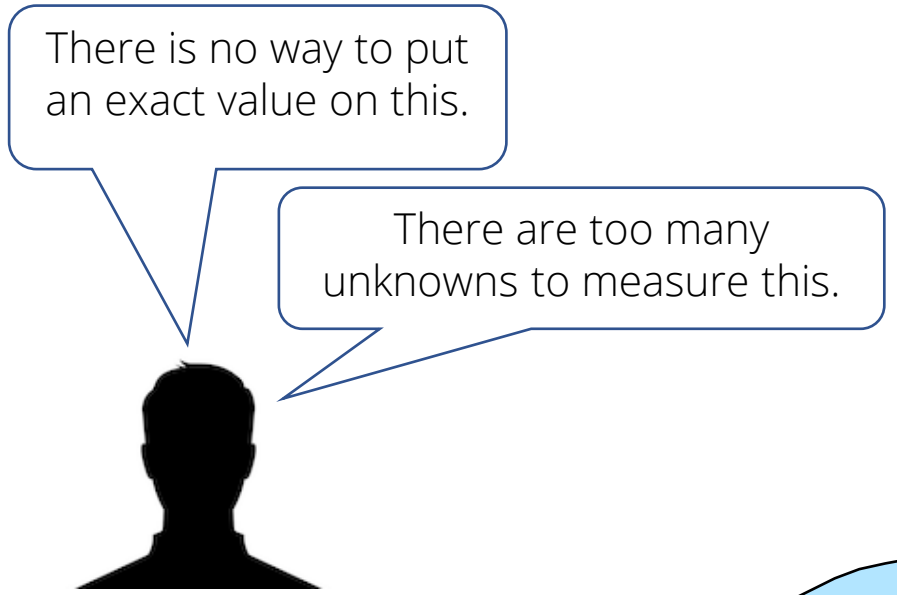
These misconceptions lead many experts to believe they lack data for assessing uncertainties or they need some ideal amount before anything can be inferred.

"Our thesis is that people have strong intuitions about random sampling…these intuitions are wrong in fundamental respects…[and] are shared by naive subjects and by trained scientists"

Amos Tversky and Daniel Kahneman, Psychological Bulletin, 1971

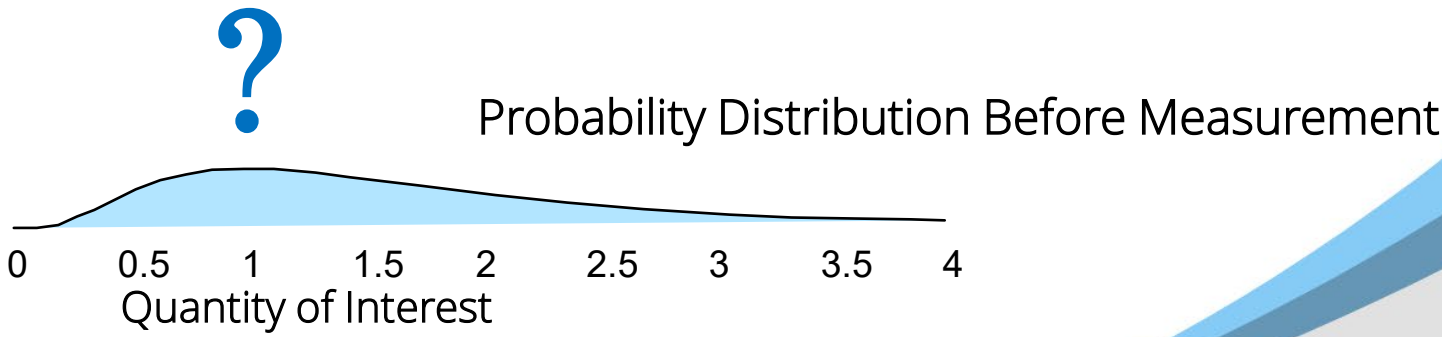
The Concept of Measurement

What Measurement Really Means



It's not a point value.

Measurement: a quantitatively expressed reduction in uncertainty based on observation.

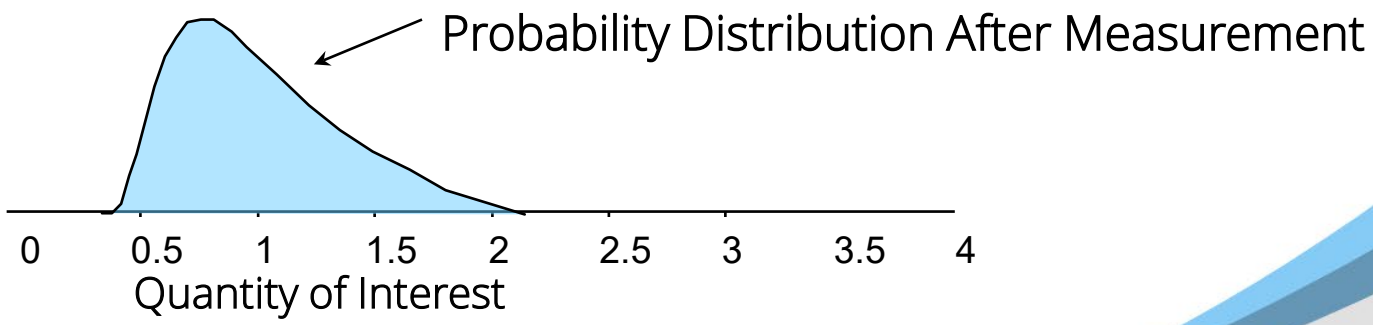


The Concept of Measurement

What Measurement Really Means

It's not a point value.

Measurement: a quantitatively expressed reduction in uncertainty based on observation.

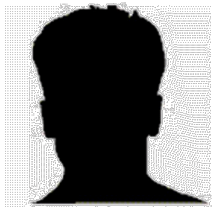


Overconfidence

The Need to be "Calibrated"

There is an 80% chance we will win this contract.

There is a 70% chance the competitor's product will get to the market before ours.

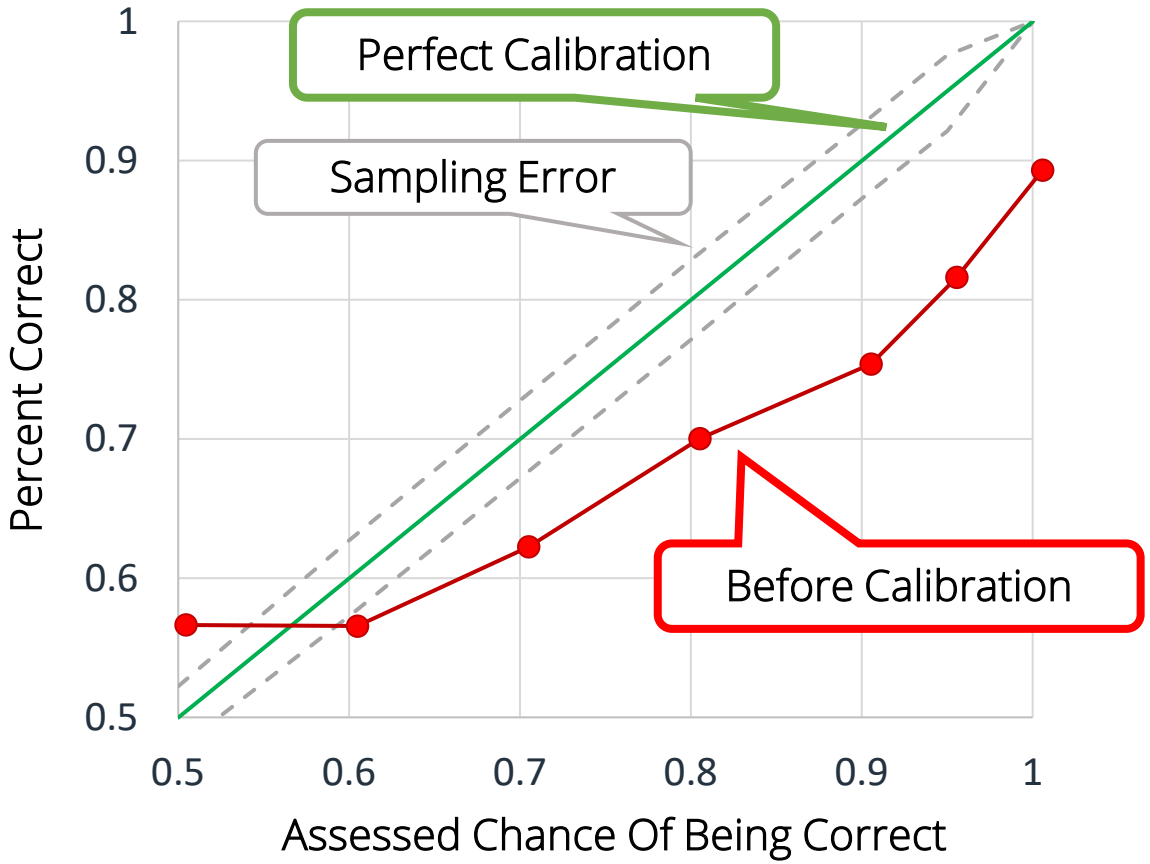


"Overconfident professionals sincerely believe they have expertise, act as experts and look like experts. You will have to struggle to remind yourself that they may be in the grip of an illusion."

Daniel Kahneman,
Psychologist, Economics Nobel

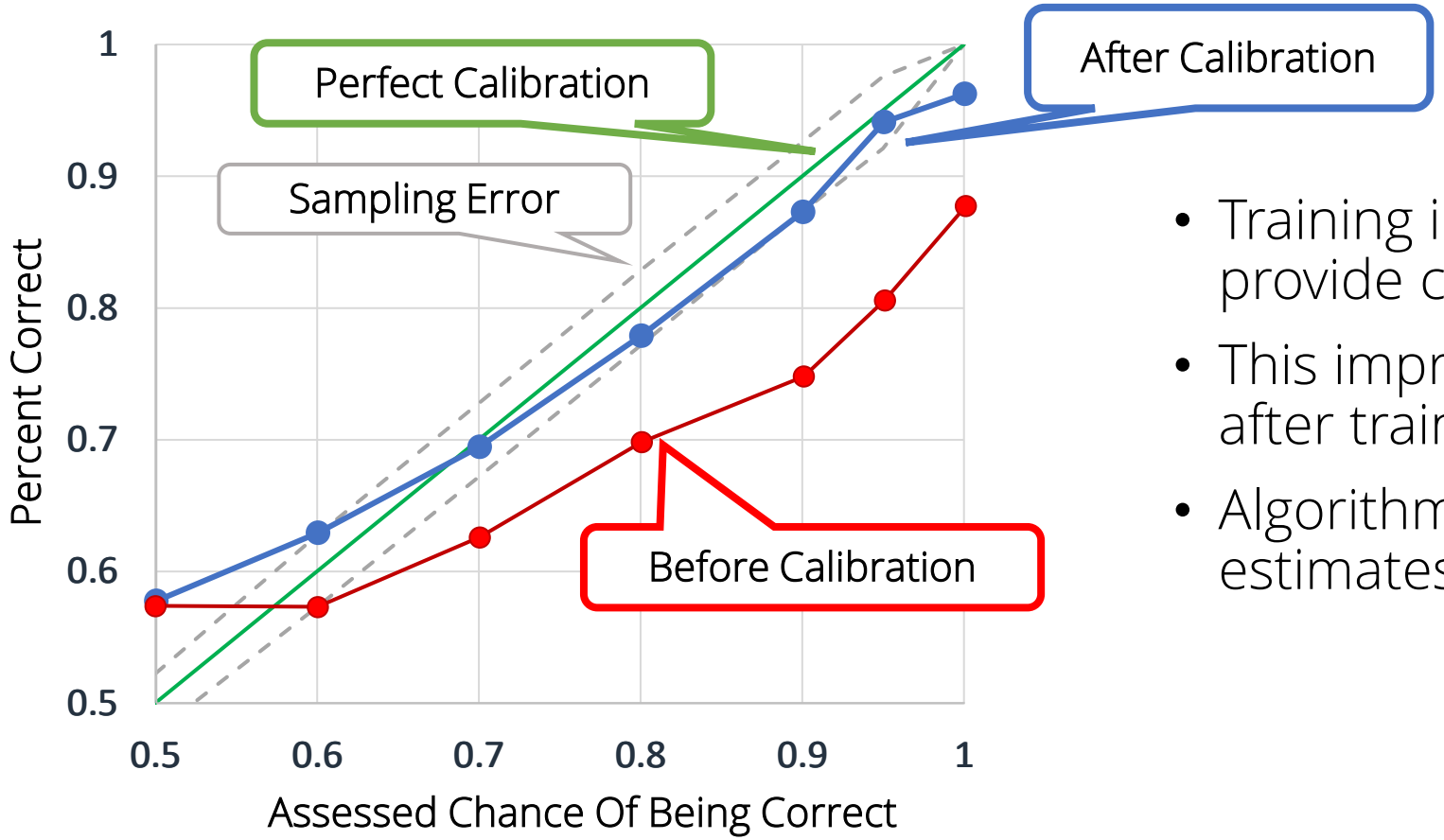
Studies also show that measuring *your own* uncertainty about a quantity is a general skill that <u>can be taught</u> with a ***measurable*** improvement.

Measuring Overconfidence



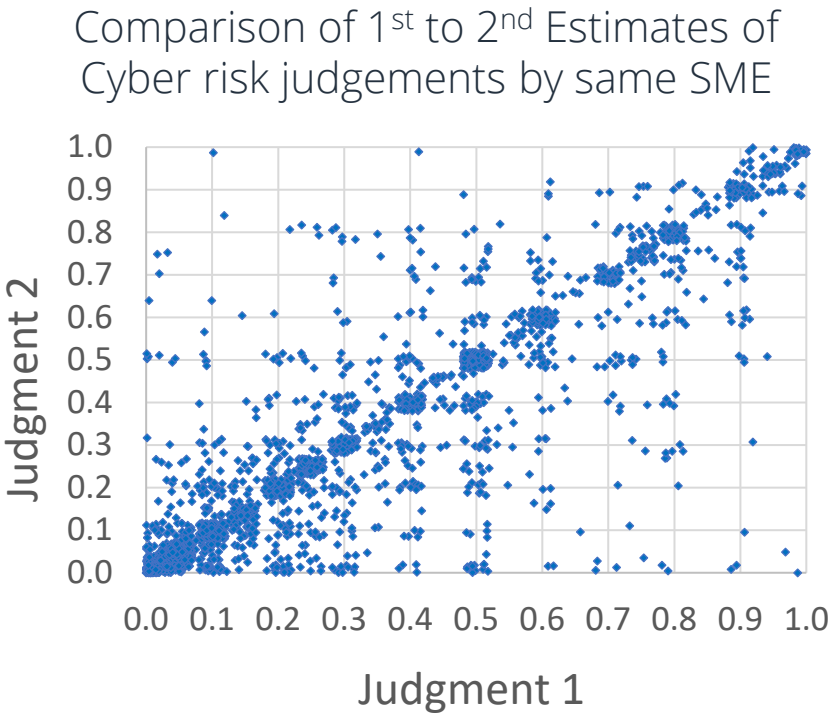
- We've trained over 2,000 individuals in subjective estimation of probabilities.
- Almost everyone is overconfident on the first benchmark test.

Measuring Calibration Training



- Training improves the ability to provide calibrated estimates.
- This improves real-world estimates after training is complete.
- Algorithms can adjust subjective estimates to further improve them.

Calibrating Expert Consistency

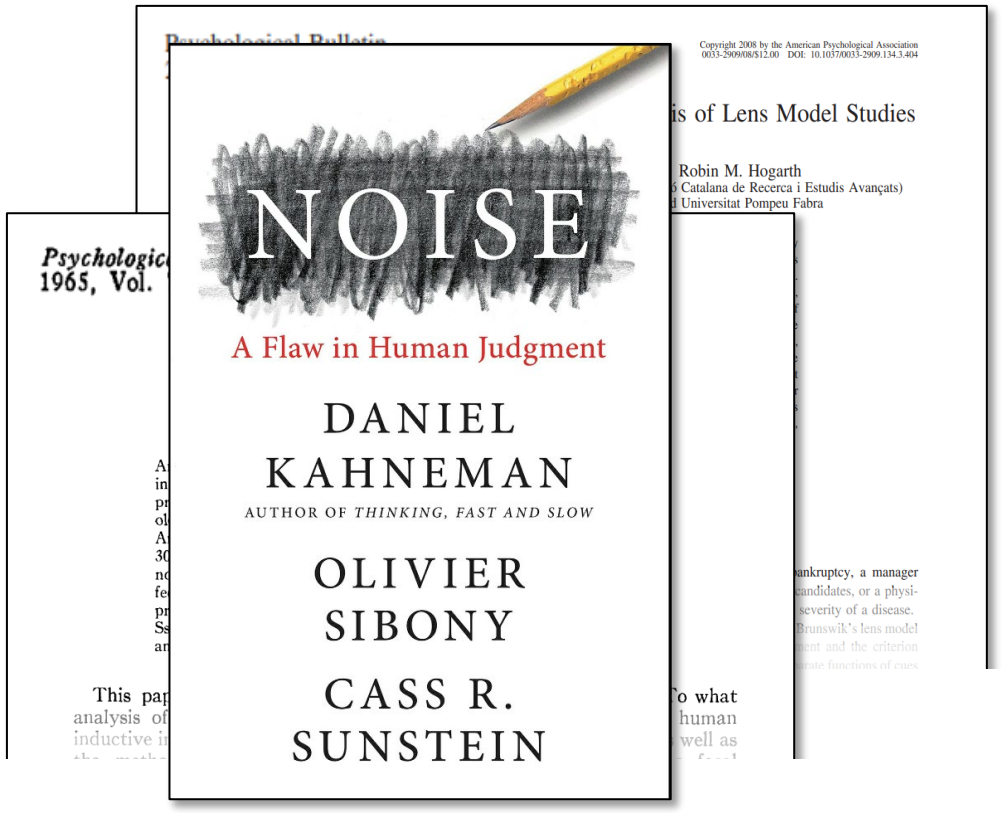


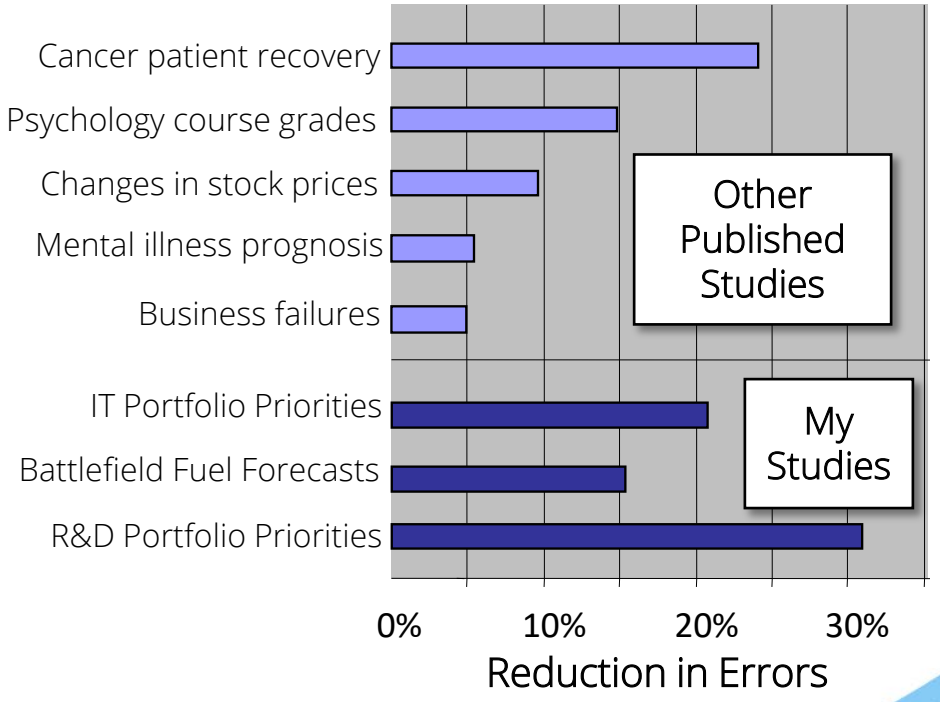
21% of variation in expert responses are explained by *inconsistency.*
(79% are explained by the actual information they were given)

- We have gathered over 30,000 individual estimates of probabilities of events from analysts in multiple organizations.
- These estimates included over 2,000 duplicate scenarios pairs.

Removing Inconsistency

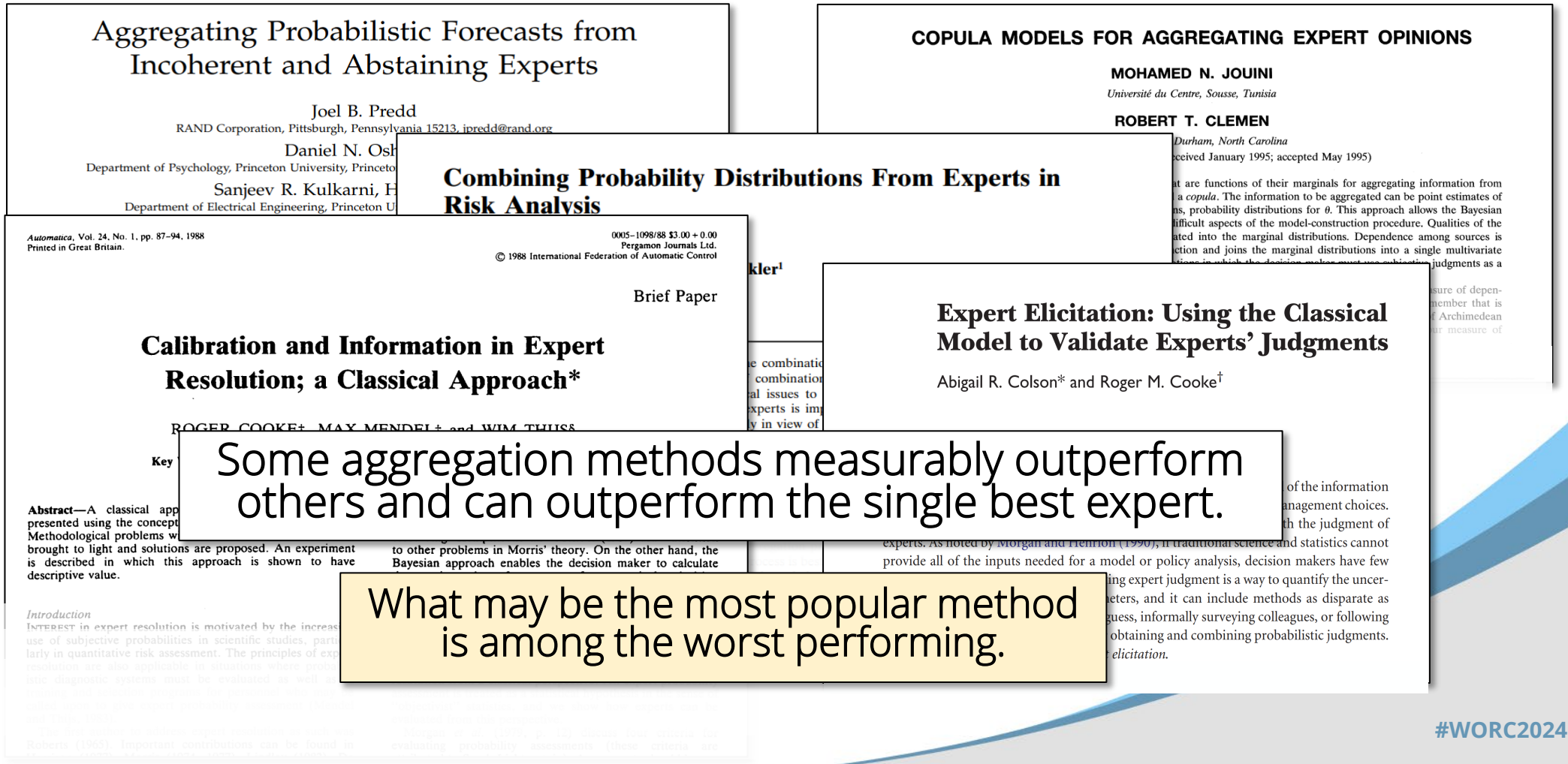
The "Lens Method" statistically "smooths" estimates of experts. Several studies for many different kinds of problems show it reduces judgement errors.



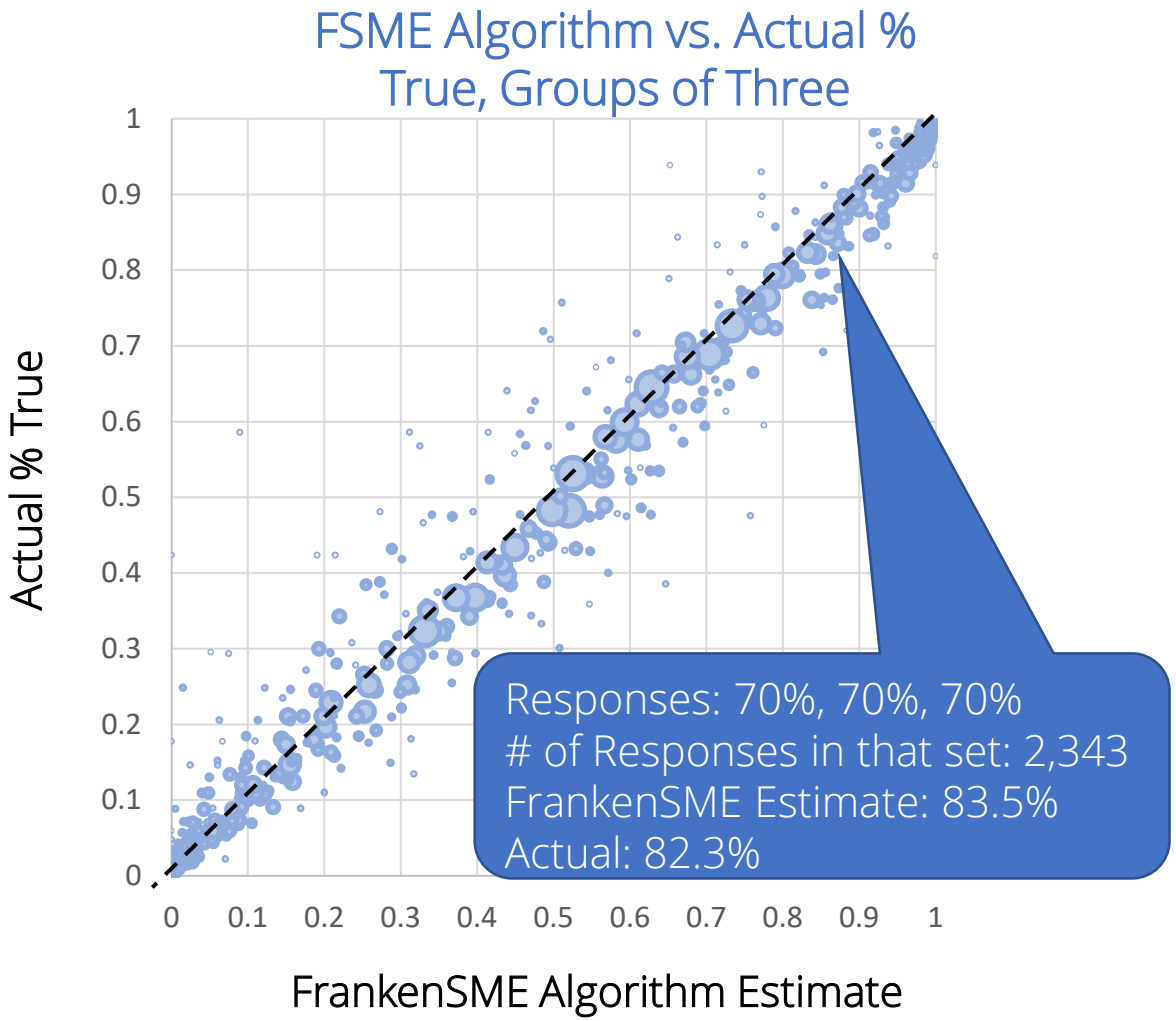


Aggregating Experts

A Lot of Research



Combining Experts: The FrankenSME

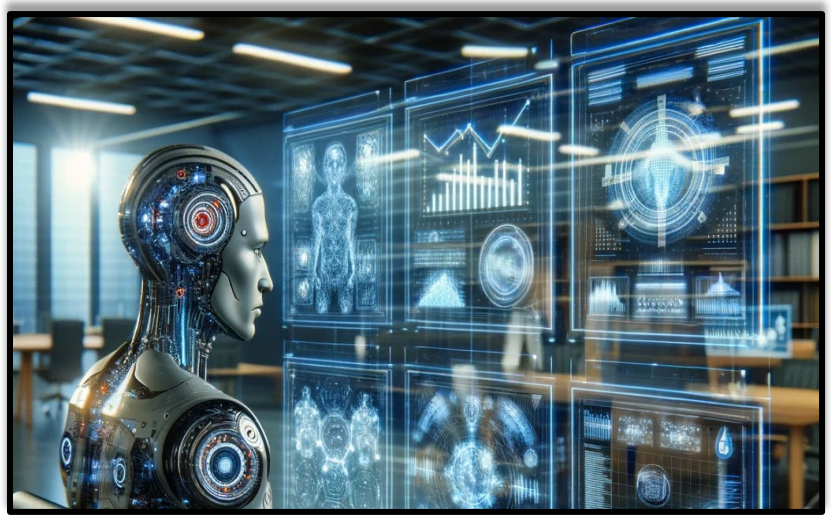


HDR has algorithms for combining experts using data from over 60,000 responses from 977 calibrated individuals grouped into 1.8 million virtual teams.

Examples of Groups of Five

Responses	Count	FSME	Actual
60%, 60%, 60%, 70%, 70%	2825	85%	86%
40%, 60%, 60%, 60%, 60%	913	67%	66%
20%, 30%, 30%, 40%, 60%	364	6%	5%

Calibrating Chat GPT Responses



ChatGPT was asked to provide 90% CI's for events like: "How much will the top-grossing film earn internationally at the box office in 2022?"

For our analyst's blog on this: https://hubbardresearch.com/is-chatgpt-as-overconfident-as-humans/

	% W/in Bounds	Sample Size (events or trivia, humans or AI sessions)
Uncalibrated Humans, General Trivia	55%	20,000+ (10+ trivia, 2000+ humans)
Calibrated Humans, General Trivia	86%	120,000+ (60+ trivia, 2000+ humans)
ChatGPT 3.5	13.5%	140 (20 events X 7 sessions)
ChatGPT 4, Temperature=1	60%	360 (20 events X 18 sessions)
ChatGPT 4, Temperature=0, Before Adjustment	64.5%	62 (62 events X 1 sessions)
Chat GPT 4, Temperature=0, After Adjustment	89.6%	31 (31 events X 1 session) trained w/k-folds

Practical Lessons

Here are a few key things I've learned measuring the "immeasurable"

- You *have more* data than you think and you *need less* data than you think.
- It's been measured before.
- Your probably need *different* data than you think.
- Decision makers understand it just fine if explained well.
- The best investment in most portfolios was better measurements of investments.

Thank you for Your Time!

Questions?

Doug Hubbard
Hubbard Decision Research
dwhubbard@hubbardresearch.com
www.hubbardresearch.com

Measure What Matters.

Make Better Decisions.

WORC 2024

World Overflight Risk Conference

2 - 4 July 2024 | Warsaw, Poland

#WORC2024

PRESENTED BY

OSPREY FLIGHT SOLUTIONS

EASA European Union Aviation Safety Agency