

# Subjective Judgements: Outperforming Your Current Best Experts

Doug Hubbard  
Hubbard Decision Research

# Addressing Judgement Problems

## Overconfidence

I'm right 95% of the time.



## Inconsistency

(Before coffee)  $P(X)=.05$   
(After coffee)  $P(X)=.2$



## Inefficient Collaboration

This control is 95% effective.

I agree.



## Slow Feedback

20 years.

## Inference Errors

So I have no data.

## Analysis Placebos

Working great.

Track Performance

Use outside data (both the object and method of measurement)

Do less math in our heads

# The Analysis Placebo

*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.

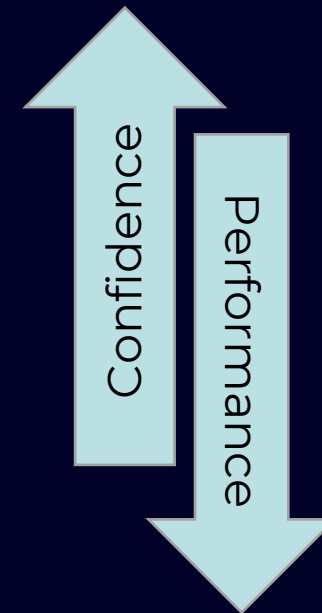
## **Interaction with Others Increases Decision Confidence but Not Decision Quality: Evidence against Information Collection Views of Interactive Decision Making**

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

When evaluating judgement methods, you can't rely on perceptions of effectiveness.



# The Need for Feedback

Experience does not automatically produce learning.

And that feedback  
has to be  
*CONSISTENT...*

*...IMMEDIATE...*

*...and  
UNAMBIGUOUS.*

To learn from  
experience, you  
need feedback.



Daniel Kahneman

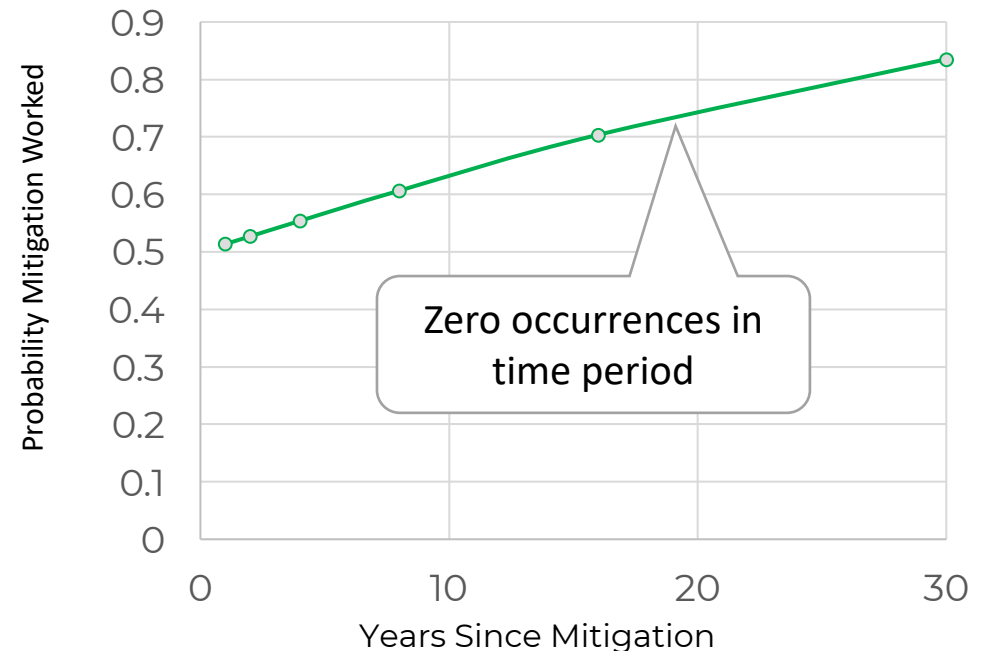


Gary Klein

# The (Very) Slow Feedback Problem

## A Bayesian Look at Mitigation Assessment Over Time

- Suppose we have an event we assess as having a 10% chance/yr of occurrence.
- We implement a mitigation that we think may reduce that chance to 5%.
- Uncertain of whether the risk will actually be reduced, we give a prior probability that there is a 50% chance the mitigation works as stated.
- How long do we have to watch our environment to see if the annualized probability went from 10% to 5%?



Solving for the probability a mitigation reduced event likelihood from 10% to 5% per year given number of occurrences in time period

# Calibrated Experts

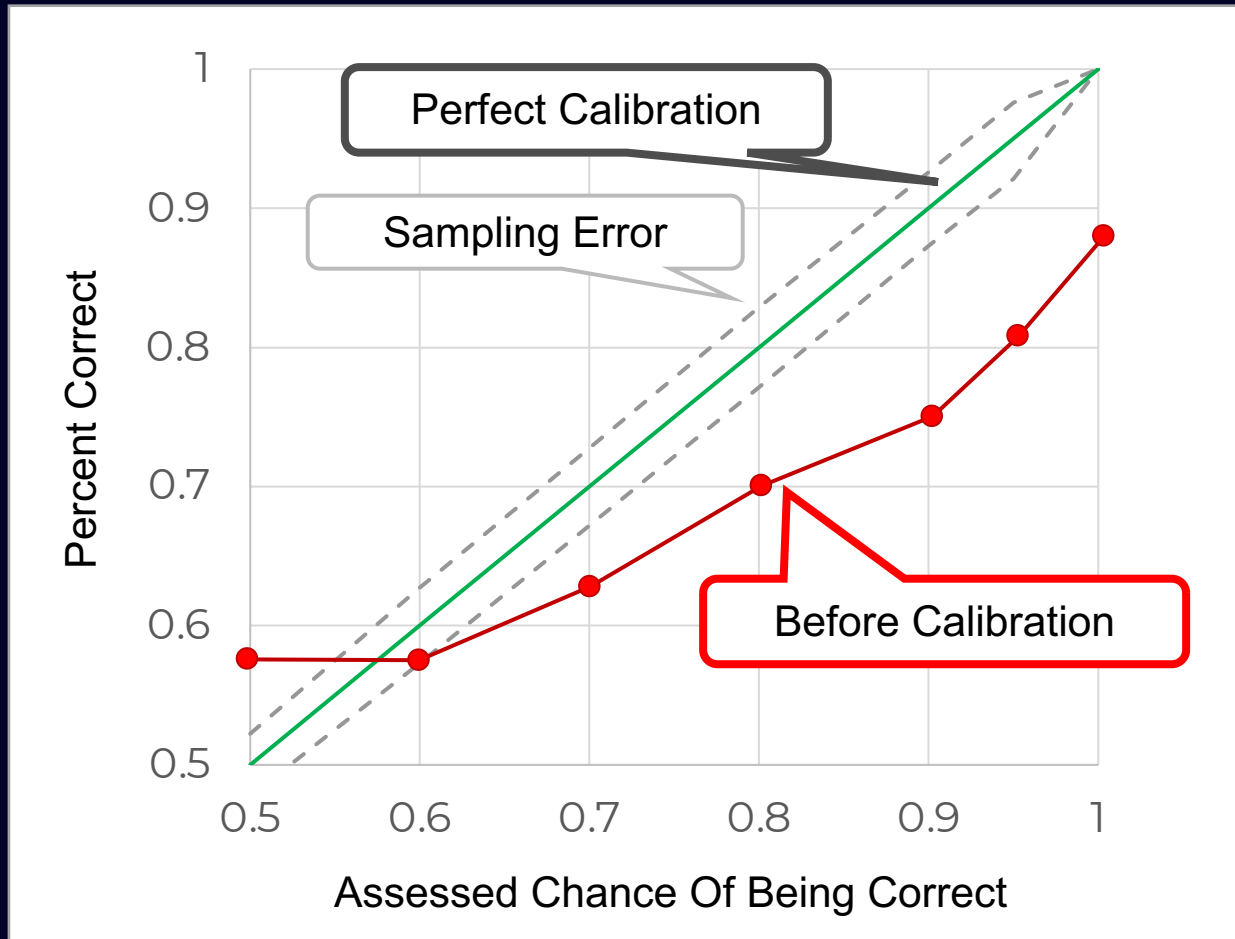
“Overconfident professionals sincerely believe they have expertise, act as experts and look like experts. You will have to struggle to remind yourself that they may be in the grip of an illusion.”



Daniel Kahneman, Psychologist, Economics Nobel

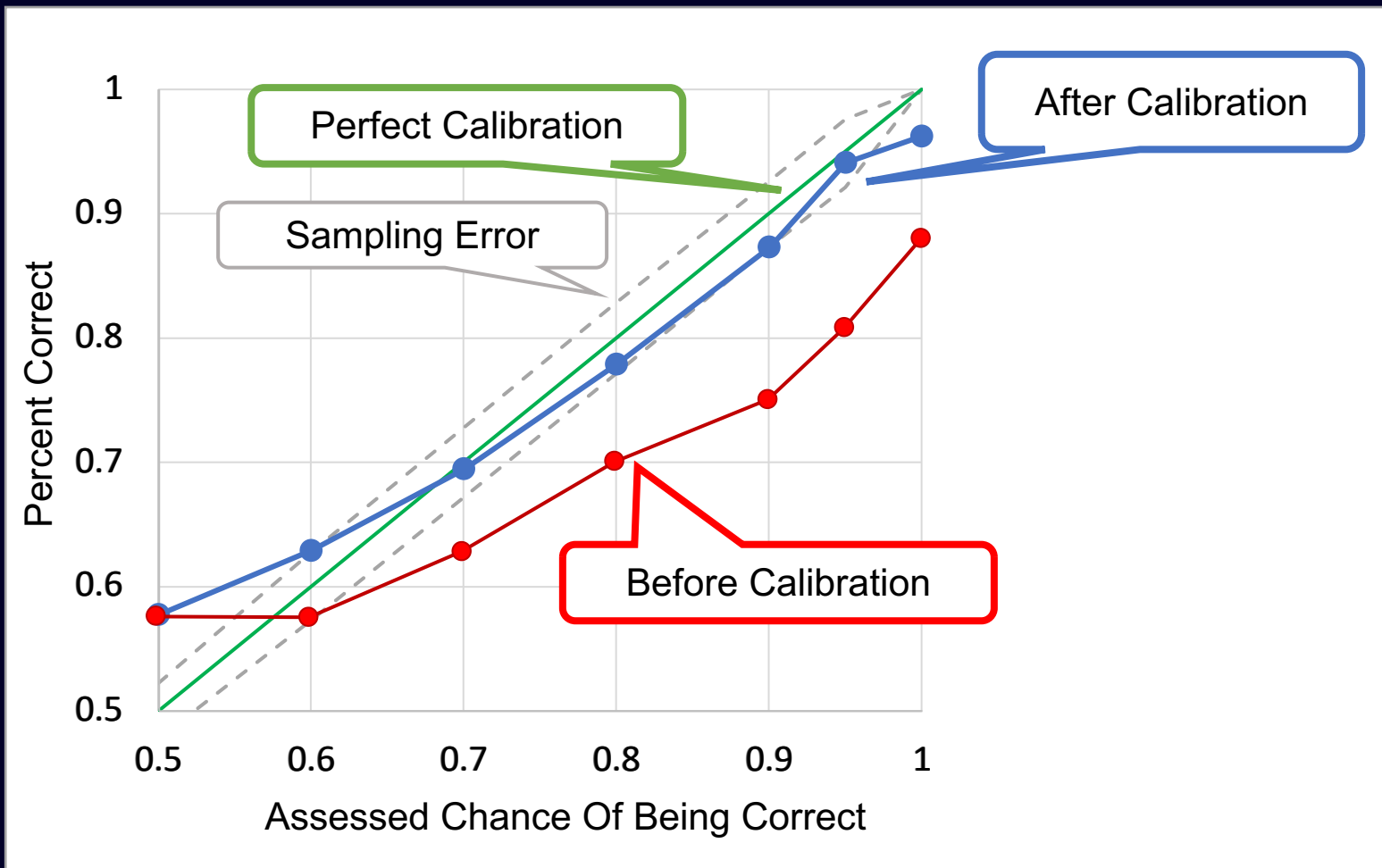
- Decades of studies show that most managers are statistically “overconfident” when assessing their own uncertainty.
- Studies also show that measuring *your own* uncertainty about a quantity is a general skill that can be taught with a **measurable** improvement.

# Measuring Overconfidence



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

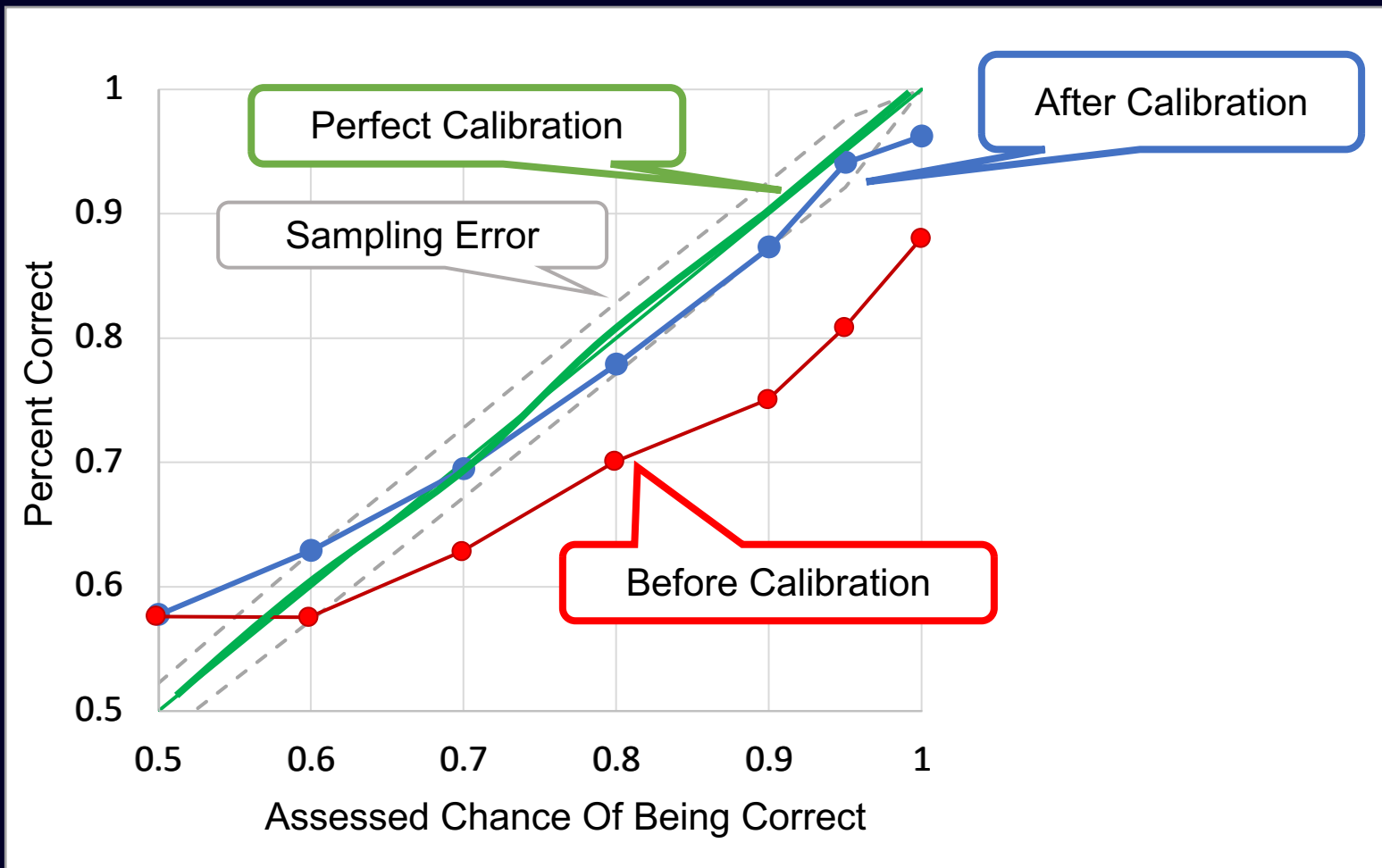
# Measuring Calibration Training



- Training improves the ability to provide calibrated estimates.
- We've done experiments which shows this training improves real-world estimates.



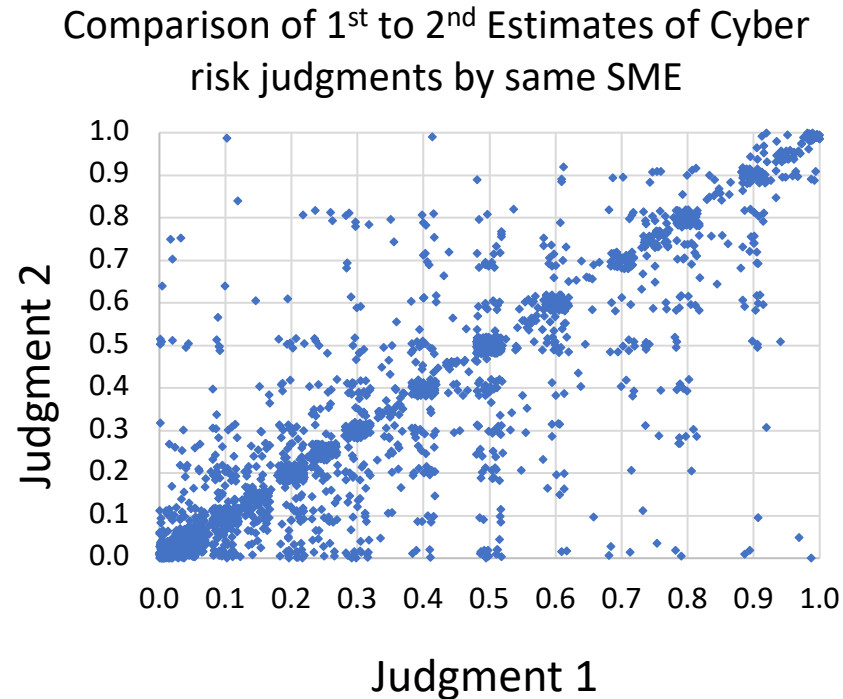
# Final Calibration Adjustments



- Further adjustments using actual performance data for individual SMEs will make them nearly perfectly calibrated.

# Measuring and Reducing Judgement Inconsistency

- We have gathered over 30,000 estimates of probabilities of various security events.
- These estimates included over 2,000 duplicate scenarios pairs.

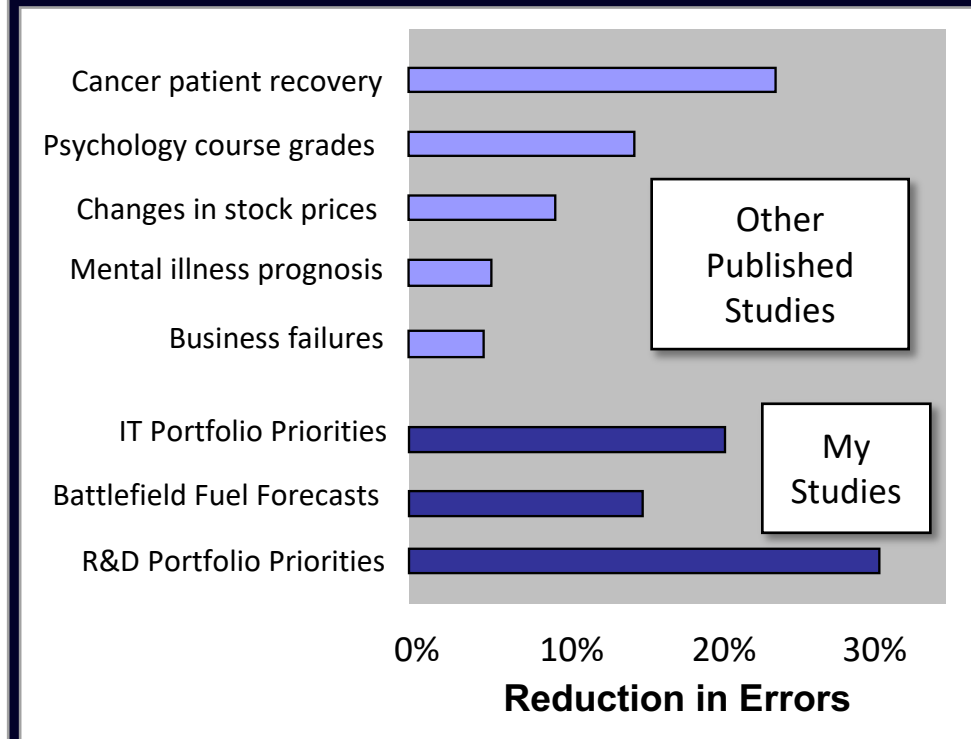
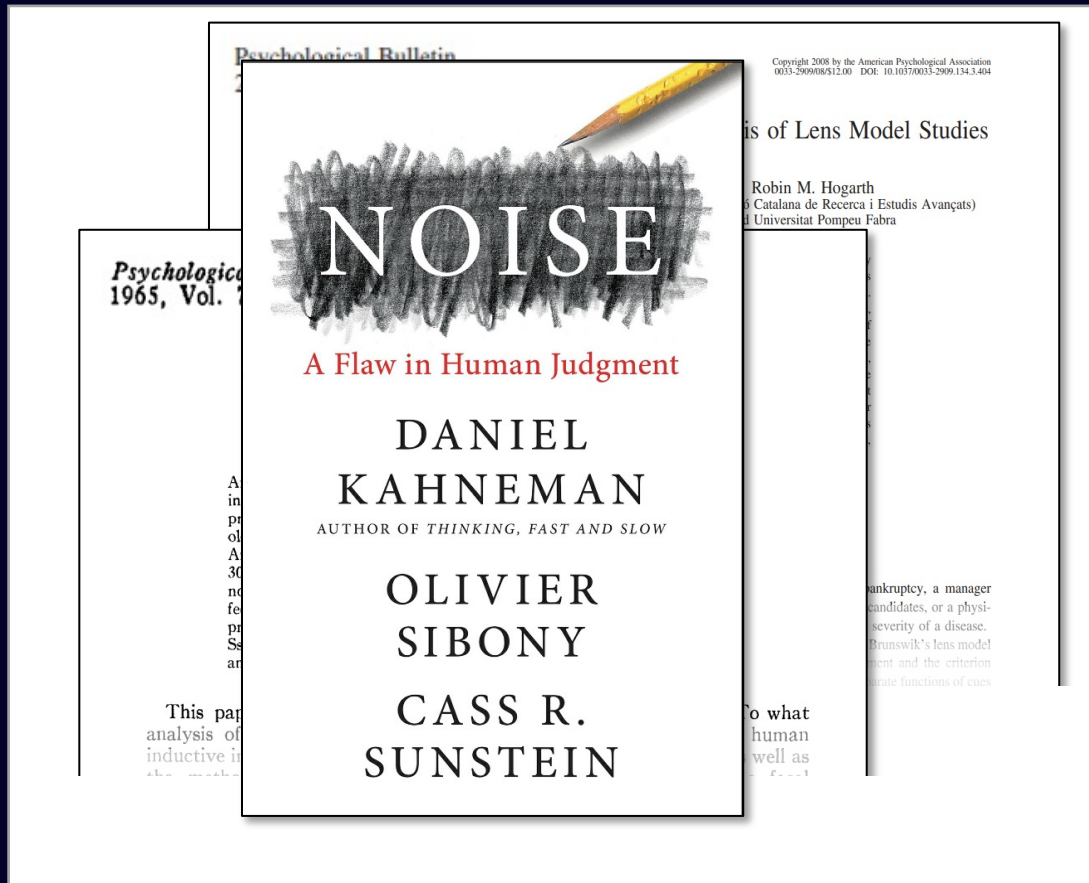


**21% of variation in expert responses are explained by *inconsistency*.**

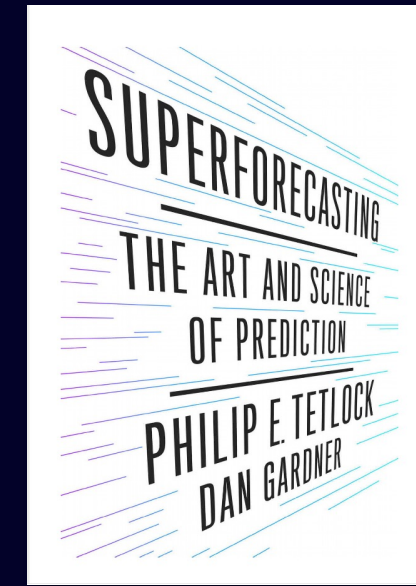
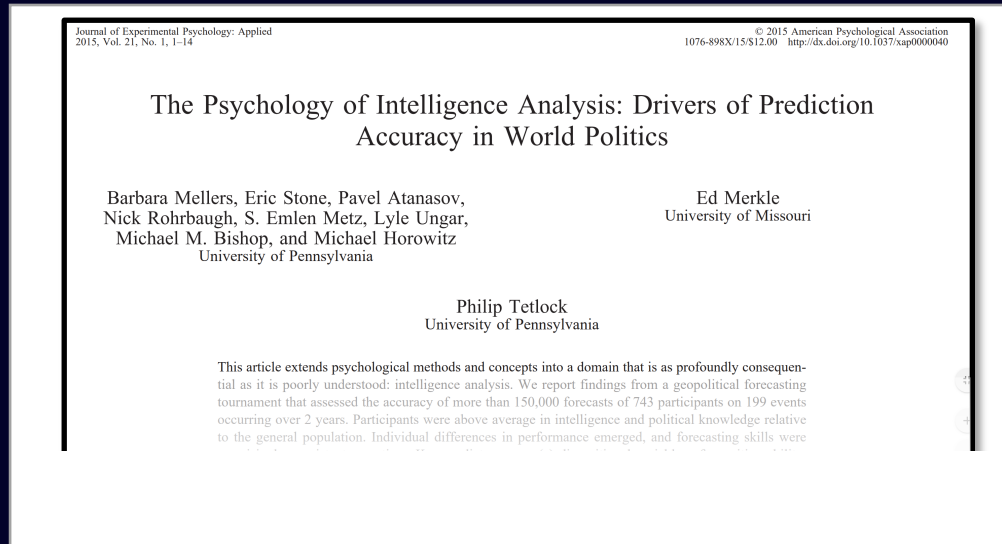
(79% are explained by the actual information they were given)

# Measuring (And Removing) Inconsistency

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

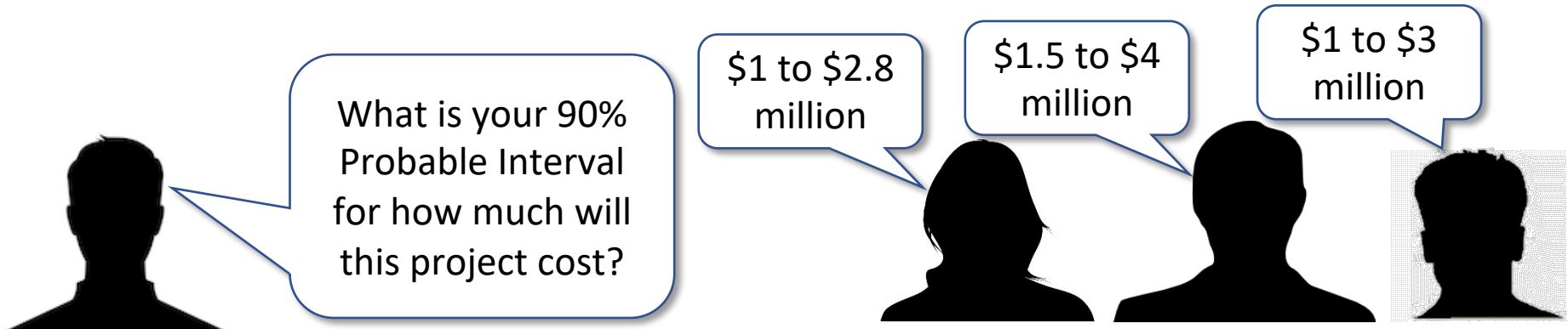
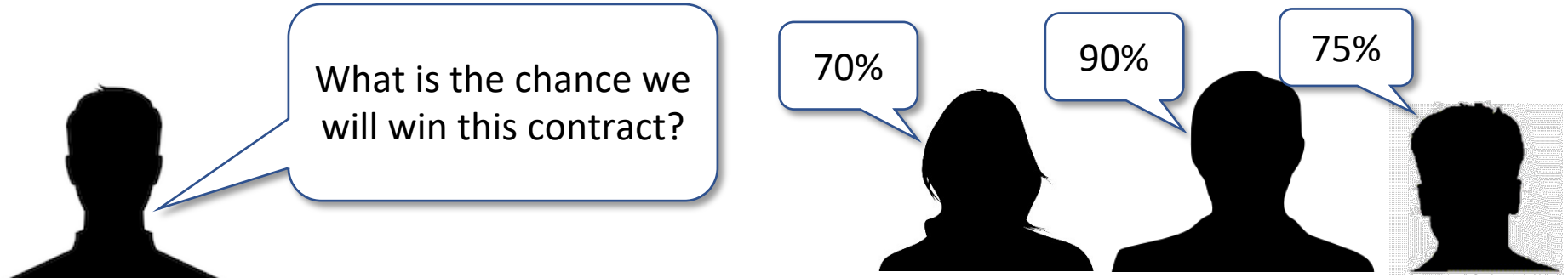


# Phil Tetlock's "SuperForecasting" Research



- **Training:** Subjects were trained in basic inference methods, using reference classes, and avoiding common errors and biases.
- **Teams of "Belief Updaters":** Teams deliberated more. But must comprise individuals who were willing to update beliefs based on new information.
- **Tracking Who is Better:** Some just had a knack for it. IQ mattered (a little).

# Do You Ask Multiple SMEs?



# Moneyball: Creating a Star Player in the Aggregate

In 2001, the Oakland A's lost their star player, Jason Giambi, to the NY Yankees. The Yankees paid \$120 million for Giambi but the A's needed an economical replacement. After hearing scouts pitch various players, the manager, Billy Beane, explained a different strategy.

Scene from Moneyball



Billy Beane: "Guys, you're still trying to replace Giambi. I told you we can't do it. ...Now, what we might be able to do is recreate him. We create him in the aggregate."

The "FrankenSME" is a way of making a team which outperforms even the best individuals.

# Aggregating Experts

Aggregating Probabilistic Forecasts from Incoherent and Abstaining Experts

COPULA MODELS FOR AGGREGATING EXPERT OPINIONS

MOHAMED N. JOUINI  
Université du Centre, Sousse, Tunisia

Combining Probability Distributions From Experts in Risk Analysis

Expert Elicitation: Using the Classical Model to Validate Experts' Judgments

Abigail R. Colson\* and Roger M. Cooke†

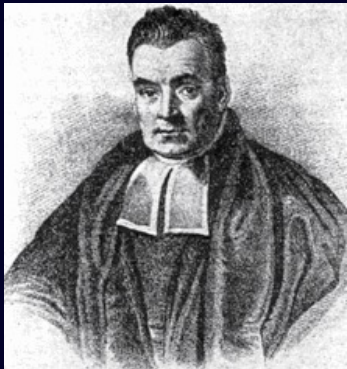
Calibration and Information in Expert Resolution; a Classical Approach\*

Some aggregation methods measurably outperform others and can outperform the single best expert.

What may be the most popular method is among the worst performing.

# Combining Data With Bayes

Bayes Theorem is a simple and powerful concept. It allows us to update our “prior” probabilities with new information – and combine the priors of experts.



Thomas Bayes, 1701-1761

$$\text{Bayes Theorem: } P(X|Y) = \frac{P(X)P(Y|X)}{P(Y)} = \frac{P(X)P(Y|X)}{\sum_i P(Y|X_i) P(X_i)}$$

$P(X)$  = the probability of X

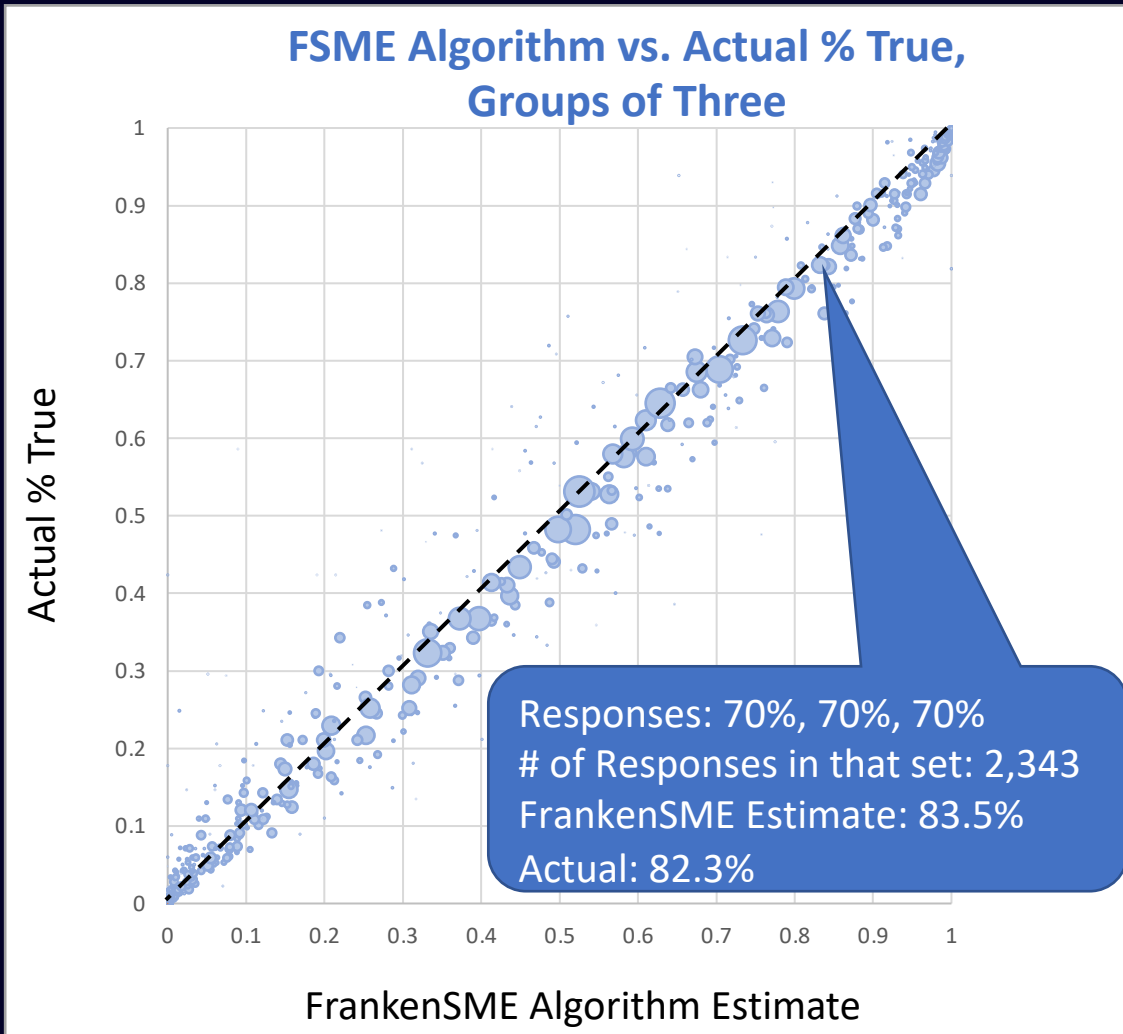
$P(X|Y)$  = the probability of X given the condition Y

$\sum P(Y | X_i) P(X_i)$  = the sum of the probability of Y under each possible condition

$$\frac{P(X|C_1 \dots C_n)}{1 - P(X|C_1 \dots C_n)} = \left( \frac{1 - P(X)}{P(X)} \right)^{n-1} \prod_{i=1}^n \frac{P(X|C_i)}{1 - P(X|C_i)}$$



# Combining Experts: The FrankenSME



- HDR has algorithms for combining experts using data from over 60,000 responses from 577 calibrated individuals grouped into 1.7 million 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%

# The Brier Score

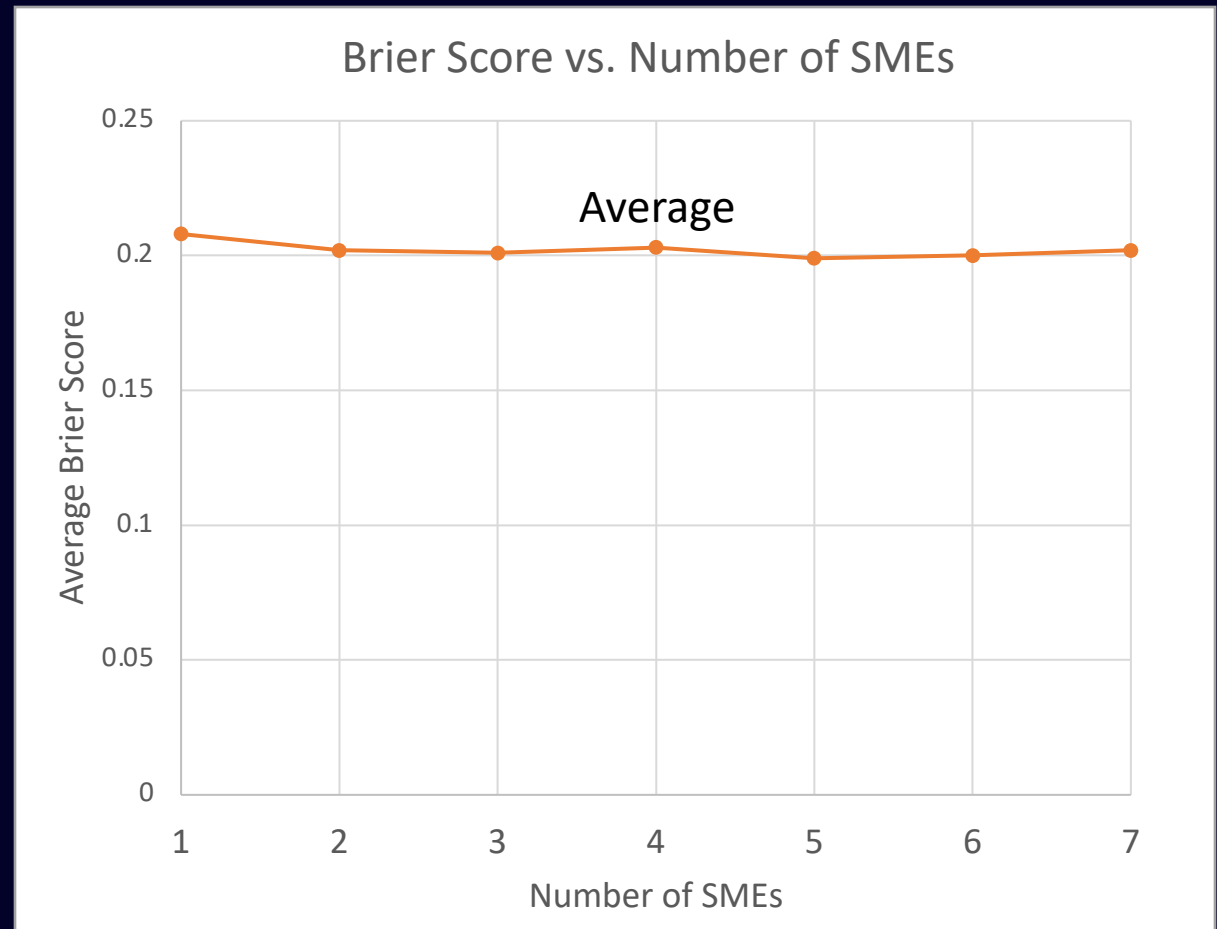
- The “Brier Score” is a way of evaluating estimates of probabilities.
- Lower is better. Lowest points are when you are certain and right, highest are when you are certain and wrong. Uncertain is in between.

Confidence	Result	Score
0.9	Correct	0.01
0.6	Correct	0.16
0.6	Incorrect	0.36
0.9	Incorrect	0.81

$$\text{Brier} = (\text{Subjective Probability} - \text{Truth})^2$$

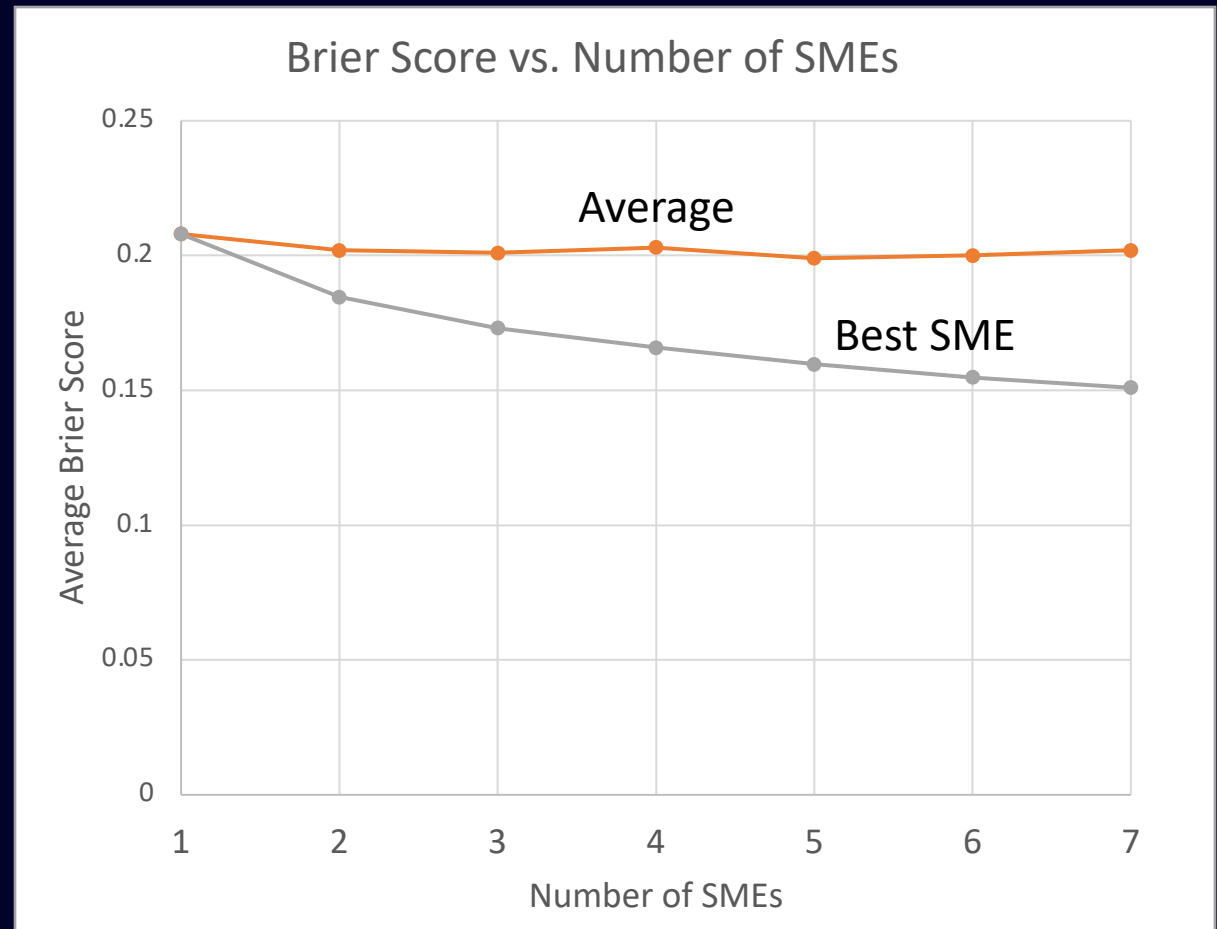
# The Value of an Additional Estimator (Binary)

- Applying the Binary FrankenSME to 10,000 samples for groups of 2-7
- Averaging doesn't improve the Brier Score much.



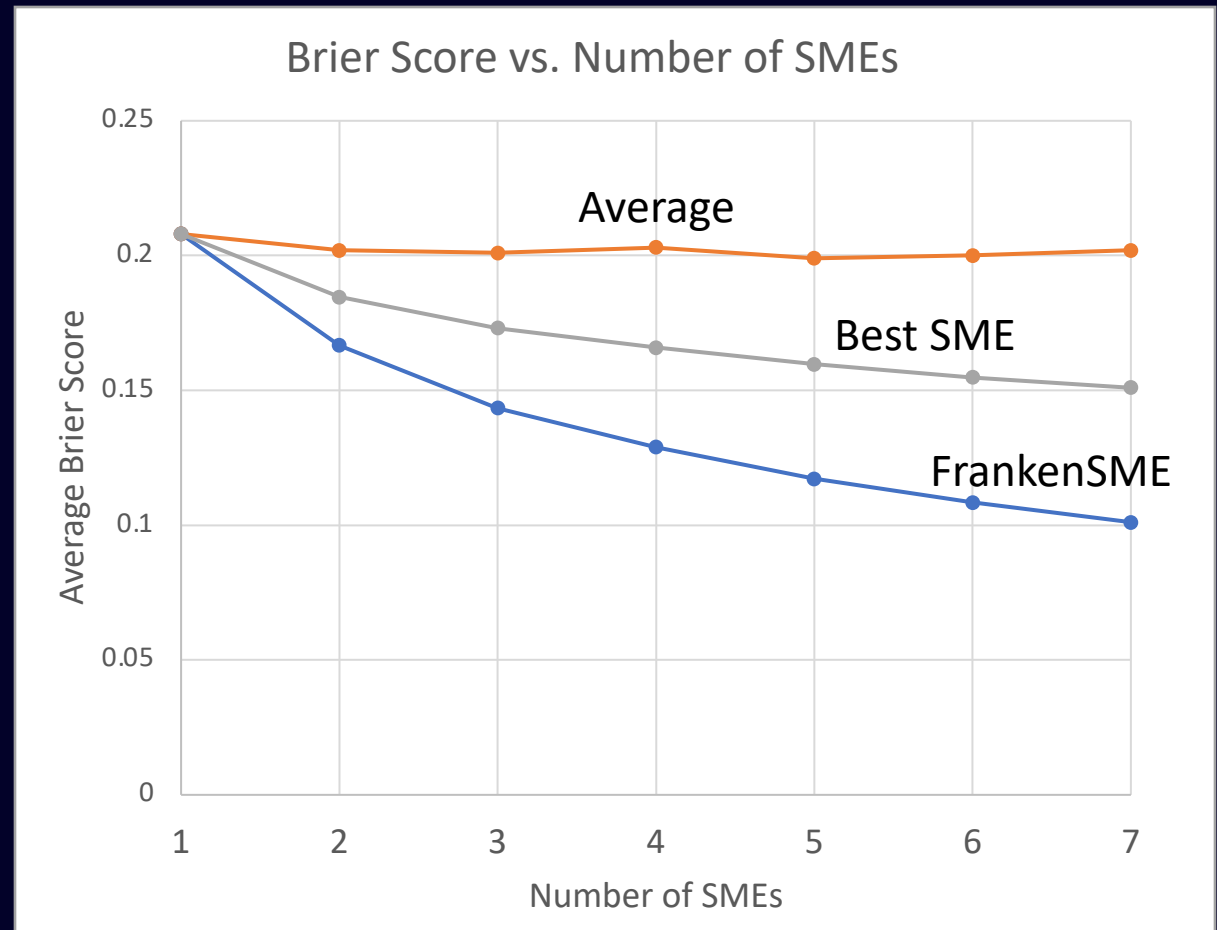
# The Value of an Additional Estimator (Binary)

- Applying the Binary FrankenSME to 10,000 samples for groups of 2-7
- Averaging doesn't improve the Brier Score much.
- The best SME out of a group of any size is better than the average of that group.

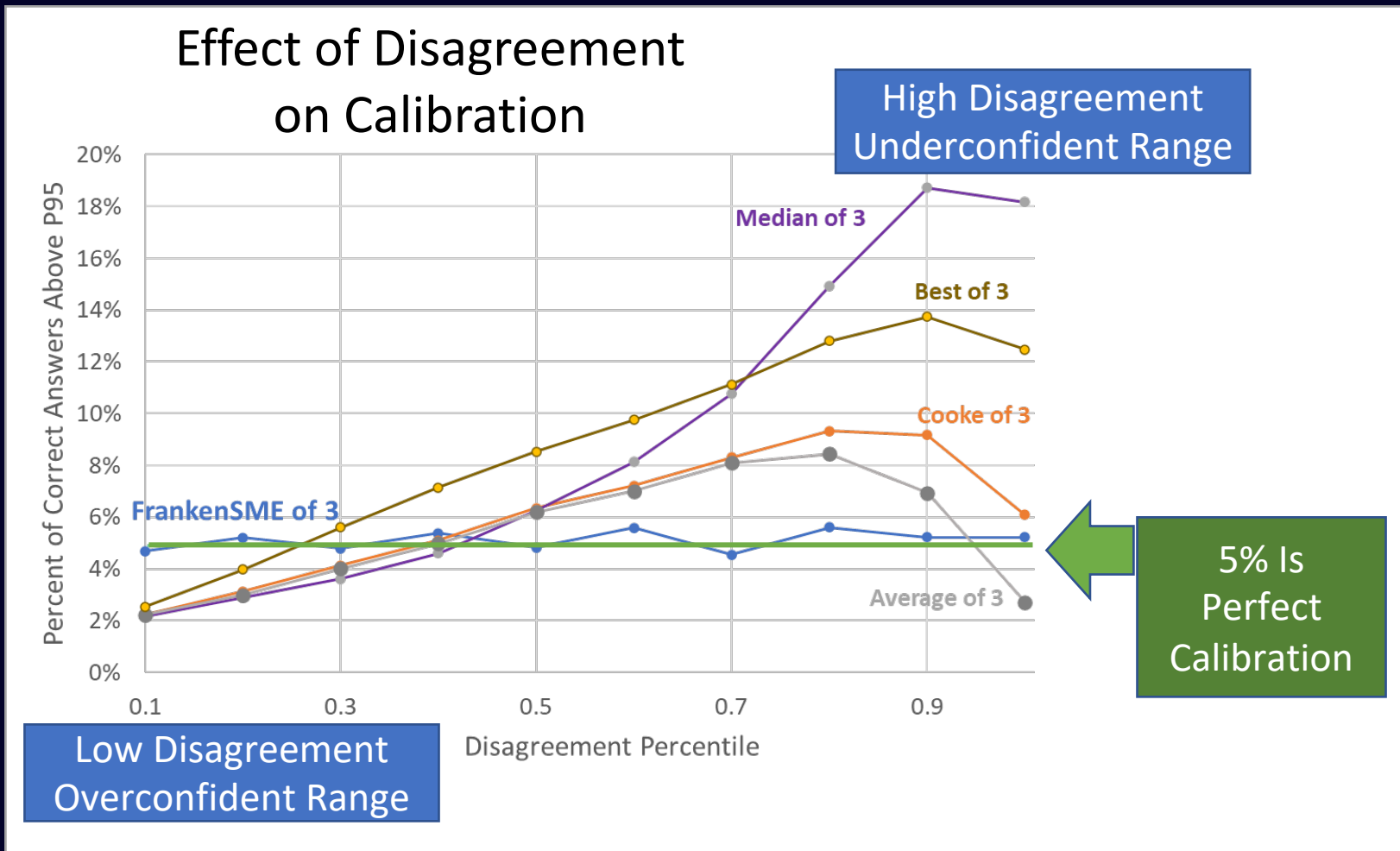


# The Value of an Additional Estimator (Binary)

- Applying the Binary FrankenSME to 10,000 samples for groups of 2-7
- Averaging doesn't improve the Brier Score much.
- The best SME out of a group of any size is better than the average of that group.
- FrankenSME improves the most with more SMEs.



# Calibration Methods vs. Disagreement



- The FrankenSME performs well regardless of level of disagreement among SMEs.

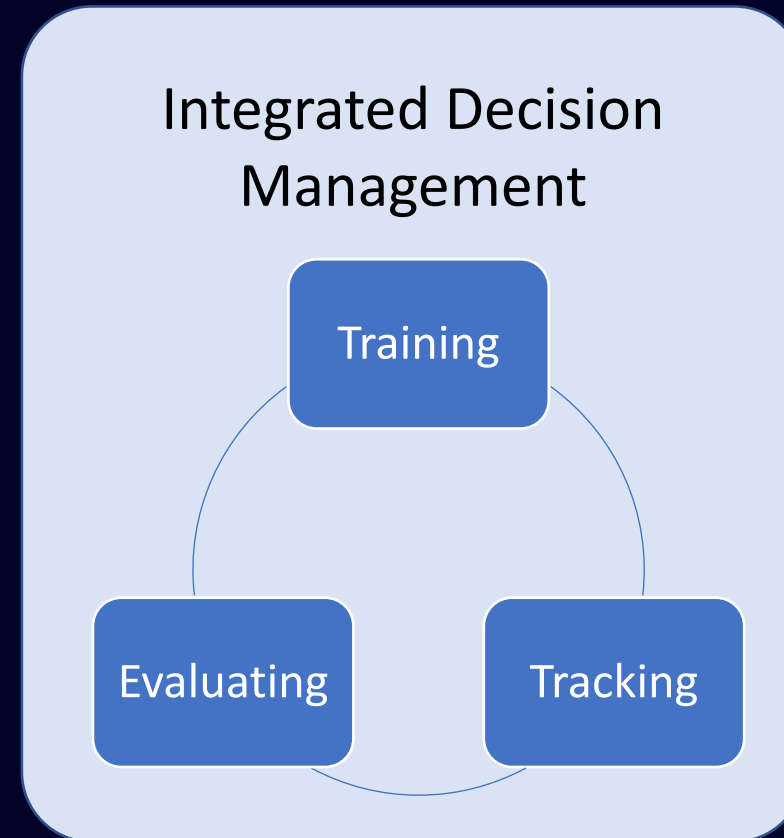
# Improving Expert Judgement

Yes, you can outperform your best current expert.

1. Calibrate SMEs.
2. Where there are repeatable judgements, use Lens methods to reduce the inconsistency error.
3. Create the best SME in the aggregate “FrankenSME” of your team.
4. With both “practical” and “practice” forecasts, measure individuals, teams, and tools and adjust.

# Integrated Decision Management

- The major components of decision making are among the least measured in any organization.
- We need some form of a dashboard for estimation performance including individuals, teams, and tools.





# Thank you for Your Time!

Doug Hubbard

Hubbard Decision Research

[dwhubbard@hubbardresearch.com](mailto:dwhubbard@hubbardresearch.com)

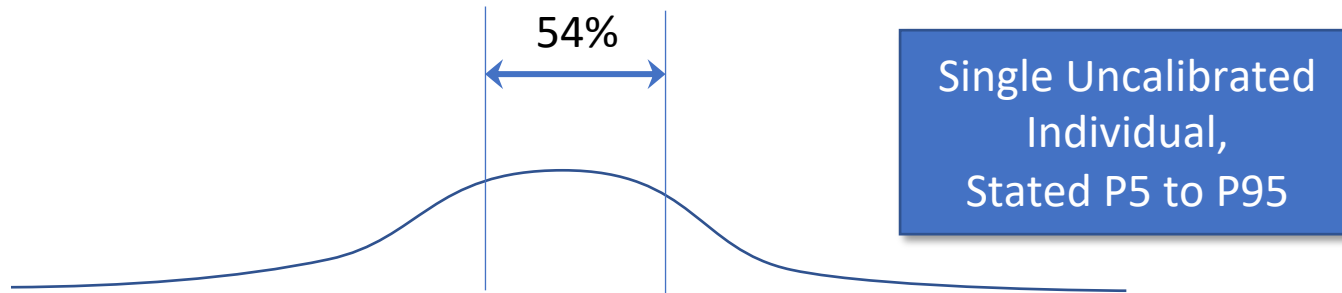
[www.hubbardresearch.com](http://www.hubbardresearch.com)

*Measure What Matters.*

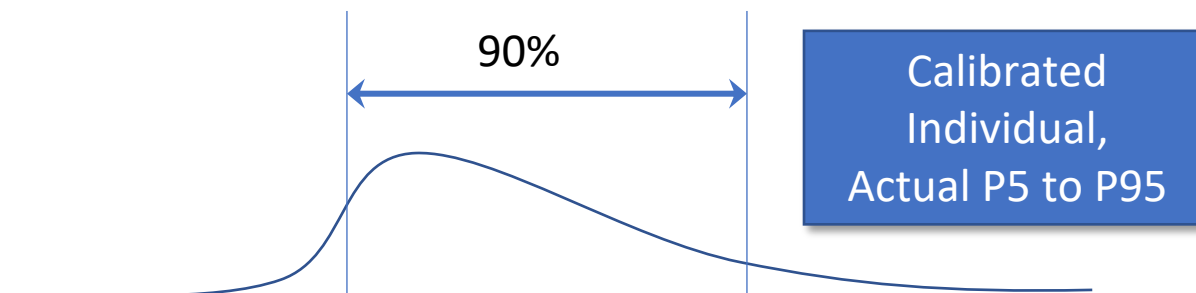
*Make Better Decisions.*

# Supplementary Material

# Uncalibrated & Calibrated Intervals

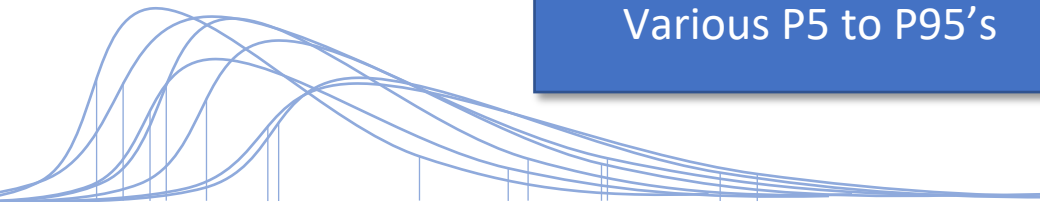


The uncalibrated SME, on average, gets about 54% of correct answers within a stated 90% probable interval.



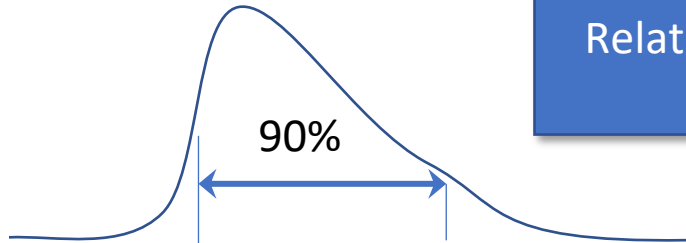
The calibrated SME, on average, gets about 86% of correct answers within the stated interval based on training alone, and 90% after further adjustments.

# Calibrated & Combined Intervals



7 Calibrated SMEs,  
Various P5 to P95's

Groups of SMEs will vary their estimates. Some variation is just personal inconsistency, and some is differences in knowledge.

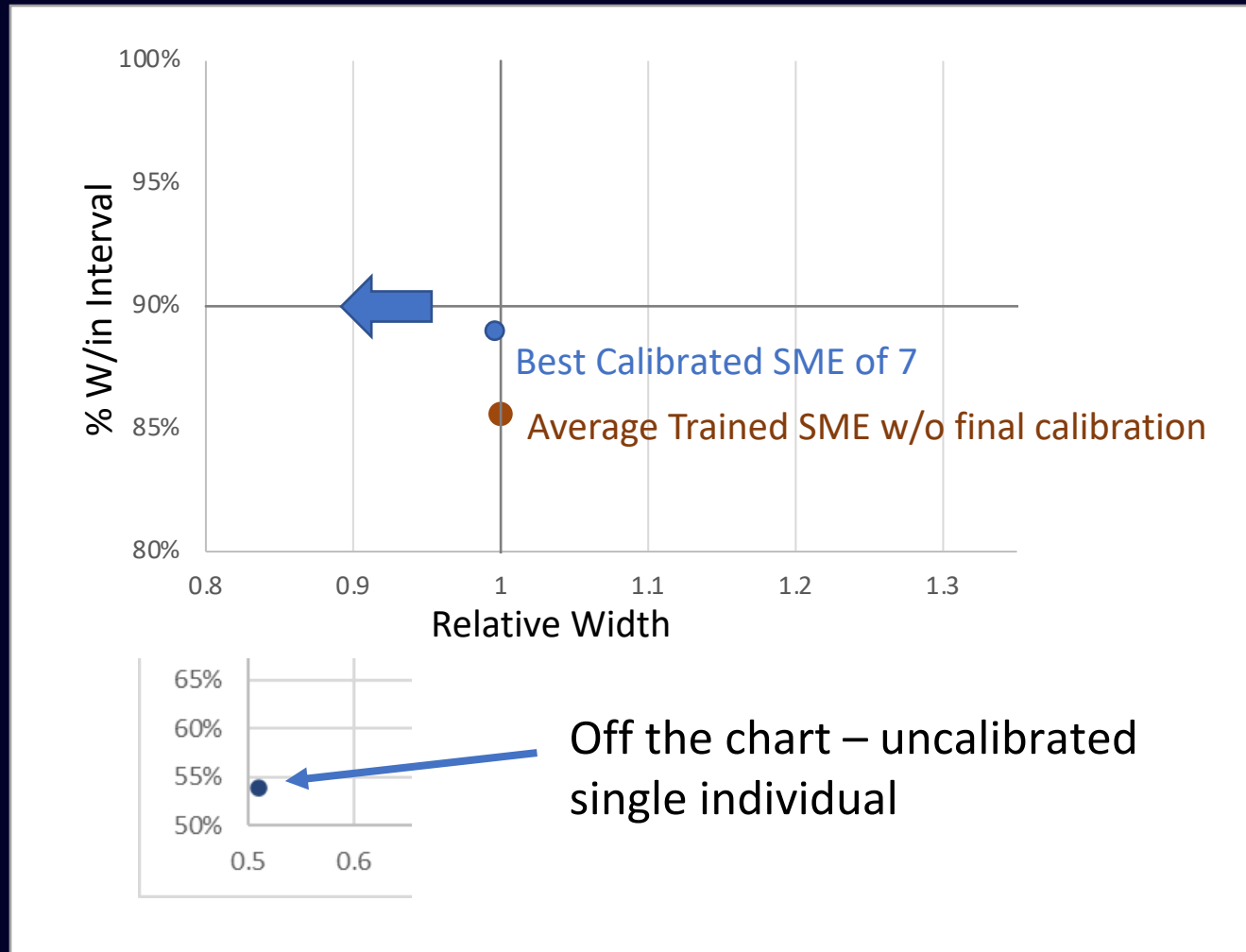


FrankenSME of 7 SMEs,  
Relative Width of P5 to  
P95

The FrankenSME is based on a hybrid Bayesian and machine learning method that looks at “patterns of agreement.”

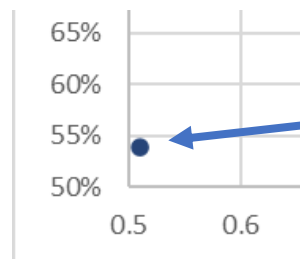
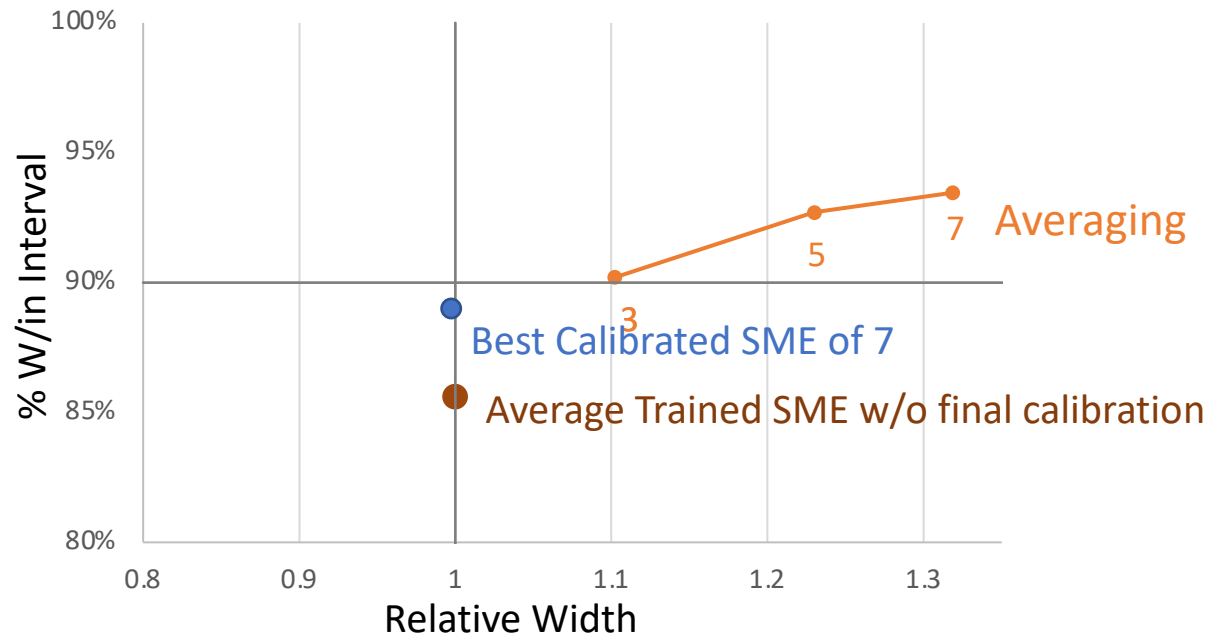
It will have an ideally calibrated 90% interval and will be narrower than the best SME on average.

# Comparing Aggregation Methods



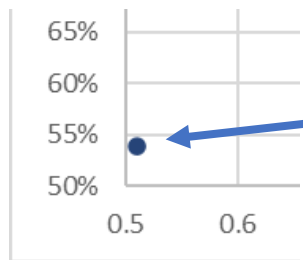
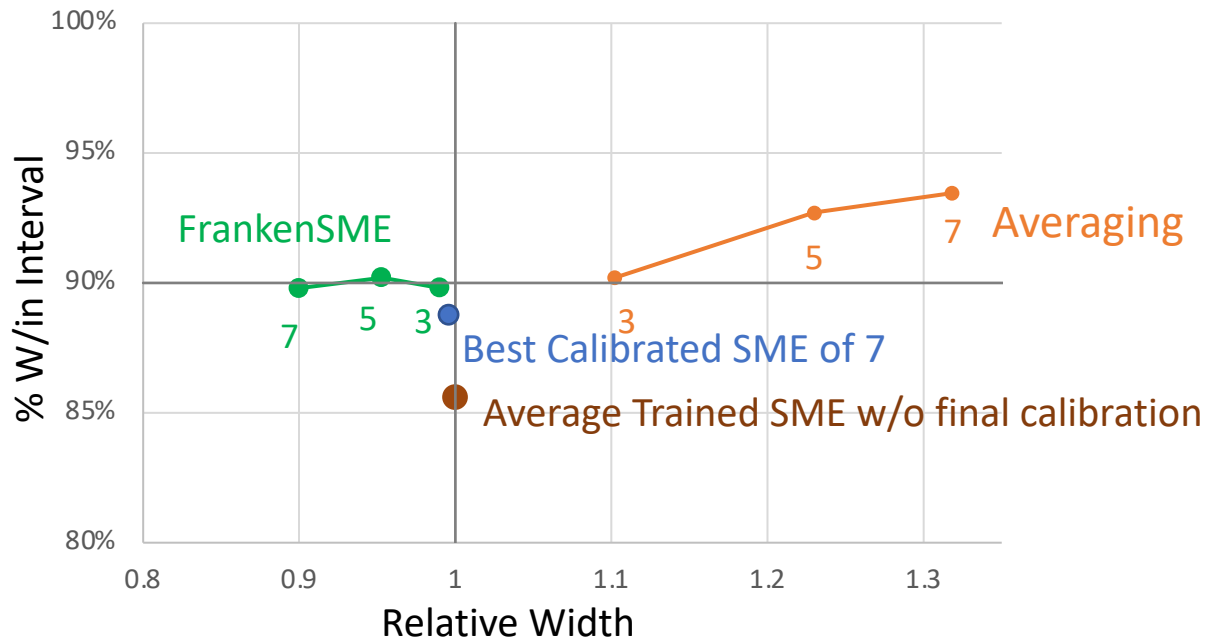
- Calibration and information (width of the interval) are measures of different aggregation methods.
- The width of confidence intervals is normalized to use the average trained SME as a width of “1.”
- The objective is to be narrow but calibrated.
- The best of 7 is about as wide as 1 on average but better calibrated (closer to 90% within the CI).
- The untrained individual has a narrow but extremely uncalibrated range.

# Comparing Aggregation Methods



Off the chart – uncalibrated single individual

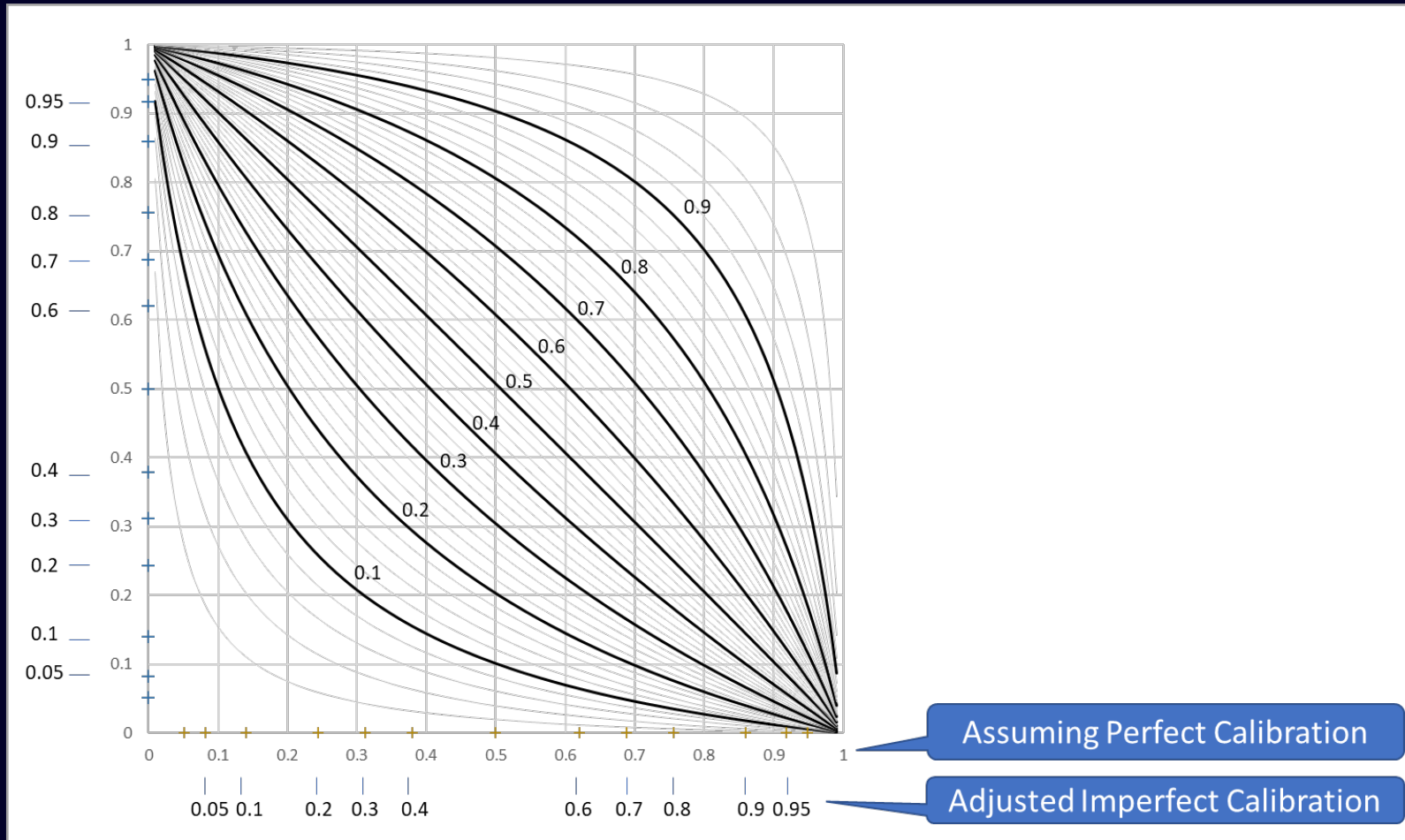
- Averaging improves calibration with 3 SMEs but makes the range wider on average.
- Averaging more than 3 tends to make the range too wide and underconfident.



Off the chart – uncalibrated single individual

- Only the FrankenSME gets narrower with more SMEs and stays calibrated.
- There are other aggregation algorithms but they also tend to create wider ranges as more SMEs are included.

# A Special Case: Aggregating Two Experts





# 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 in 2016 is 16%.”

What if we could prioritize security investments based on a “Return on Mitigation”?

	Expected Loss/Yr	Cost of Control	Control Effectiveness	Return on Control	Action
DB Access	\$24.7M	\$800K	95%	2,832%	Mitigate
Physical Access	\$2.5M	\$300K	99%	727%	Mitigate
Data in Transit	\$2.3M	\$600K	95%	267%	Mitigate
Network Access Control	\$2.3M	\$400K	30%	74%	Mitigate
File Access	\$969K	\$600K	90%	45%	Monitor
Web Vulnerabilities	\$409K	\$800K	95%	-51%	Track
System Configuration	\$113K	\$500K	100%	-77%	Track

