Advanced Calibrated Probability Assessments

Module 1: Basic Calibration Tools and Procedures for Trainers

Hubbard Decision Research
2 South 410 Canterbury Ct
Glen Ellyn, Illinois 60137
www.hubbardresearch.com

Advanced Calibration Outline

Modules of The Computer Based Training Course

- Basic Calibration Tool Use and Procedures
- Dealing with Challenges
- Optional Next Steps

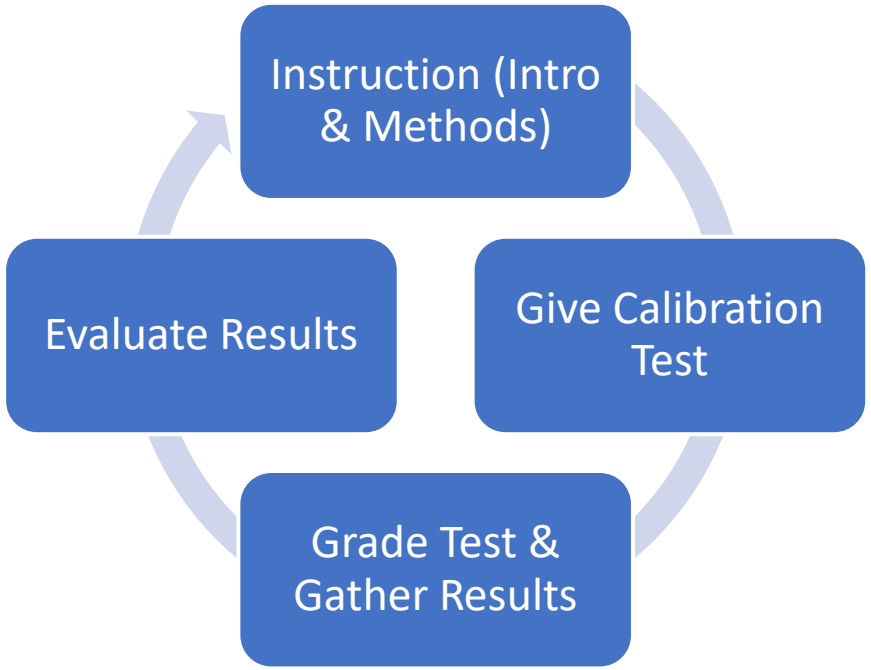
Advanced Calibration Outline

Modules of The Computer Based Training Course

- Basic Calibration Tool Use and Procedures
- Dealing with Challenges
- Optional Next Steps

The Training Cycle

What a Calibrated Group Should Look Like



- Calibration training uses 4 to 6 exercises.
- Each exercise starts with a presentation component (the first one is the introduction and each one thereafter introduces a new technique).
- When everyone is done with each test, give out the “scoring code” provided for each test.
- After the scoring code is entered by participants on their test sheets, collect these three numbers from each participant:
  - On range questions get “answers within ranges” (e.g. “7 of 10”)
  - On the True/False (Binary) tests get “predicted correct” and “actual correct”
- Enter the results into the summary sheet and review them Discuss group and individual progress.

Working the Spreadsheet – Basic Features

What a Calibrated Group Should Look Like

- Two sections: range and true/false
- Sections for participant input are colored yellow – all other cells are protected
- Answers appear after participants input the grading code

Short B

	Range Questions	90% Confidence Interval Lower Bound	90% Confidence Interval Upper Bound	Correct Answer
1	How many countries are in NATO?	7	40	26
2	What is the average Fahrenheit temperature in Boston, MA in April?	40	70	48
3	On behalf of the US, President Grover Cleveland accepted the Statue of Liberty as a gift from France in what year?	1800	1940	1886
4	What were the total number of gold medals won by the USA in the 2008 Bejing Summer Olympics?	32	38	36
5	What is the percentage of Americans without healthcare insurance?	12%	32%	15.3%
6	The Earth is 93 million miles from the sun. How far is Venus from the sun (in millions of miles)?	20	70	67.2
7	What year did *Gunsmoke* premier on television?	1940	1980	1955
8	How many billions of dollars did Microsoft earn in revenue in 2008?	2	100	51.1
9	How many stories tall is the Empire State Building?	60	120	102
10	What is the weekly food expense (in dollars) for the average US household with children under 18 (per person)?	$30.00	$32.00	$31.25

	True/False Questions	T/F	% Confidence	Answer	
1	The percentage of households which own their homes is higher in North Carolina than New York.	T	50%	T	
2	In high humidity, baseballs tend to be hit further than in low humidity.	F	80%	F	
3	Alpha Centauri is closer than Andromeda.	T	60%	T	
4	When Churchill said "Never in the field of human conflict was so much owed by so many to so few," he was referring to the soldiers of D	T	90%	F	X
5	The US has competed in the soccer World Cup	T	60%	T	
6	Adjusted for inflation, Hurricane Andrew was more costly than Hurricane Katrina	F	90%	F	
7	Nuclear fusion involves splitting helium into hydrogen.	F	70%	F	
8	In 2012 Sam's club (Walmart subsidiary) sales were greater than Amazon.com sales	F	60%	F	
9	President John Adams was a lawyer.	T	95%	T	
10	The Yangtze River is the longest river in Asia.	T	80%	T	

Grading Code:	4321
Answers in Given Ranges	10 of 10
Predicted Correct:	7.4
Actual Correct:	9

Range Adjustment	0.74
Binary Adjustment:	+17%

Spreadsheet Grading and Scoring

What a Calibrated Group Should Look Like

Short B

	90% Confidence Interval		Correct
	Lower Bound	Upper Bound	Answer
1	7	40	**26**
2	40	70	**48**
3	1800	1940	**1886**
4	32	38	**36**
5	12%	32%	**15.3%**
6	20	70	**67.2**
7	1940	1980	**1955**
8	2	100	**51.1**
9	60	120	**102**
10	$30.00	$32.00	**$31.25**

	T/F	% Confidence	Answer	
1	T	50%	**T**	
2	F	80%	**F**	
3	T	60%	**T**	
4	T	90%	**F**	x
5	T	60%	**T**	
6	F	90%	**F**	
7	F	70%	**F**	
8	F	60%	**F**	
9	T	95%	**T**	
10	T	80%	**T**	

Grading Code:	4321
Answers in Given Ranges	10 of 10
Predicted Correct:	7.4
Actual Correct:	9

Range Adjustment	0.74
Binary Adjustment:	+17%

Give participants the grading codes for the corresponding quizzes *only* after presenting that section's material or reviewing techniques.

Scores appear at the bottom of each test after the grading code is entered.

Grading Code:	4321
Answers in Given Ranges	10 of 10
Predicted Correct:	7.4
Actual Correct:	9

The Other Scores…

What a Calibrated Group Should Look Like

Short B

	90% Confidence Interval		Correct
	Lower Bound	Upper Bound	Answer
1	7	40	26
2	40	70	48
3	1800	1940	1886
4	32	38	36
5	12%	32%	15.3%
6	20	70	67.2
7	1940	1980	1955
8	2	100	51.1
9	60	120	102
10	$30.00	$32.00	$31.25

	T/F	% Confidence	Answer	
1	T	50%	T	
2	F	80%	F	
3	T	60%	T	
4	T	90%	F	x
5	T	60%	T	
6	F	90%	F	
7	F	70%	F	
8	F	60%	F	
9	T	95%	T	
10	T	80%	T	

Grading Code:	4321
Answers in Given Ranges	10 of 10
Predicted Correct:	7.4
Actual Correct:	9

Range Adjustment	0.74
Binary Adjustment:	+17%

The Range Adjustment is the factor by which a participant needs to expand their range to be calibrated.

Range Adjustment	0.74
Binary Adjustment:	+17%

The Binary Adjustment is the average increase or decrease in stated confidence required to achieve calibration.

The Odds: The Benchmark Test

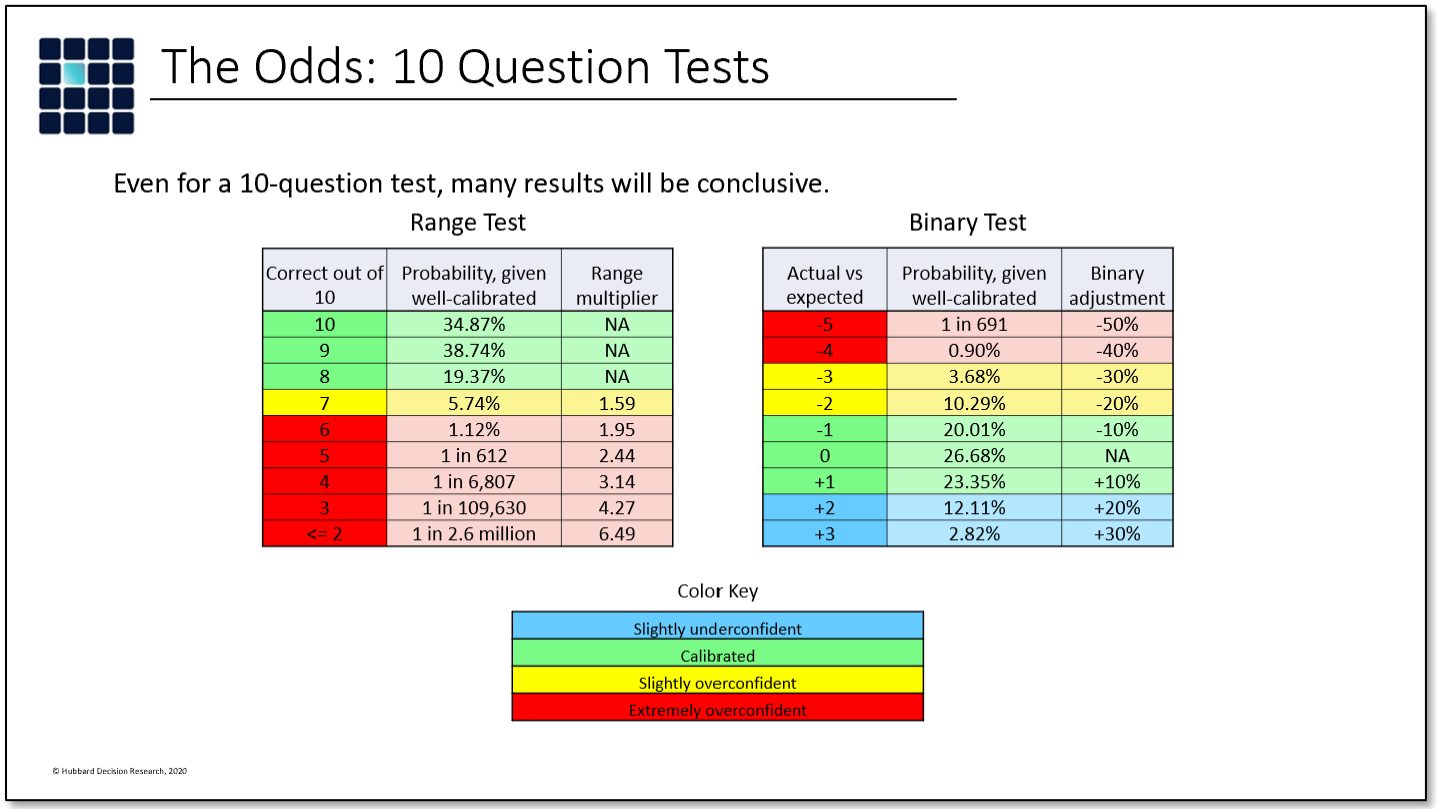
What a Calibrated Group Should Look Like

- You must help participants to connect the dots – that they are not getting less range or binary questions than expected *just by chance* (some will either consciously or subconsciously believe it is bad luck or bad questions) *but rather because they aren't calibrated*.
- You need to be blunt – usually most participants will need to be shocked into understanding that they are fundamentally overconfident.

- Point out that in a group of ten calibrated people, nine plus would get 8-10 out of 10 range questions on a range test. Similarly for the binary tests nine or more people would be within 2 of their expected.

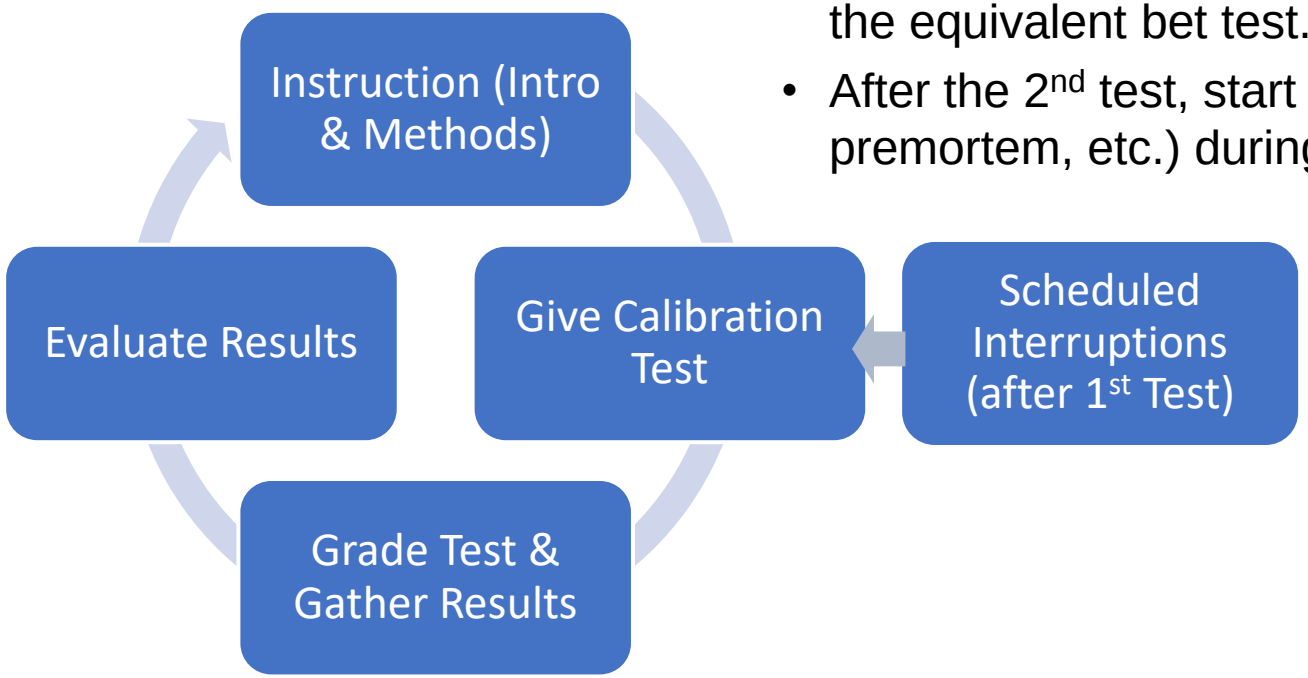
- Point out that in a group of 6,800 calibrated people, **only one person** would get 4 or less on a ten question range test just "by chance." For the binary test we would only expect one out of a group of 691 to get more than a difference of 5 from their expected.

- Similar odds exist for the 20-question test.



The Scheduled Interruption

A Necessary Reminder



- Just before the 2nd test, announce that you will be using a “scheduled interruption”. Wait about 2 minutes and ask who is using the equivalent bet test. Remind everyone that it is a key method.
- After the 2nd test, start asking about additional tactics (Klein's premortem, etc.) during the scheduled interruption.

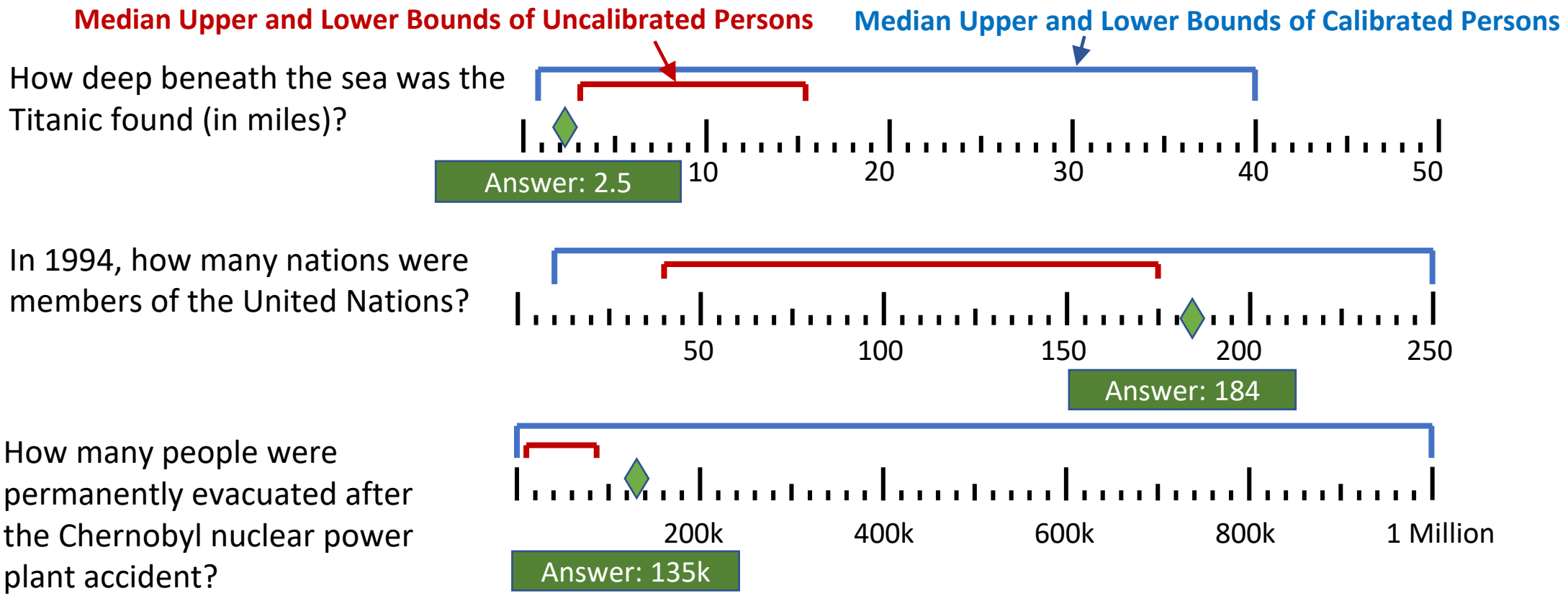
What to Cover in Each Iteration

A Reference Table

Lecture	Test (After Lecture)	Scheduled Interruption	# of test items	Other Notes
Introduction/Objective	Test A		10 Range, 10 T/F	The first small benchmark test is sufficient to see overconfidence, especially in ranges.
Introduction to grading, past research, overconfidence, and Equivalent Bet	Test B	✓	10 Range, 10 T/F	Don't read too much into improvements at this point – especially for binary.
Avoiding to anchoring	Test C	✓	20 Range, 20 T/F	Intervene early if performance is bad (range score< 13 correct).
Klein's premortem	Test D	✓	20 Range, 20 T/F	Absolute minimum number of tests even with nearly perfect scores up to this point
Applying calibration adjustments	Test E	✓	20 Range, 20 T/F	Recommended number of tests
Optional: Do's and don'ts, review of effects of calibration	Test F	✓	20 Range, 20 T/F	This many tests are needed if performance is not good

Three Relative Calibration Ranges

Individuals scoring between 17 and 19 on range tests are not just better at trivia. They are simply willing to use wider ranges. Generally about 2 to 10 times wider than the people scoring far below average.



The Odds: 10 Question Tests

Even for a 10-question test, many results will be conclusive.

Range Test

Correct out of 10	Probability, given well-calibrated	Range multiplier
10	34.87%	NA
9	38.74%	NA
8	19.37%	NA
7	5.74%	1.59
6	1.12%	1.95
5	1 in 612	2.44
4	1 in 6,807	3.14
3	1 in 109,630	4.27
<= 2	1 in 2.6 million	6.49

Binary Test

Actual vs expected	Probability, given well-calibrated	Binary adjustment
-5	1 in 691	-50%
-4	0.90%	-40%
-3	3.68%	-30%
-2	10.29%	-20%
-1	20.01%	-10%
0	26.68%	NA
+1	23.35%	+10%
+2	12.11%	+20%
+3	2.82%	+30%

Color Key

Slightly underconfident
Calibrated
Slightly overconfident
Extremely overconfident

The Odds: 20 Question Test

A 20-question test will have slightly better resolution – but still better at detecting overconfidence than under-confidence

Range Test

Correct out of 20	Probability, given well-calibrated	Range multiplier
20	12.2%	NA
19	27.0%	NA
18	28.5%	NA
17	19.0%	NA
16	9%	1.28
15	3.2%	1.43
14	0.89%	1.59
13	0.20%	1.76
12	0.036%	1.95
11	0.005%	2.18
<= 10	1 in 126,135	2.4 – 3.6

Binary Test

Actual vs expected	Probability, given well-calibrated	Binary adjustment
<= -5	0.68%	-25%
-4	2.78%	-20%
-3	7.16%	-15%
-2	13.04%	-10%
-1	17.89%	-5%
0	19.16%	0%
1	16.43%	+5%
2	11.44%	+10%
3	6.54%	+15%
4	3.08%	+20%
>= 5	1.20%	+25%

Color Key

Slightly underconfident
Calibrated
Slightly overconfident
Extremely overconfident

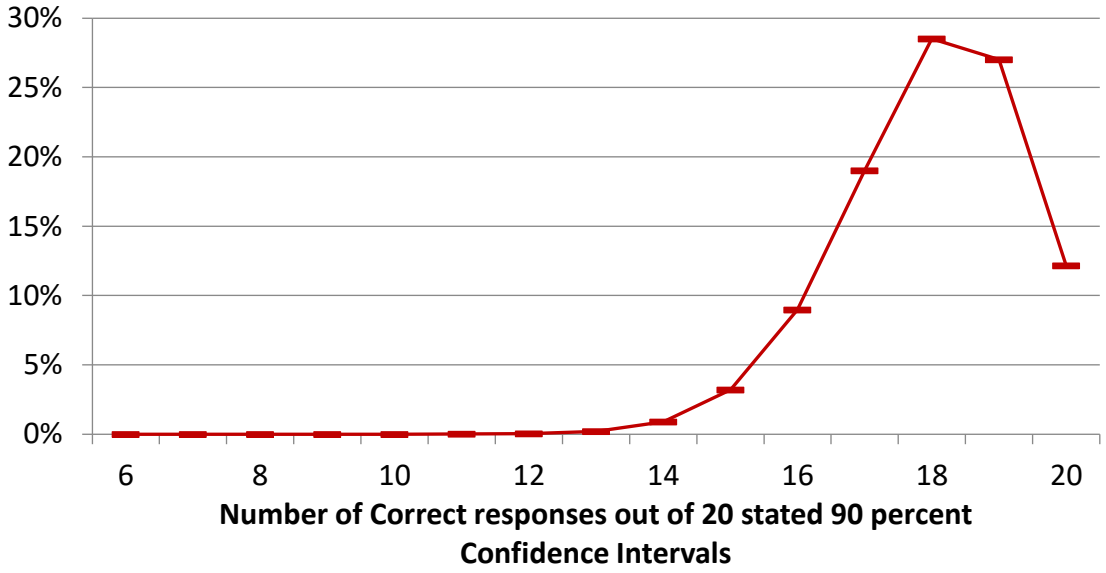
Expected Results

What a Calibrated Group Should Look Like

About 15% may fail to be calibrated. Out of a large group, some will *appear* not to be calibrated but some of that will just be expected random variation.

Final Range Test Expected Distribution

About 3/4 of the students should get a 17,18 or 19 out of 20



Final Binary Test Expected Distribution

About 9/10 of the students should be within +/-3 between expected and actual

Actual vs expected	Probability, given well-calibrated
-5	1 in 691
-4	0.90%
-3	3.68%
-2	10.29%
-1	20.01%
0	26.68%
+1	23.35%
+2	12.11%
+3	2.82%

Using Calibration in Estimation Workshops

What a Calibrated Group Should Look Like

The point of calibration is to better estimate uncertainties in real decisions. You may be called on to facilitate workshops where the goal is to estimate various quantities. Here are some guidelines for those meetings:

- Redirect the “Storyteller”: There is often a strong temptation for people to explain in detail complicating factors, exceptions, historical background, etc. It’s a given that participants have uncertainty. Push them to provide a range.
- Remind them to not assume wide ranges are useless: If it represents uncertainty fairly, that’s the range we want. Whether that range needs to be narrowed is another step in the Applied Information Economics process.
- Resist “Infinite Decomposition”: You can always compute a value based on other more detailed values but at some point, you have to just provide a range.
- Remind them that they are calibrated: Their performance skill at assessing odds has be proven quantitatively.

Module 1 Summary

What a Calibrated Group Should Look Like

- You have just reviewed the basics of using the calibration procedure and the two spreadsheets – the calibration exercises sheets and the calibration results summary sheet.
- Now you can take the quiz for the first module.
- When you are done you can begin Module 2, “Dealing With Challenges.”

Supplementary Material

Slides from Calibration Below

Course Objective

Learn how to assess odds like a bookie

This skill is called “Calibrated Probability Assessment”.

When you say you are 90% confident, you will have a 90% chance of being right!

Calibrated Experts

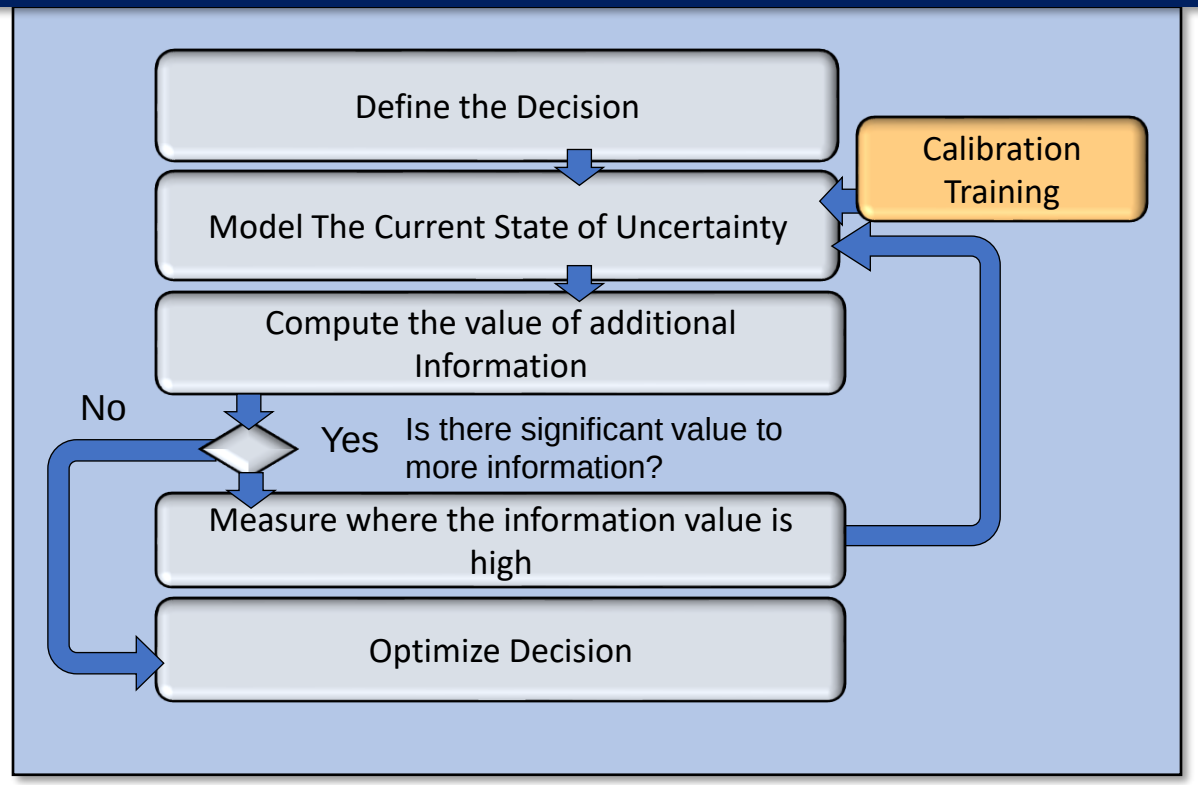
"Overconfident professionals sincerely believe they have expertise, act as experts and look like experts. You will have to struggle to remind yourself that they may be in the grip of an illusion."

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 <u>can be taught</u> with a ***measurable*** improvement.

A Process that Utilizes Experts

Applied Information Economics treats subject matter experts as key measurement instruments that must be calibrated before use.



Expected vs. Actual

To determine your level of calibration, we need to compare *actual* outcomes to *"expected"* outcomes.

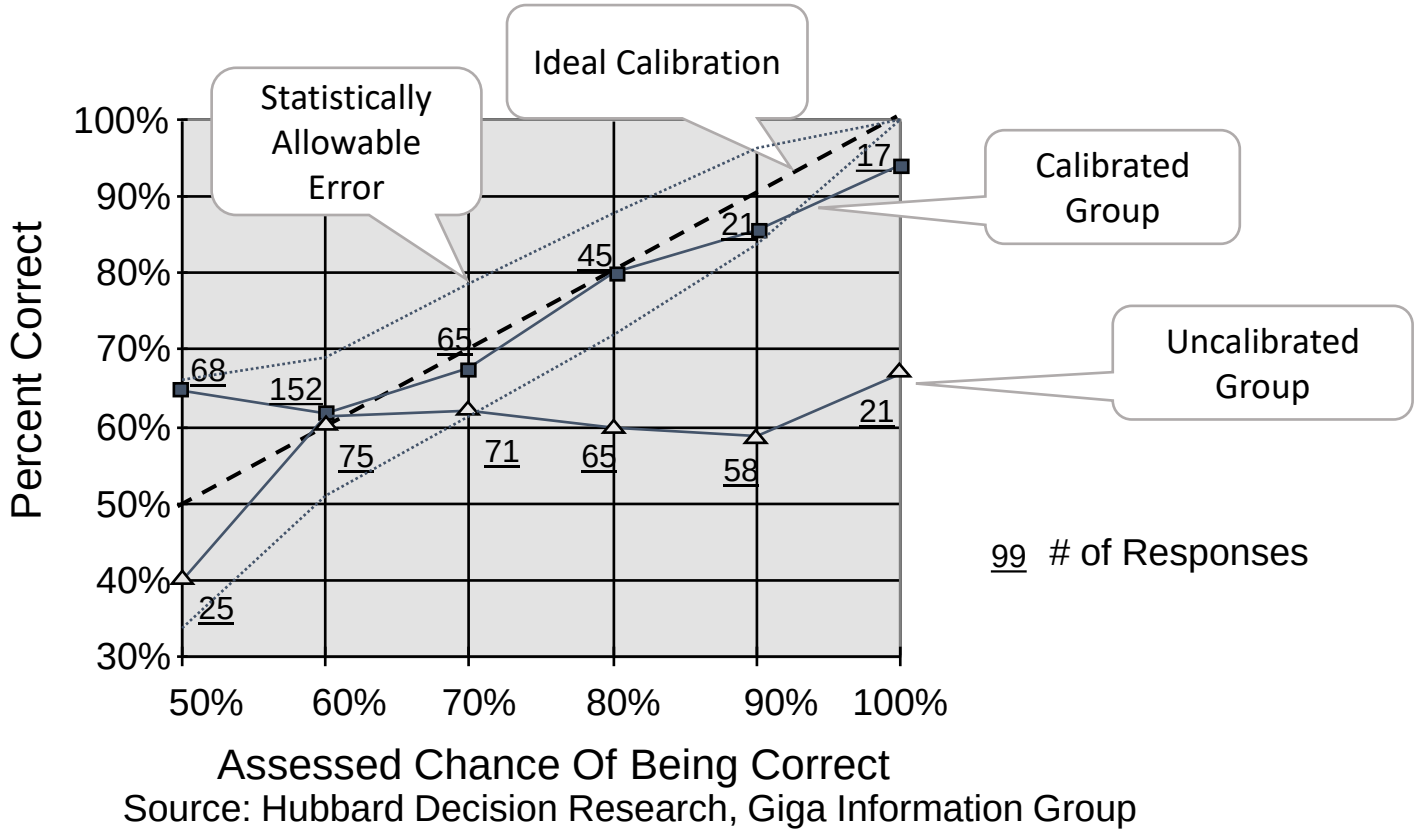
In decision analysis, the word **"expected"** literally means "probability weighted average".

- For the questions that ask for a 90% confidence interval, you expect to get 90% between your upper and lower bounds, by definition.

- For the true/false questions, your expected number correct is equal to the total confidence on your answers. That is, if you were 50% confident on each, you expected to get half right; if you were 100% confident on each, you expect to get them all right, and so on.

Training Experts to Give Calibrated Probabilities

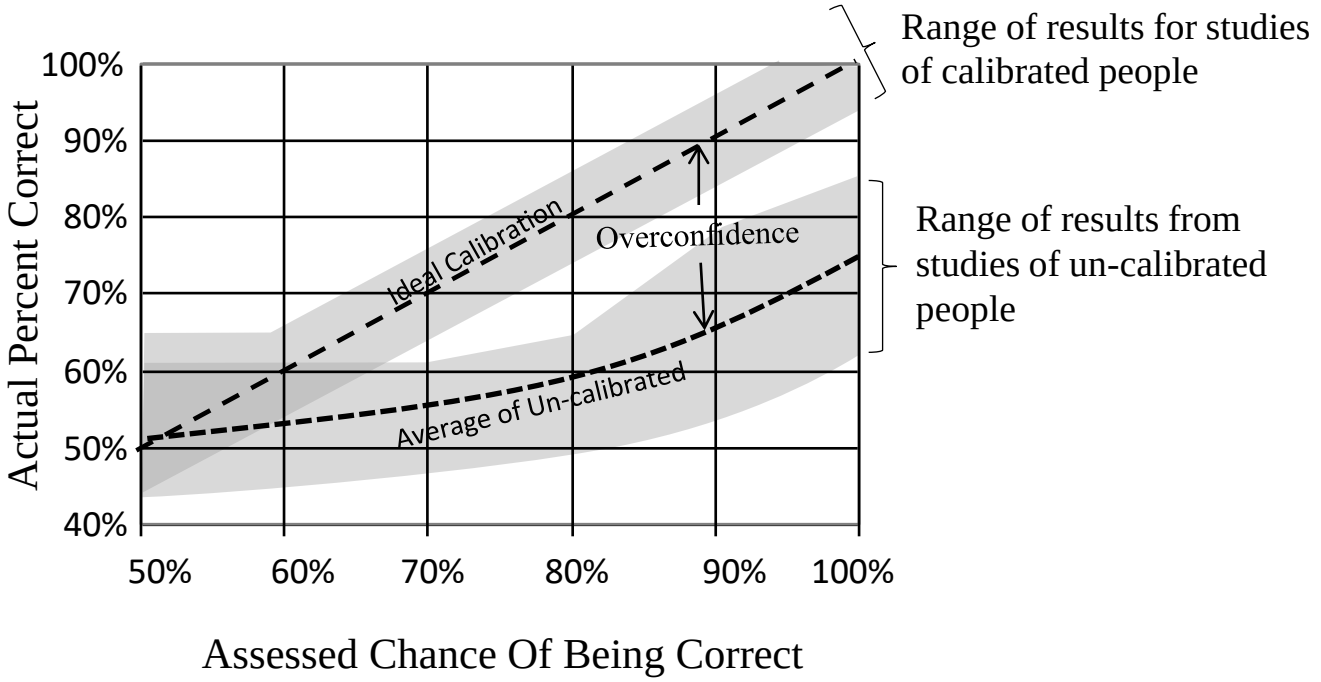
Training can “calibrate” people so that of all the times they say they are 90% confident, they will be right 90% of the time.



Source: Hubbard Decision Research, Giga Information Group

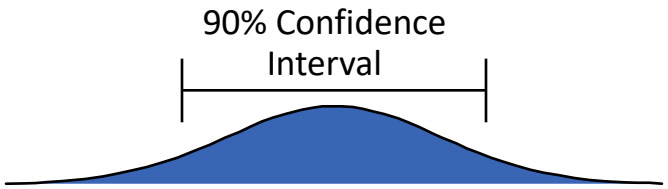
Overconfidence

- This is the aggregate of 11 studies in how well people subjectively assess odds
- The overwhelming evidence shows that everyone is systematically “overconfident”
- Fortunately, training and other techniques exist that adjust for this error
- Unfortunately, almost nobody uses those methods



Overconfidence in Ranges

The same training methods apply to the assessment of uncertain ranges for quantities like the duration of project, the impact of a major data breach, etc.



Group	Subject	% Correct (target 90%)
Harvard MBAs	General Trivia	40%
Chemical Co. Employees	General Industry	50%
Chemical Co. Employees	Company-Specific	48%
Computer Co. Managers	General Business	17%
Computer Co. Managers	Company-Specific	36%
AIE Seminar (before training)	General Trivia & IT	35%-50%
AIE Seminar (after training)	General Trivia & IT	~90%

The Odds: 10 Question Tests

Even for a 10-question test, many results will be conclusive.

Range Test

Correct out of 10	Probability, given well-calibrated	Range multiplier
10	34.87%	NA
9	38.74%	NA
8	19.37%	NA
7	5.74%	1.59
6	1.12%	1.95
5	1 in 612	2.44
4	1 in 6,807	3.14
3	1 in 109,630	4.27
<= 2	1 in 2.6 million	6.49

Binary Test

Actual vs expected	Probability, given well-calibrated	Binary adjustment
-5	1 in 691	-50%
-4	0.90%	-40%
-3	3.68%	-30%
-2	10.29%	-20%
-1	20.01%	-10%
0	26.68%	NA
+1	23.35%	+10%
+2	12.11%	+20%
+3	2.82%	+30%

Color Key

Slightly underconfident
Calibrated
Slightly overconfident
Extremely overconfident

Calibration Aid: “The Equivalent Bet”

For 90% Confidence Interval questions, which game would you rather play?

- **Game A**: Win $1,000 if your interval contains the correct answer
- **Game B**: Spin a dial with a 90% chance to win $1,000

For the Binary Confidence questions, which game would you rather play?

- **Game A**: Win $1,000 if your answer is correct
- **Game B**: Spin a dial with a chance to win $1,000 equal to your stated confidence

The Equivalent Bet Cheat Sheet

For 90% Confidence Interval questions, which game would you rather play?

- **Game A**: Win $1,000 if your interval contains the correct answer → Narrow your range!
- **Game B**: Spin a dial with a 90% chance to win $1,000 → **Widen** your range!

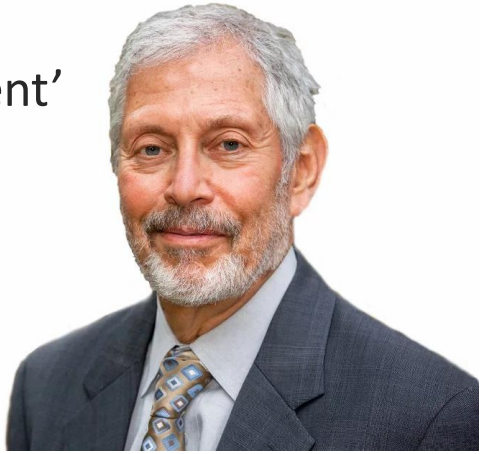
For the Binary Confidence questions, which game would you rather play?

- **Game A**: Win $1,000 if your answer is correct → **Increase** your confidence!
- **Game B**: Spin a dial with a chance to win $1,000 equal to your stated confidence → Decrease your confidence!

Klein's Premortem: a Prospective Hindsight Approach

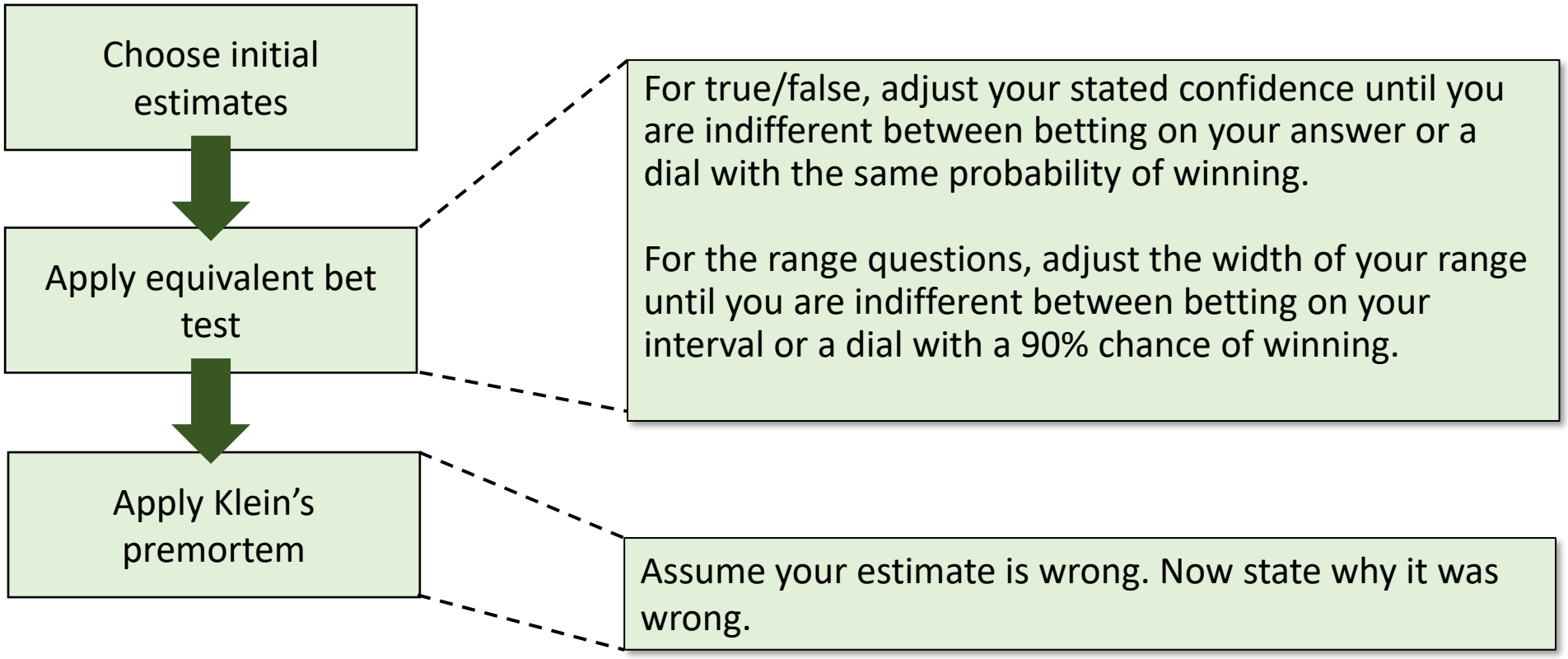
"Unlike a typical critiquing session, in which project team members are asked what *might* go wrong, the premortem operates on the assumption that the 'patient' has died, and so asks what *did* go wrong."

Gary Klein, Psychologist, in an article for Harvard Business Review



Initial estimates → **Assume your answer is wrong** → Explain why you were wrong → Update your estimates

Calibration Process #1



Anchoring

Our initial thought tends to be sticky and our subsequent range centers around it – even if that initial thought came from an unrelated piece of information that we think we remember.

Apply “The Equivalent Bet” to Each Bound

For lower bound estimates, which game would you rather play?

- **Game A**: Win $1,000 if the correct answer is *above your lower bound*
- **Game B**: Spin a dial with a 95% chance to win $1,000

For upper bound estimates, which game would you rather play?

- **Game A**: Win $1,000 if the correct answer is *below your upper bound*
- **Game B**: Spin a dial with a 95% chance to win $1,000

The Equivalent Bet for Each Bound Cheat Sheet

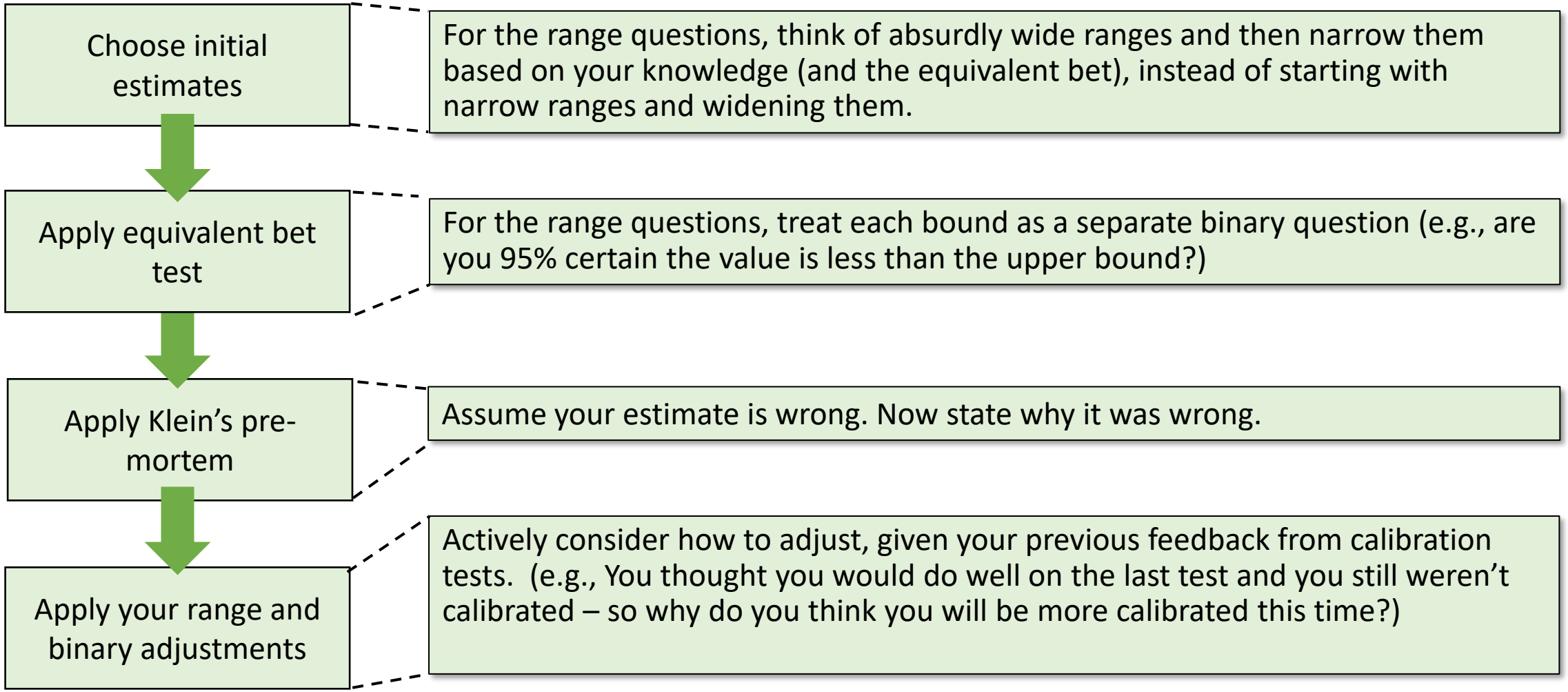
For lower bound estimates, which game would you rather play?

- **Game A**: Win $1,000 if the correct answer is *above your lower bound* → **Increase** your lower bound!
- **Game B**: Spin a dial with a 95% chance to win $1,000 → Decrease your lower bound!

For upper bound estimates, which game would you rather play?

- **Game A**: Win $1,000 if the correct answer is *below your upper bound* → Decrease your upper bound!
- **Game B**: Spin a dial with a 95% chance to win $1,000 → **Increase** your upper bound!

Calibration Process #2



Calibration Do’s and Dont’s

Don’ts

- Don’t think of this as a test of trivia knowledge.
- Don’t presume that wide ranges are useless.
- Don’t hang on to traditional expectations of “+/- 10%” ranges.
- Don’t think of your answers as “guesses.”

Do’s

- Do think of this as a test of assessing your uncertainty (whatever your level of knowledge).
- Do give a wide range if it realistically represents your uncertainty.
- Do remember to use the calibration process.
- Do remember that your estimates have measurably improved and will be aggregated with other SMEs.

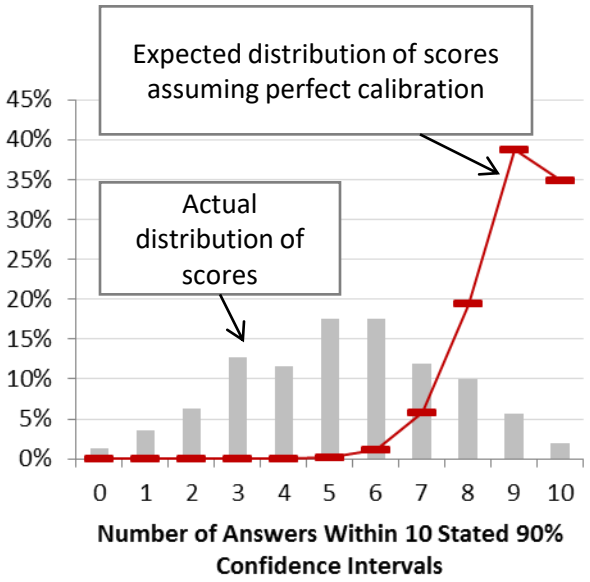
Calibration Improvement

With over 880* subjects who have taken the same calibration tests, and over 100,000 individual responses, a clear pattern emerges:

Training has a major impact on 90% CI tests.

***Now over 1400 subjects, as of April 2019**

Initial 10-Question 90% CI Test



Final 20-Question 90% CI Test