OBESITY EXPERTISE

A first-principles curriculum in 196 books

BOOK 10 11th of 196 in the series

Epistemics: philosophy of science and calibration · Rung 1

How strong is the evidence, how sure should you be, and how to say you were wrong

CONFIDENCE THAT FOLLOWS THE EVIDENCE, WRITTEN DOWN, DATED AND CHANGED OUT LOUD

Dr. Harsh Maheshwari

MD Community Medicine Resident, AIIMS Raipur

drharshmaheshwari.com · version 1.0

OBESITY EXPERTISE · BOOK 10

Epistemics: philosophy of science and calibration · Rung 1

How strong is the evidence, how sure should you be, and how to say you were wrong

Dr. Harsh Maheshwari

MD Community Medicine Resident, AIIMS Raipur

contact@drharshmaheshwari.com · drharshmaheshwari.com · ORCID 0000-0001-8848-1918

SERIES	Obesity Expertise: A first-principles curriculum in 196 books
BOOK	Book 10, Epistemics: philosophy of science and calibration · Rung 1 (11th of 196 in the series)
VERSION	1.0
LAST UPDATED	14 October 2026
AUTHOR	Dr. Harsh Maheshwari, MD Community Medicine Resident, AIIMS Raipur
CONTACT	contact@drharshmaheshwari.com
ORCID	https://orcid.org/0000-0001-8848-1918
NEWEST VERSION	drharshmaheshwari.com

© 2026 Harsh Maheshwari.

This book is shared free, for anyone to read, under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International licence (CC BY-NC-SA 4.0). You may copy, share and adapt it for non-commercial purposes, if you credit the author and share what you make under the same licence.

It quotes and adapts material from OpenStax textbooks (Rice University), which are licensed CC BY-NC-SA 4.0; that is why this book carries the same licence. It also quotes short passages from laws, reports and other works, each credited where it is used; those passages remain under their owners' terms. Fonts: Source Serif 4, Inter and JetBrains Mono, SIL Open Font License 1.1.

HOW THIS BOOK WAS MADE

Written for the author's own study, then shared. Drafted and checked with AI assistance (Claude, by Anthropic) under the author's direction. Every factual claim was checked against a source the checker had open, every quotation against its source text, and every worked answer recomputed. Mistakes can remain; the feedback page says how to report one.

NOT ADVICE

This book is for learning. It is not medical, legal or financial advice, and it does not replace a qualified professional who knows your situation.

About the author



Dr. Harsh Maheshwari

MD Community Medicine Resident, AIIMS Raipur

Harsh Maheshwari completed his MBBS at King George's Medical University (KGMU), Lucknow, and is a resident in MD Community Medicine at the All India Institute of Medical Sciences (AIIMS), Raipur. His MD thesis, on chrononutrition and obesity, was selected by the Indian Council of Medical Research (ICMR) for funding in 2026.

He learns best from first principles: rebuilding an idea from what came before it until the formula feels like the only reasonable answer. This series grew out of that habit.

His long-term aims are to reduce obesity in India and to help make India healthy.

Email	contact@drharshmaheshwari.com
Website	drharshmaheshwari.com
ORCID	0000-0001-8848-1918

Introduction

Why this book exists

A clinician or researcher is asked every day how sure they are, and usually answers with a tone of voice. This book treats confidence as something that should follow the strength of the evidence, and teaches the difference between a hypothesis, a finding and a belief. It shows how to put a confidence on a claim in plain words and say why, and why being wrong in public is a normal event in science rather than a disgrace.

How to read it

Read Book 0 first, or at least the sections this book names as it goes. Then read this one in order. The book asks you to build one thing: a dated written record of five beliefs you hold about your own field, each with a confidence attached. Start it when the book tells you to and keep it; the next book in this subject scores it.

Every section has the same shape. **Definition** states the idea exactly. **In plain terms** says it again in everyday words. **Illustration** works through one real case, usually with a source you can open yourself. **Where this picture breaks** says where the illustration stops being true. **Must-know points** are what should change what you do. Then come **exercises** and, in most sections, a graded **practice set**.

Do the exercises before you look anything up. The worked answers are in the appendix at the back, under the same numbers. The references for each Part follow that Part. The glossary lists every term of art in the plain words it was first given. The last pages list all 196 books in the series and show where this one sits.

If you are new to this

Book 0, in particular percentages (A4), ratios (A5), what a probability is (D1), the two-way table (D3), average and spread (D5), sampling (D6), random and systematic error (D7), reading a table (F1), checking a reference (F3) and the shape of an argument (F4). The mathematics is arithmetic with a calculator: turning a percentage into a fraction or odds and back, counting proportions in a small table, and filling a two-way table in natural frequencies. No algebra, no formula for Bayes' theorem and no statistics beyond a mean, a range and, in the section on calibration, Book 0's square-root law from sampling (D6), used to see how far chance moves a small tally of claims. You need something you have taught or written, and a notebook or file you will keep for the five dated beliefs.

On a first reading, read each section's definition, plain-terms explanation, illustration and must-know points, and do the exercises. Worked derivations can be skimmed and revisited once the idea feels familiar. Keep a calculator beside you: every worked answer can be followed with pen, paper and a calculator. Anything beyond the stated prerequisites is explained where it first appears, or marked as beyond this book. The **Symbols used in this book** page lists every sign and Greek letter with how to read it and where it is first used.

How to send feedback

If something is wrong, unclear, or could be better, please write to **contact@drharshmaheshwari.com**. Put the book number, the version and the section in the subject line (for example: *Book 10, v1.0, section C5*). Say what the text says, what you think it should say, and, if you can, the source.

Corrections are made in new versions. The version number and the date of the last update are on the copyright page of every book, and the newest version of every book is at **drharshmaheshwari.com**.

> **WHERE THIS BOOK SITS**
> Book 10 of the series' 196 books (Book 0 and 195 rung books). Before it: Book 9, Scientific writing, visualisation and public communication · Rung 1. After it: Book 11, Adjacent specialist domains: recognise and route · Rung 1. The full list is at the back.

Symbols used in this book

Every sign and Greek letter this book prints, how to read it aloud, what it means in this series, and the section where it first appears. Each is also explained where it is first used.

Symbol	Read as	Meaning	First in
SIGNS			
⊕	a circled plus, GRADE's 'filled circle'	One level of certainty of evidence reached in a GRADE rating: ⊕⊕⊕⊕ is high, ⊕◯◯◯ very low	7 · GRADE certainty: four levels, where you start, and moving down or up
◯	an empty circle (GRADE)	A level of certainty of evidence not reached in a GRADE rating	7 · GRADE certainty: four levels, where you start, and moving down or up
–	minus	Subtraction, or a negative number	9 · No evidence of an effect is not evidence of no effect
MARKS ON LETTERS			
x^2, 10^{-3}	superscript	Small raised characters are powers: $x^2 = x \times x$, and $10^{-3} = 0.001$	4 · Belief beyond the evidence

Contents

Before you start 9

Rung 1 — Introductory 10

1 · Three kinds of statement: a hypothesis, a finding and a belief 10
2 · A hypothesis is testable: the prediction, and the result that would count against it 15
3 · One finding, several explanations 20
4 · Belief beyond the evidence 24
5 · Two kinds of support for a causal claim: a difference between groups, and a mechanism 31
6 · Why evidence comes in strengths 38
7 · GRADE certainty: four levels, where you start, and moving down or up 43
8 · Certainty of the evidence is not the strength of a recommendation, and certainty has words 50
9 · No evidence of an effect is not evidence of no effect 55
10 · Confidence as a probability: putting a number on how sure you are of one claim 60
11 · The same word read as different numbers 67
12 · Calibrated scales: fixing what the words mean, and putting the number beside the word 74
13 · Stating how confident you are, and saying why 82
14 · The base rate: how often claims like this one hold up 88
15 · Updating: changing your confidence by as much as the evidence warrants 95
16 · Confidence is not accuracy: what being calibrated means 104
17 · Changing a stated position out loud 113
18 · How science corrects its record: correction, expression of concern, retraction, republication 119
19 · Being wrong in public is common, and owning it costs less than hiding it 124
20 · Saying “I was wrong about that” 131
21 · The dated belief register: what you believe, how sure you are, and when you said it 137
22 · Worked journey: One claim, from a belief to a dated line: “eating breakfast helps you lose weight”. 143
References *151*

Appendix · Worked answers 181

Glossary 217

The series: all 196 books 222

Before you start

This book uses the following Book 0 sections. Each is summarised where it is first needed; work through Book 0 itself if a summary is not enough.

- Book 0, A4 · Percentages, and the percentage of a percentage
- Book 0, A5 · Ratios, rates and proportions
- Book 0, D1 · What a probability is
- Book 0, D3 · Conditional probability through the two-way table
- Book 0, D5 · Average and spread, and why the average is not the person
- Book 0, D6 · Sampling: how a part can tell you about a whole
- Book 0, D7 · Random error and systematic error as different things
- Book 0, F1 · Reading a table
- Book 0, F3 · Checking a reference
- Book 0, F4 · Argument: premise, inference, conclusion

Rung 1 — Introductory

1 · Three kinds of statement: a hypothesis, a finding and a belief

DEFINITION. A hypothesis is a proposed explanation, stated so that a study can be designed to support or refute it.

A finding is what a particular study observed, stated with the people and setting it was observed in and with the uncertainty that comes with it. It reports; it does not on its own reach beyond the study.

A belief is a statement a person holds to be true, held with some strength, on whatever grounds that person has.

The three are uses of a statement, not kinds of sentence.

Sherman Kent, writing in 1964 for intelligence analysts, sorted statements on a second axis, by how close they come to what is known. The first kind states something known, as close to indisputable fact as one can come. The second is a judgement on evidence, which the evidence makes likely but cannot make certain. The third is a judgement made almost without evidence. The two sortings ask different questions and cross each other. [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]

IN PLAIN TERMS. Take one sentence: skipping breakfast makes people put on weight.

Offered to be tested, it is a hypothesis. Reporting what a study saw, in certain people, it is a finding. Simply held true, it is a belief.

Illustration 1. This book follows one claim from start to finish: eating breakfast helps you keep weight off, and skipping it puts weight on. Here it is three times. Sort each one before you read on.

The first is made up for this section. A resident writes in her thesis protocol:

> *We propose that students who skip breakfast on most days gain more weight over one year than students who eat it.*

The second is real. Brown and colleagues' 2013 paper "Belief beyond the evidence" is at https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/. They pooled the observational studies of breakfast and weight that four earlier reviews had gathered — studies that watch people as they are, without changing anything. They took 88 estimates from 58 studies, done in 30 countries. Pooling like this is a meta-analysis: a study that combines the results of earlier studies of the same question into one answer. In their Results they write:

> *There is a clear association between breakfast omission and excess weight as summarized in our cumulative meta-analysis (see RLPV below), and this association supports the PEBO but does not show causation.*

An association means the two things go together: excess weight is found more often among people who skip breakfast. PEBO is their short name for the claim that eating breakfast protects against obesity. A cumulative meta-analysis pools the estimates again each time a new year's studies are added, here year by year from 1994 to 2011. RLPV is their short name for "research lacking probative value": studies that cannot add to what is known, such as repeats of a question already answered.

The third is a line Brown and colleagues quote from the Mayo Clinic. It sits in their table of public statements, whose sources include health websites, a newspaper and the US Surgeon General:

> *In fact, skipping breakfast actually increases your risk of obesity.*

Now sort them with three questions.

	Hypothesis	Finding	Belief
What it does	offers an explanation to be tested	reports what one study saw	states what someone holds true
Its grounds	a reason to think it worth testing	the study's own data, in its own people	whatever persuaded the person
A result against it	is what the test is for	is another finding to set beside it	ought to change how strongly it is held

The protocol sentence is a hypothesis. Its job is to be tested, and it is useful even if it turns out wrong. Ask of it: what result would count against it?

Brown's sentence is a finding. It is about an association, seen in the people those studies measured. Notice that it says so itself: "does not show causation". Ask of it: who was studied, and what exactly was seen?

The Mayo Clinic line is a belief. It reaches further than the finding: it says skipping breakfast increases the risk, which is a claim about cause. As Brown and colleagues quote it, it names no study. Its words "in fact" and "actually" dress it as a finding. Ask of it: what is it resting on? Brown and colleagues' own answer, in their abstract, is that this claim "is only presumed true". (An abstract is the short summary printed before a paper.)

Figures quoted in this illustration are from [14].

WHERE THIS PICTURE BREAKS. The three kinds are uses, not fixed labels. One person can believe a claim and still test it as a hypothesis. And a hypothesis that passes many fair tests comes to be held as a belief, on good grounds. The sorting tells you what a statement is doing, not whether it is true.

Brown's finding covers the studies they pooled, published up to 2011. This book has not checked which countries those 30 were, so it says nothing here about Indian studies.

Illustration 2. Kent's essay "Words of estimative probability" is at https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf. It opens with an officer reporting a reconnaissance photograph. Read his three statements and sort them by how much the officer actually knows.

> *Its longest runway is 10,000 feet.*

> *It is almost certainly a military airfield.*

> *It is possible that they will.*

The last one answers whether the other side will lengthen the runways and make the field a base.

Here is Kent's own sorting. The first statement is, in his words, "as close as one can come to a statement of indisputable fact": the photograph shows it. The second is a judgement on evidence. The aircraft on the field are all military, and the support buildings look like other military bases. Kent says that evidence "makes the case, say, 90 percent of the way". It "cannot justify a flat assertion", so the officer adds "almost certainly". The third is a judgement made almost without evidence. Kent calls it "an estimate of something that no man alive can know": the other side may not have decided yet.

Now carry Kent's sorting to breakfast. Here are three sentences a presenter might say at a journal club; the last one is made up.

- "This paper pooled 88 estimates from 58 studies." You can check it by opening the paper: Kent's first kind.
- "In the countries these studies came from, people who skip breakfast more often have overweight." Many studies support it. But it reaches beyond the people counted, so it is a judgement on evidence, Kent's second kind.
- "Indian schoolchildren who skip breakfast today will have more obesity in 2040." No study in hand speaks to it: Kent's third kind.

The two sortings ask different questions, so use both. The first asks what a statement is doing. Kent's asks how close it sits to what is known. A finding usually holds a first-kind part, what was

counted, and a second-kind part, what it means beyond the people counted. A belief can sit at either of Kent's last two kinds. The words a speaker chooses should show which.

Figures quoted in this illustration are from [15].

WHERE THIS PICTURE BREAKS. Kent's "indisputable fact" is the closest a statement comes, not a guarantee. A photograph can be misread and a count can be wrong, as Book 0's section on random and systematic error showed. His 90 per cent is an example he gives to show a judgement has a degree, not a measured figure. How to put such numbers on your own judgements comes later in this book.

MUST KNOW POINTS FOR YOU.

- A finding is what one study saw, in its own people, with its own uncertainty. When you quote one, say whose it is and whom it was seen in, in the same sentence.
- Before you agree or disagree with a claim, sort it: hypothesis, finding or belief. Then ask the question that kind needs. For a hypothesis, what result would count against it. For a finding, who was studied and what was seen. For a belief, what it rests on.
- Watch for a belief dressed as a finding: "in fact", "studies show", "research proves", with no study named.
- When a resident's protocol says "this study will prove that…", have her rewrite it as a hypothesis: "we will test whether…".
- When a district health officer asks what is known, separate your sentences the way Kent did. Say what was counted, what the evidence makes likely, and what is a guess.
- Sorting a statement tells you what it is doing and what to ask next. It does not tell you whether it is true.

COMMON MISREADING. It is tempting to read the three as steps on one ladder of sureness. A hypothesis is a guess, a belief is firmer, and a finding is proven. That reading is wrong.

They differ in what the statement is doing, not in how sure anyone is. A hypothesis can rest on strong earlier evidence and still be a hypothesis, because it is offered to be tested. A finding is limited to its study, however large. And a belief can be held very strongly on almost no evidence. Brown and colleagues found the breakfast claim "only presumed true", although health websites stated it as fact. How sure to be is a separate question, and Kent's sorting is a first step towards it.

EXERCISE 1 (retrieval). From memory: in one sentence each, what are a hypothesis, a finding and a belief? Then name Kent's three kinds of statement. Before you check, say how sure you are.

EXERCISE 2 (critique). Sort each statement as a hypothesis, a finding or a belief. Then write the one question you would ask of it next. All four are made up.

1. A newspaper health column: "Studies prove that children who skip breakfast develop obesity."
2. A thesis protocol: "We hypothesise that a school breakfast programme lowers the share of children with overweight over two years, compared with schools without one."

3. A results section: “Of 412 students surveyed, 127 said they skipped breakfast on most days.”
4. A district health officer at a review meeting: “Breakfast skipping is surely behind the rise in obesity in our district.”

EXERCISE 3 (teaching). For a first-year resident, ten minutes with a whiteboard. Her draft protocol says:

> *This study will prove that skipping breakfast causes obesity in medical students.*

Tell her what kind of statement she has written, and help her rewrite it.

EXERCISE 4 (interpretation). In Kent’s airfield passage, the officer says the field is “almost certainly” military. Why does he not say “it is a military airfield”? And which of your own sentences, in your last lecture or journal club, should have carried a word like his?

2 · A hypothesis is testable: the prediction, and the result that would count against it

DEFINITION. A hypothesis is testable when a prediction follows from it: a statement of the form "if the hypothesis is true, then this observable result will follow", fixed before the result is seen. It is falsifiable when some possible result would count against it.

A test the hypothesis survives supports it and does not prove it.

A claim can be held in three different stances, which do not imply one another. It is accepted when it is taken as the best available description of what it describes. It is used when it is taken as an adequate tool for a practical job. It is pursued when it is thought worth developing and testing further. [16, 17, 2, 2, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32]

IN PLAIN TERMS. Turn it into an "if... then..." sentence: if this is true, then we should see that. Then name, before you look, the result that would count against it.

If you cannot name such a result, the hypothesis can never be caught out, and so it tells you nothing.

When the predicted result fails to show up in a fair test, that counts against the hypothesis. When it does show up, that supports the hypothesis but does not prove it, because something else could have produced the same result.

Illustration 1. Start with a hypothesis that cannot fail. It is made up for this section, but you will have heard its shape. A colleague says:

> *Breakfast helps weight control, as long as it is the right breakfast eaten the right way.*

Try to catch it out. Suppose a trial finds that people told to eat breakfast lose no more weight than people told nothing. He answers that they ate the wrong breakfast. Suppose a survey finds that breakfast eaters weigh more. He answers that they ate it the wrong way. Every result fits, because "right" is decided after the result is in.

So ask him one question: what result would make you drop it? If he cannot say, his sentence is not a hypothesis. It is a belief, and no study can touch it.

Now repair it. Fix in advance who, what, compared with what, measured how, and over how long:

> *Adults trying to lose weight who are advised to eat breakfast lose more weight over 16 weeks than adults given the same advice on healthy eating with nothing said about breakfast.*

Write the prediction as an "if... then..." sentence:

> *If this is true, then at 16 weeks the group advised to eat breakfast will, on average, have lost more weight than the group told nothing about breakfast.*

And name the result that would count against it: the two groups lose about the same, or the group told nothing loses more. The hypothesis can now be caught out. That is what makes it worth testing.

WHERE THIS PICTURE BREAKS. The repaired hypothesis is testable, which does not make it true or even likely. It also does not settle how big a difference would count as "more", or how many people a trial needs to see it. A test too small to see a real difference is not a fair test, and the book returns to that.

Illustration 2. Here is a real hypothesis that was put to a test. Dhurandhar and colleagues' 2014 trial is at https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/. It took adults with overweight or obesity who were trying to lose weight, at five research centres in the United States and Denmark. Chance put 309 of them into three groups. One was advised to eat breakfast, one to skip it, and a control group was given neither piece of advice. The trial ran for 16 weeks.

An earlier small trial had suggested that changing your usual breakfast habit, either way, helps weight loss. So the authors state this hypothesis in their introduction:

> *[we] hypothesized that individuals who were advised to switch their usual breakfast eating habits due to their group assignment would lose more weight than the control group.*

Write it as a prediction yourself. If the hypothesis is true, then people told to switch lose more weight than the control group. The result that counts against it: they lose about the same.

The abstract reports:

> *Treatment assignment did not have a significant effect on weight loss, and there was no interaction between initial breakfast eating status and treatment.*

"Not significant" is a statistical test's verdict: roughly, no difference too large to put down to chance. "Significant" means unlikely to be produced by chance alone, not large. The number behind it comes in section 4.

The trial sorted people by habit before assigning them. Usual eaters ate breakfast four or more times a week; usual skippers, three or fewer. Switchers were usual eaters told to skip and usual skippers told to eat. "No interaction" means the advice's effect on weight did not differ between usual eaters and usual skippers.

Switchers lost a little more than the control group: 0.05 kg more among usual skippers, 0.18 kg among usual eaters. But the authors' test found no difference beyond what chance easily produces. The predicted result did not clearly appear.

Set it out as an argument, as Book 0, F4 taught:

	The test that carries	The test that does not
First premise	if the hypothesis, then switchers lose more	if the hypothesis, then switchers lose more
Second premise	switchers did not lose clearly more	switchers lost clearly more
Conclusion	so the hypothesis is counted against	so the hypothesis is true

The left column is what happened. The inference is valid. But the conclusion is only as good as the first premise, and that premise quietly assumes a fair test: that people followed the advice, and that the trial was long enough and large enough to see a difference. The authors name two doubts themselves. Sixteen weeks "may not have been long enough" to see a very small effect. And some usual skippers in the control group began eating breakfast on their own, which "may have weakened our power to detect differences". So the result counts against the hypothesis, in adults like these, over 16 weeks. It does not abolish it.

The right column is the trap, and it is worth seeing although it did not happen. Had switchers lost more, the hypothesis would have passed a test. It would not have been proved, because something else could also predict that result. The next section lists what.

Figures quoted in this illustration are from [33].

WHERE THIS PICTURE BREAKS. This trial tested advice about breakfast, in adults in the United States and Denmark who were trying to lose weight, for 16 weeks. It did not test what breakfast does in children, in India, or over years. Reading its result as "breakfast does not matter" goes beyond it. How far a result like this one reaches is a question this book takes up later.

Illustration 3. Last, separate testing a claim from believing it. An open textbook by Barseghyan, Overgaard and Rupik, Introduction to History and Philosophy of Science, names three stances towards a claim. Its section 1.3 is at https://human.libretexts.org/Bookshelves/Philosophy/Philosophy_of_Science/Introduction_to_History_and_Philosophy_of_Science_(Barseghyan_Overgaard_and_Rupik).

Stance	What it means	The question that shows it
Accepted	taken as the best available description	Is this our best account of how things are?
Used	taken as an adequate tool for a practical job	Does this work well enough for the task in hand?
Pursued	thought worth developing and testing further	Is this worth more work?

The textbook writes of a scientific community's stances towards a theory. The same three fit one person and one claim.

The book's own example of use without acceptance is engineering. Physicists no longer accept classical physics (Newton's) as the best description of the world; general relativity replaced it. Bridges are still built with its equations, because they are simpler and good enough for a bridge. In its words, "It is possible to accept one theory but use another unaccepted theory in practice."

Now picture a resident who chooses breakfast skipping and weight as her thesis question. She is pursuing the hypothesis: she thinks it worth testing. She need not accept it, and she may privately doubt it. The book is plain on this: "a pursued theory is not necessarily accepted or

useful." If her result counts against the hypothesis, she has not failed. She has done what pursuing a hypothesis is for.

The stances can be combined in any way. The textbook says they "are not mutually exclusive". So when someone asks whether you believe your own hypothesis, answer the question they asked. "I am testing it" is a stance. "I think it is the best account we have" is a different one.

WHERE THIS PICTURE BREAKS. The three stances describe how a claim is held, not how good the evidence for it is. Accepting a claim as the best available does not make it true. In the textbook's words, a fallibilist community, one that holds any accepted claim may turn out wrong, "only accepts theories as the best on the market, not as infallibly true". The bridge example is the textbook's, from physics. In medicine, using a rule you do not accept needs the same honesty about why it is good enough for the job.

MUST KNOW POINTS FOR YOU.

- Write every hypothesis with its prediction, as "if this is true, then we will see that".
- "The result matched the prediction, so the hypothesis is proved" is the error. Write "supports" or "is consistent with", never "proves".
- When a claim is rescued after every result ("it was the wrong breakfast"), ask its owner what result would make them drop it.
- A result against the prediction counts against the hypothesis only as far as the test was fair. Before you write "the trial disproved it", check that people followed the advice.
- Tell a resident that pursuing a thesis hypothesis does not oblige her to believe it.
- When a committee asks whether you "believe in" a measure, say which stance you hold.
- A testable hypothesis need not be testable by experiment. When no one can assign the exposure, the thing whose effect is in question, test it by what it predicts you will observe.

COMMON MISREADING. "Falsifiable" sounds like "probably false", and "the hypothesis was counted against" sounds like "the researcher failed". Both readings are wrong.

A hypothesis is falsifiable when some possible result would count against it. That is the property that makes a test worth running. It says nothing about whether the hypothesis is false. And a result against a hypothesis is a finding like any other. Their trial counted against Dhurandhar and colleagues' hypothesis, and their report of that is what made the trial useful.

REPORTING IT. We hypothesise that [exposure, the thing whose effect is in question] leads to [outcome, what is measured to see the effect] in [people] over [time]. If this is true, then [the predicted result]. A result of [the result named in advance] would count against it.

EXERCISE 1 (retrieval). From memory: what makes a hypothesis testable, and why does a result that matches the prediction not prove it? Then name the three stances towards a claim, with one line each.

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (critique). A newspaper health column, made up for this exercise, says:

> *Breakfast switches on your metabolism, so people who skip it gain weight. Some studies find that skippers do not gain weight. That is because they make up for it with extra snacks later in the day.*

Find the step that stops this claim from ever being caught out. Then rewrite the core claim as a hypothesis with a prediction and a result that would count against it.

EXERCISE 3 (design). Take the main hypothesis of your own thesis, or of a study you would like to do. Write it in one sentence. Write its prediction as "if... then...". Write the result that would count against it. Then write which stance you hold towards it today, accepted, used or pursued, and one sentence on why.

EXERCISE 4 (teaching). For a journalist, three minutes. She asks: "So the breakfast trial proved that breakfast makes no difference to weight?" Answer her, and give her one sentence she can quote.

3 · One finding, several explanations

DEFINITION. An association is a finding that two things go together: one is found more often, or in larger amounts, where the other is present. An association observed in a study has several rival explanations, any of which can produce it.

Chance: sampling variation (Book 0, D6) produced a pattern in the sample that is not present in the population.

Bias: a systematic error (Book 0, D7) in how people were chosen or measured pushed the result one way.

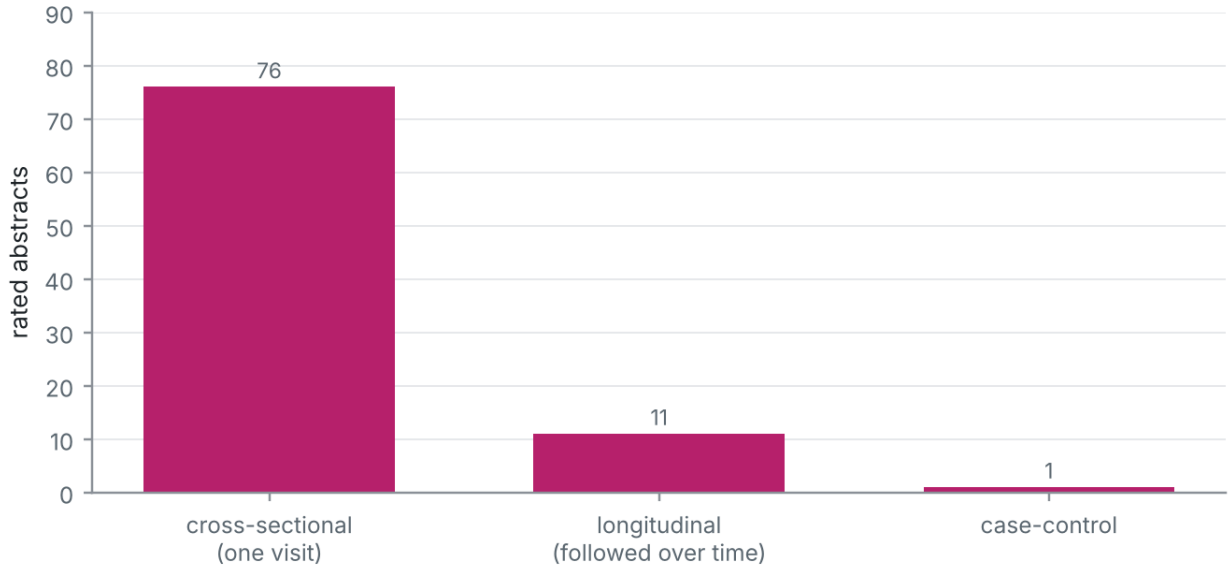
A third factor: something else is linked to both, and produces the association although neither causes the other. Epidemiologists call it a confounder.

Reverse direction: the second thing causes the first, not the first the second.

The proposed cause: the first thing changes the second. [34, 2, 4, 35, 5, 36, 37, 38]

IN PLAIN TERMS. Suppose a study finds that people who skip breakfast more often have overweight. Before you conclude that skipping breakfast puts weight on, ask what else could have produced the same pattern.

Write all five down. Then cross off only the ones the evidence actually rules out. Whatever is left is still a live explanation, however many studies found the pattern.



The breakfast studies repeated one kind of design. Brown and colleagues (2013) rated 88 abstracts, all of observational studies. Of these abstracts, 76 reported a study that measured breakfast habits and weight at the same visit, 11 a study that followed people over time and 1 a case-control study. Repeating a design makes chance a weaker explanation; it leaves the third factor and the reverse direction where they were.

ILLUSTRATION. Brown and colleagues' 2013 paper, at https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/, collected public statements about breakfast. One is from a newspaper, the Daily Mail:

> *Skipping breakfast to lose weight makes you fatter - and far more likely to raid the vending machine.*

That is a claim about cause. Now look at what the evidence under it was. Brown and colleagues pooled the studies of breakfast and weight, and found:

> *There is a clear association between breakfast omission and excess weight as summarized in our cumulative meta-analysis [...], and this association supports the PEBO but does not show causation.*

PEBO, as the first section said, is their name for the claim that eating breakfast protects against obesity.

Explanation	The question to ask	What these studies let you say
Chance	Could sampling variation alone have made it?	Hardly: it held in study after study
Bias	Were people picked or measured in a way that pushes one way?	Not settled by repeating the study; a bias the studies share repeats with them
A third factor	Does something else go with both breakfast and weight?	Open
Reverse direction	Could weight, or trying to lose it, change breakfast habits?	Open: most studies measured both at one visit
The proposed cause	Does skipping breakfast put weight on?	Open, until the three above are dealt with

Chance first. Brown and colleagues added the studies year by year. Their pooled association was already clear after the first year, from only 3 groups of people, and stayed clear every year to 2011. A pattern made by luck shrinks as more samples are pooled (Book 0, D6); this one did not. So chance is a weak explanation.

Now the third factor. You can supply candidates yourself, as questions to check rather than facts: a regular daily routine, or how much sleep a person gets.

Then reverse direction. Brown and colleagues had gathered 92 articles. 88 of them had an abstract in English, and they rated those. This 88 is a different count from the 88 pooled estimates of the first section; the match is chance. Of those 88 abstracts, 76 reported a cross-sectional study and 11 a study that followed people over time. The last 1 reported a case-control study, which starts from people with and without the condition and looks back.

A cross-sectional study measures both things at the same time: one visit, one count. It cannot show which came first. So a person who has put on weight and starts skipping breakfast to lose it would produce exactly the same pattern.

Brown and colleagues write: “All identified studies were observational.” Repeating that design study after study made chance a poor explanation. It did nothing to the third factor or the reverse direction, because every repeat carried the same weakness.

In a randomised controlled trial, chance decides who is told to eat breakfast and who is not, and then the groups are compared. Chance spreads routines, sleep and everything else evenly between the groups, apart from chance differences. And the advice comes before the weight change, so the direction is fixed.

Randomising decides who gets which advice. It does nothing about how weight is measured or who drops out, so bias can remain. Trials guard against it by keeping the people who weigh participants unaware of their group.

Figures quoted in this illustration are from [14].

WHERE THIS PICTURE BREAKS. The table is a checklist of rivals, not a ranking of how likely each is.

And a randomised trial deals with the third factor and the reverse direction, but chance stays. Any one trial can mislead by luck, which is one reason trials are repeated.

MUST KNOW POINTS FOR YOU.

- Before you accept that one thing causes another, write down the five explanations: chance, bias, a third factor, reverse direction and the proposed cause. Beside each, write what rules it out. Accept the cause only when the other four have something written beside them.
- “Dozens of studies agree, so it must be causal” is the error. Repeating the same design makes chance a weaker explanation and leaves every rival the design shares.
- A cross-sectional survey measures both things at one visit, so it cannot tell you which came first.
- When a resident presents an association at journal club as a cause, stop the discussion of mechanism. Ask her first to name one third factor and the reverse direction.
- To a journalist, say “people who skip breakfast more often have overweight, but these studies cannot tell us whether skipping breakfast is the reason”.
- The list tells you which rivals are still open, not which is most likely. Do not treat the one with most studies behind it as the winner.

COMMON MISREADING. It is tempting to think a third factor or a reverse direction means the study was done badly. It does not.

A survey can pick its people at random, measure them perfectly and count without a slip. Its association can still come from a third factor. The rival lives in the world the study watched, not in a mistake the researchers made. That is why Brown and colleagues could call the association “clear” and, in the same sentence, say it “does not show causation”. Better measurement of the same design does not remove the rival. A different design does.

REPORTING IT. Across many observational studies, people who skip breakfast more often have overweight than people who eat it. These studies show that the two go together; they cannot show that skipping breakfast is the cause.

EXERCISE 1 (retrieval). From memory: name the five explanations for an association. Which one does repeating a study with the same design make weaker, and which ones does it leave standing?

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (critique). A resident's survey, made up for this exercise, reports:

> *We surveyed first-year medical students in one college. Those who usually slept after midnight more often had overweight than those who slept earlier. We conclude that late sleep causes weight gain in medical students.*

List the rivals to her conclusion. For each, write one question you would ask her or one thing you would check.

EXERCISE 3 (interpretation). Brown and colleagues write that the observational association between skipping breakfast and excess weight "supports the PEBO but does not show causation". A randomised trial is then run. Which of the five explanations can such a trial remove, and which can it not? Say why for each.

EXERCISE 4 (teaching). For a health secretary, one page. An official proposes a state campaign telling adults to eat breakfast to prevent obesity. The note supporting it says "dozens of studies show breakfast eaters weigh less". Answer first: can the note's evidence carry the campaign's message?

4 · Belief beyond the evidence

Checked against the sources as they stood on 2026-10-10.

DEFINITION. A belief goes beyond its evidence when it is held or stated as a stronger kind of claim than the evidence can carry.

Brown, Bohan Brown and Allison (2013) measured this in one literature: the claim that skipping breakfast causes weight gain. They pooled 88 estimates from 58 observational studies and found the association between skipping breakfast and overweight or obesity firmly established. By 2011 the pooled P value, the number a statistical test prints beside a comparison, was below 10^{-42}. A P value that small means that if there were no association at all, results like these would almost never turn up by chance. They also found few randomised trials and no consistent evidence from them that breakfast affects weight. Their conclusion was that the claim is "only presumed true".

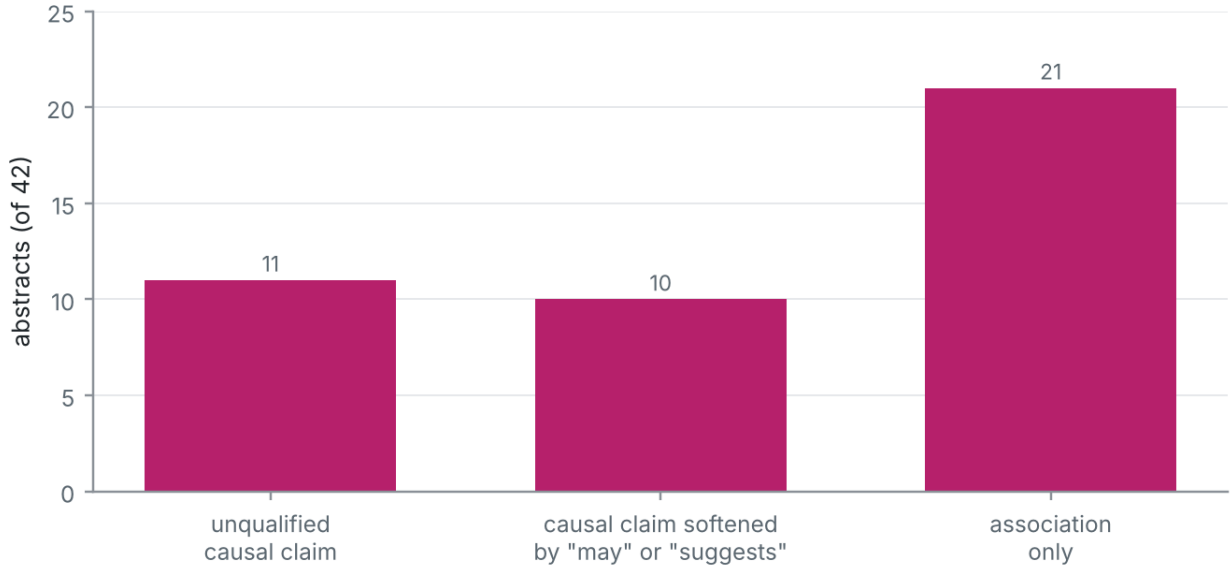
They named four ways the literature reported more than its evidence showed. Two are in how authors described their own results: a biased reading of them, and causal language for an observational finding. Two are in how authors cited other people's work: misleading citation of a trial's result, and causal language for another study's association. [3, 39, 40, 41, 5, 5, 42, 43, 44, 45, 46, 47, 48, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65]

IN PLAIN TERMS. Suppose many studies find that people who skip breakfast more often have overweight or obesity. That is an association: two things found together. It does not show that skipping breakfast causes the weight.

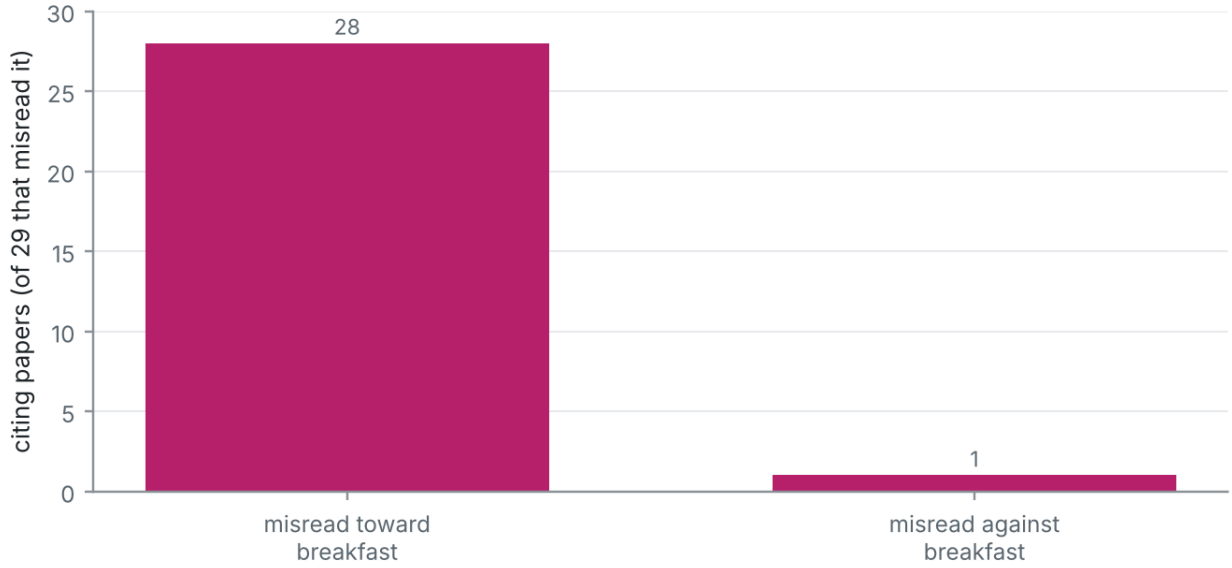
Now suppose people repeat the claim as a cause. Brown and colleagues suggest, without having tested it, that a claim met everywhere comes to feel proven. The evidence underneath has not changed.

Brown and colleagues measured it for breakfast and weight in 2013. Then a trial tested advice based on the claim and found no detectable effect of the advice on weight.

Your defence is two questions for any claim. What kind of statement is this? And what does it rest on?



Brown and colleagues (2013): how 42 abstracts of observational studies of breakfast and obesity worded their conclusion. 11 made an unqualified causal claim, 10 a causal claim softened by “may” or “suggests”, and 21 kept to an association. The authors judged causal language unwarranted for these designs, so between 26 and 50 per cent claimed more than their studies could show.



Brown and colleagues (2013) on citations of Schlundt’s 1992 trial, which found no main effect of breakfast. Of 46 papers that cited it about breakfast and weight, 29 misread it, and 28 of those 29 leaned toward breakfast.

Illustration 1. Start with sentences you could meet in any health column. Brown and colleagues collected them in 2013. Their article is at https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/. Find Table 2. These three come from the Mayo Clinic, the Academy of Nutrition and Dietetics, and the US Surgeon General.

> *In fact, skipping breakfast actually increases your risk of obesity.*

> *Want to trim your waist? Try eating breakfast!*

> *Eating a healthy breakfast is a good way to start the day and may be important in achieving and maintaining a healthy weight.*

Put the two questions to each one. First, what kind of statement is it? The first two are stated as facts. The third is softer: "may be important". Second, what does each rest on? None of them says. Each claims a cause, though: change what you eat in the morning, and your weight will change.

Now look at what was underneath. Brown and colleagues gathered the studies that earlier reviews had found. Search their Methods for "58 of 92".

> *In total, 58 of 92 studies fit these criteria with a total of 88 independent OR estimates.*

OR stands for odds ratio, one way of measuring how strongly two things go together. Independent means each estimate comes from a separate group of people, such as the men and the women of one study.

Every one of these studies was observational. The researchers watched what people already did and measured their weight. Nobody was given breakfast or asked to skip it. The studies came from 30 countries on 5 continents, and together they showed the association again and again.

Now search for "42 abstracts". Brown and colleagues also rated the abstracts of the articles they had gathered: 88 had one, and 42 of those drew a conclusion about breakfast and obesity.

> *of the 42 abstracts that included conclusions about breakfast and obesity, 11 abstracts made unqualified causal claims about breakfast consumption and obesity, whereas an additional 10 abstracts used qualifiers to introduce uncertainty into the causal statements; 21 articles limited the abstract conclusions to noncausal language*

conclusion of the abstract	abstracts
an unqualified causal claim	11
a causal claim softened by "may" or "suggests"	10
an association only	21
total	42

Work out the two shares the authors report as a range.

```
11 divided by 42 = 0.262, about 26 per cent
causal claims, flat or softened: 11 plus 10 = 21
21 divided by 42 = 0.5, which is 50 per cent
```

So between 26 and 50 per cent of these abstracts claimed more than their designs could show. Brown and colleagues judged causal language unwarranted for every one of these observational studies. The lower figure counts only the flat causal claims. The upper one adds the softened ones.

The trap is that the causal word sits in the abstract, where a reader who reads only the abstract meets it. A reader who sees 21 abstracts saying “causes” or “may cause” counts 21 votes for a cause. In fact, those 21 studies showed one association, measured many times.

A biased reading leans the conclusion towards the answer the authors favoured. Of the studies with a result favouring breakfast, 65 per cent put it in the abstract’s conclusion. Of the studies without one, only 39 per cent did.

Figures quoted in this illustration are from [14].

WHERE THIS PICTURE BREAKS. The counts describe one literature, breakfast and obesity, in the studies four earlier reviews had found, up to 2011. They show that causal language for an association happened and how often it happened there. They are not a rate for nutrition research as a whole, or for any Indian journal.

Two of the authors rated the abstracts separately and settled disagreements by discussion. The authors call the softened claims a judgement, which is why they give a range, not one figure.

Illustration 2. The second kind of overstatement happens one step later, when a paper cites another. Brown and colleagues followed the citations of one small trial, by Schlundt and colleagues in 1992. It was one of very few trials on the question.

First find out what the trial found. Women on a 12-week weight-loss plan were told to eat breakfast or to skip it. They were also split by whether they usually ate breakfast before the study. Search Brown’s Table 3 for “Schlundt”.

> *Habitual breakfast eaters assigned to skip breakfast and habitual breakfast skippers assigned to eat breakfast lost more weight than subjects who maintained their breakfast habits; no main effect of breakfast on weight loss.*

“No main effect” means this: eating breakfast, taken across everyone, did not change weight loss. The women who changed their usual habit, in either direction, lost more. That hint was weak. Brown reports its P value as below 0.06. The later trial’s authors give it as 0.06 and call it “only suggestive of significance”. Either way, it was not shown to fall below the usual cut-off of 0.05.

Now see how others cited it. Search Brown’s Results for “only 17%”.

> *Of the PEBO-relevant citations, only 17% of articles cited the results accurately. Of the 29 articles that misleadingly cited Schlundt et al (10), 28 articles mildly or explicitly cited Schlundt et al (10) positively toward breakfast with only 1 article that cited Schlundt et al (10) negatively toward breakfast*

("PEBO" is the authors' short name for the claim that eating breakfast protects against obesity. "(10)" is the number of Schlundt's paper in Brown's reference list.) Of 46 papers that cited the trial about breakfast and weight, only 17 per cent cited it accurately, and 29 misread it. The rest cited it in other ways. All but one of those misreadings leaned the same way: towards breakfast. A trial that found no main effect of breakfast was cited, again and again, as support for breakfast.

This is the failure Book 0's section on checking a reference (Book 0, F3) warned about: a citation that opens and is real, and does not say what it was cited for. You catch it only by the fourth act of that section, reading the passage.

Nor had anyone shown how breakfast could change weight. Brown and colleagues found "no clearly and consistently shown physiologic mechanism" for the claim.

Now the test. Brown and colleagues found one registered trial of the claim in October 2012: a trial whose plan was recorded in a public list of trials before it began. It was published in 2014 by Dhurandhar and colleagues, with David Allison, one of Brown's co-authors, among them (https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/). Chance decided which advice each person got: eat breakfast, skip it, or general advice on healthy eating alone. Search its abstract for "283 of the 309".

> *A total of 283 of the 309 participants who were randomly assigned completed the intervention. Treatment assignment did not have a significant effect on weight loss*

The advice worked on behaviour. People reported following it on 93.6 per cent of days in the breakfast group and 92.4 per cent in the skipping group. It made no detectable difference to weight over 16 weeks. The authors offer two readings of the earlier hint. It may have been a chance finding. Or Schlundt's tighter weight-loss plan left people less free to make up later in the day for what they did at breakfast.

Figures quoted in this illustration are from [14, 33].

WHERE THIS PICTURE BREAKS. The trial tested advice, not breakfast itself. Its 309 adults had overweight or obesity, were trying to lose weight, were recruited at sites in the United States and Denmark, and were followed for 16 weeks. It says nothing directly about children, about Indian adults, about longer periods, or about what kind of breakfast is eaten. And a trial that finds no clear effect has not shown that there is none; a later section of this book reads this result with care.

The citation counts cover one trial's citations, found in two citation databases in May 2012.

MUST KNOW POINTS FOR YOU.

- "Dozens of studies agree, so it must be true" is the error. Repeating a study of the same kind makes chance a weaker explanation of the association. It does not turn it into evidence of a cause.

- Read the verb before you read the number. “Increases your risk”, “protects against”, “leads to” and “helps you lose” all claim a cause. Over an observational finding, write “people who skip breakfast more often have overweight or obesity” instead.
- “May cause” is still a causal claim. It does not turn an association into evidence of a cause.
- When a slide or a paper cites a study for a claim, open the study and read what it found before you repeat the claim.
- When a causal claim turns out to rest only on observational studies, say so out loud: “that is an association; nobody has shown the cause yet”.
- Showing that a belief went beyond its evidence does not show that the belief is false. Until Dhurandhar’s 2014 trial, the right answer was “we do not know”, not “it is a myth”.
- Brown’s counts come from one question, breakfast and weight. They are not a rate for nutrition research in general.

COMMON MISREADING. “Brown and colleagues proved that breakfast does not matter for weight.” They did not try to. They showed that the belief was stronger than its evidence: the association was firmly established, and the cause was not. The trial that followed tested one thing: advice to eat or skip breakfast, in adults trying to lose weight. Over 16 weeks the advice had no detectable effect on weight. Replacing “breakfast helps you lose weight” with “breakfast is useless” swaps one belief beyond the evidence for another.

REPORTING IT. Observational studies have repeatedly found that people who skip breakfast more often have overweight or obesity. This is an association. One large trial gave adults trying to lose weight advice to eat or skip breakfast. Over 16 weeks the advice made no detectable difference to weight (Dhurandhar et al., 2014).

EXERCISE 1 (retrieval). From memory: name the four ways Brown and colleagues found the breakfast literature stating more than its evidence showed. Which two are about authors’ own results, and which two about citing others?

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (critique). A slide in a postgraduate lecture reads (made up for this exercise):

> *Skipping breakfast causes weight gain (Schlundt 1992). Eating breakfast helps people keep weight off: 78% of the National Weight Control Registry eat it daily (Wyatt 2002).*

Brown and colleagues describe both studies. The registry study, by Wyatt and colleagues, is a case series. A case series describes one group of people who share something, here weight lost and kept off. It has no comparison group. So it can show only that two things occur together in these people, not that they go together more than they do in others. Say what each cited study found, what is wrong with how the slide uses it, and rewrite the slide.

EXERCISE 3 (interpretation). Brown and colleagues followed 72 papers that cited the registry study by Wyatt and colleagues about breakfast and weight. By their count, 29 per cent of these papers described breakfast and kept-off weight as “simply co-occurring”. Another 26 per cent

called them “associated”. And 22 per cent linked them as cause and effect. The rest used softened wording. What does each wording claim, and which is right for a case series?

EXERCISE 4 (teaching). For a journalist, three minutes, no jargon. She is writing a column for a Sunday paper and asks: “Is it true that skipping breakfast makes you fat?” Give her one quotable sentence, say how sure you are in plain words, and say why.

5 · Two kinds of support for a causal claim: a difference between groups, and a mechanism

DEFINITION. A causal claim says that changing one thing, the exposure, changes another, the outcome.

Evidence of a difference between groups compares groups that differ in the exposure and asks whether the outcome differs between them. Other things equal, it is strongest when chance decided which group each person joined. Howick, Glasziou and Aronson (2009) call this direct evidence: evidence that an association between an exposure and an outcome is causal and not spurious, that is, not produced by chance, bias or a third factor.

Mechanistic evidence is evidence for the chain of steps by which the exposure is said to produce the outcome. Proposing a mechanism is not evidence that it operates.

A mechanism with no measured difference between groups is a hypothesis about the outcome, however well each step is understood. Howick and colleagues hold that mechanistic evidence should play a supporting role to direct evidence. They also name a third kind, parallel evidence: other studies, replications and similar ones, that give similar results.

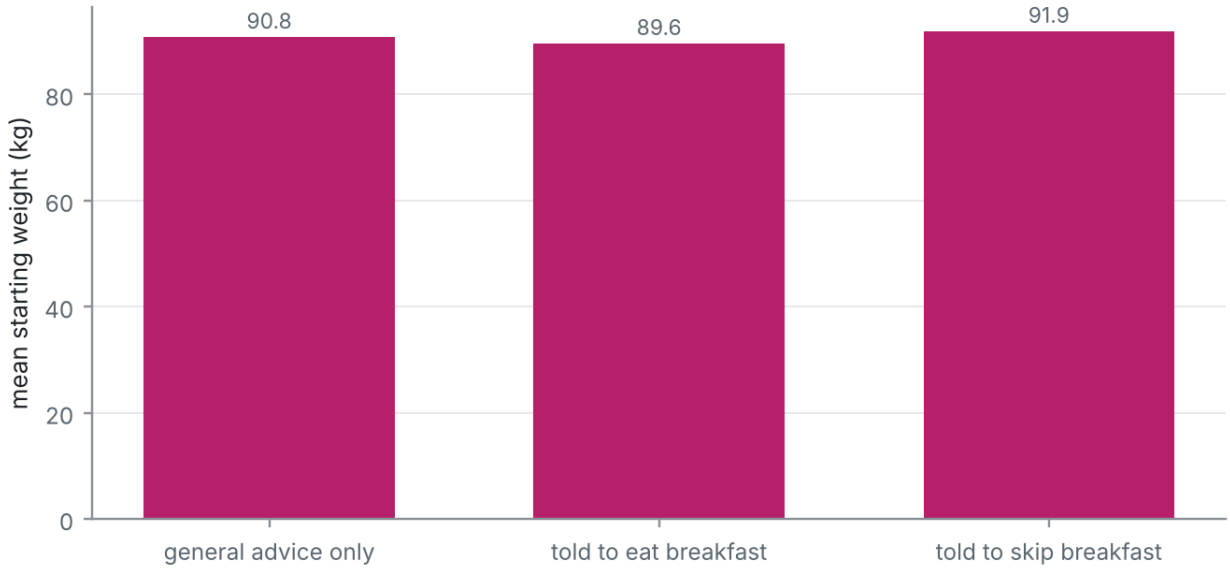
Their three kinds are a regrouping of the considerations Austin Bradford Hill listed in 1965 for judging whether an association is causal. [66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 77, 78, 79, 37, 37, 80, 25, 81]

IN PLAIN TERMS. There are two ways to show that one thing causes another.

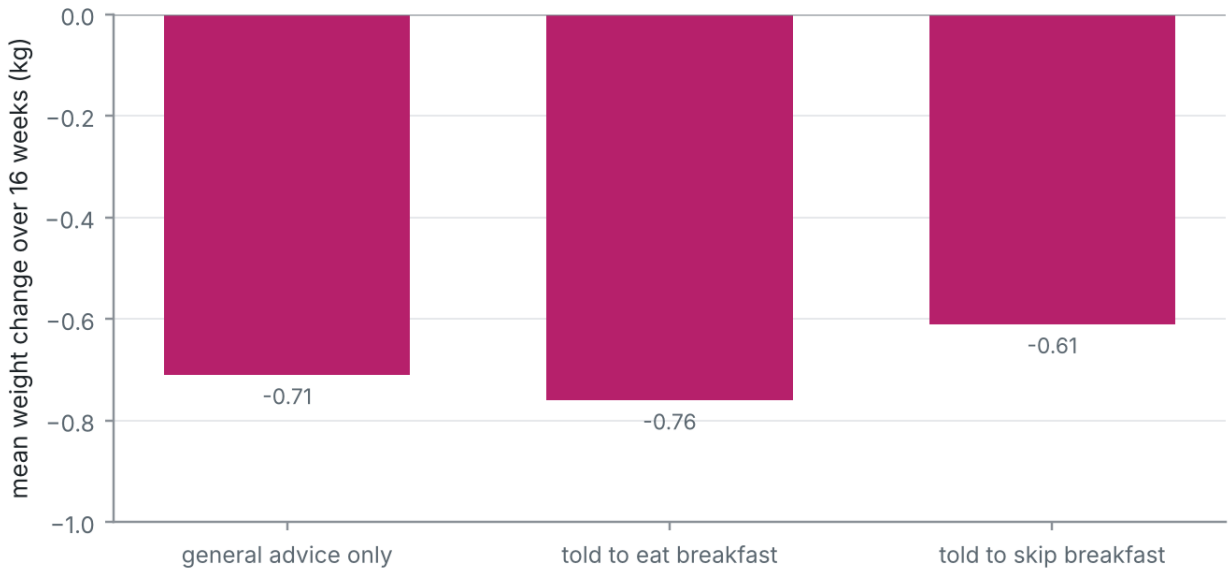
The first is to compare. Take two groups that differ in the thing, and see whether the result differs. Let chance decide who goes into which group.

The second is to explain how. Lay out the steps by which the thing would act in the body, and find evidence for each step.

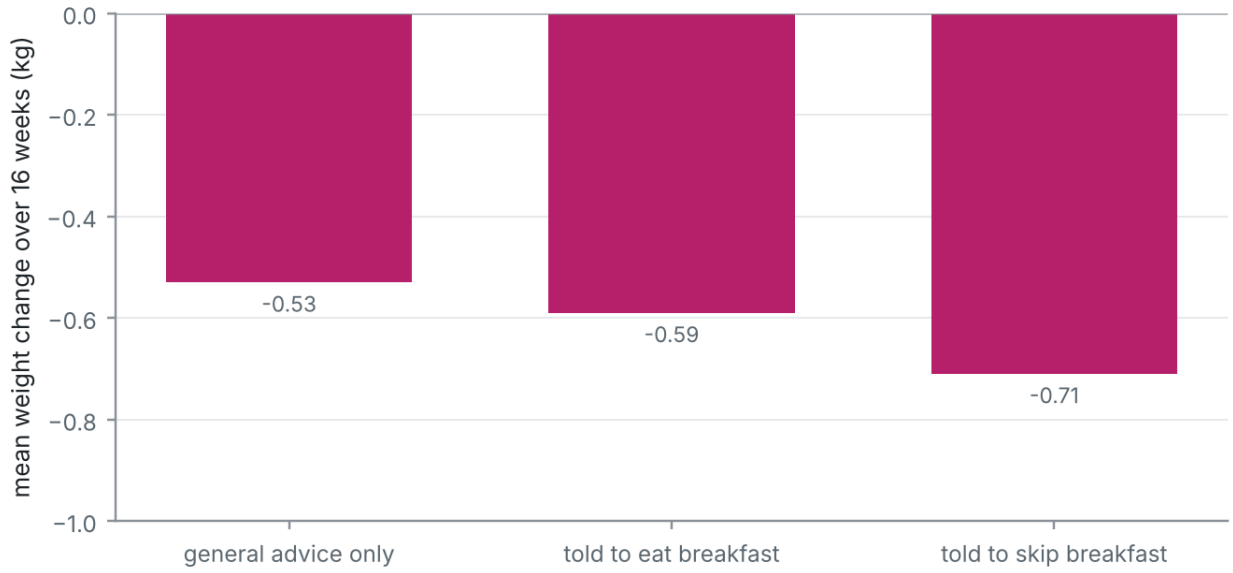
Until someone compares groups and finds the difference, the story is a hypothesis.



Dhurandhar and colleagues (2014), Table 1: mean starting weight in each arm. 90.8 kg with general advice, 89.6 kg told to eat breakfast, 91.9 kg told to skip it. Chance made the groups, and they started out close: within 2.3 kg of each other.



Usual breakfast skippers in Dhurandhar and colleagues' trial: adjusted mean weight change over 16 weeks. Told to eat breakfast, they lost 0.15 kg more than those told to skip it.



Usual breakfast eaters in the same trial. Here it went the other way: told to skip breakfast, they lost 0.12 kg more than those told to eat it. The authors' test of all three arms together found no difference larger than chance (P = 0.77).

Illustration 1. The section before this one showed that the breakfast claim rested on observational association. Its defenders also had a mechanism. Build it first, as an argument in the sense of Book 0's section on argument (Book 0, F4): a chain of links, the claim the conclusion.

step	link
1	Eating breakfast makes you less hungry later in the day.
2	So you eat less at lunch and afterwards.
3	So you eat less over the whole day.
4	Eating less over days and weeks lowers weight.
conclusion	Eating breakfast lowers weight.

It reads well. Every step sounds like physiology. Now test each link, because each is a claim that needs its own evidence.

Brown and colleagues (2013) collected the short studies that measured these links. Their article is at https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/. Search it for "subsequent eating occasion".

> *Although some studies showed a decreased energy intake at a subsequent eating occasion after eating compared with skipping breakfast, none of the studies showed a decrease in the sum of calories consumed across the study period when breakfast was eaten rather than skipped (Table 4).*

So step 2 had some support. Step 3 had none. In two of the studies in their Table 4, by Kral and by Levitsky, total intake was higher when breakfast was eaten. In the second of Levitsky's experiments, people who skipped breakfast ate more at lunch, but not enough to make up the break-

fast they had missed. Brown and colleagues sum it up: "there is currently no clearly and consistently shown physiologic mechanism".

Notice where the argument broke. Every link up to step 2 can hold and the conclusion still fail, because the breakfast itself can add more than eating less at lunch saves. A mechanism argument is a chain, and it is only as strong as its weakest link.

Now notice what the chain could never do on its own. Even with every link supported, it would tell you how breakfast might lower weight. It would not tell you that it does, or by how much, in real people living their lives. Only a comparison of groups can tell you that.

WHERE THIS PICTURE BREAKS. The studies of these links were short: in Brown's Table 4, most lasted one day and the longest two. A link that fails over one day might hold over months, or the reverse. That is one reason a mechanism, however studied, cannot stand in for a measured difference in weight.

The four-step chain is this section's own summary of the energy-intake mechanism Brown and colleagues describe. Other mechanisms were proposed, such as effects of the time of day on metabolism. Brown's verdict covers those too.

Illustration 2. Now the other leg. Dhurandhar and colleagues (2014) compared groups whose breakfast advice was decided by chance (https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/). Start with why that matters.

In an observational study, breakfast eaters choose to eat breakfast. They may differ from skippers in other ways that affect weight: sleep, work hours, activity, income. The trial's authors say this in their introduction:

> *observational evidence does not preclude the possibility that breakfast eaters tend to weigh less because of other weight-related factors associated with breakfast eating.*

Set out the comparison as an argument too.

step	premise
1	The groups differ in the exposure: here, the breakfast advice.
2	The groups are alike in everything else that affects the outcome.
3	The outcome differs between the groups by more than chance alone would give.
conclusion	The exposure changed the outcome.

If all three premises are true, the conclusion follows. In an observational comparison, premise 2 is the one in doubt, for the reason the trial's authors give.

When chance decides the group, those other factors are spread between the groups by chance, not by choice. It uses the same tool as a random sample in Book 0's section on sampling (Book 0, D6): an actual chance process, not a person, decides. Check that it worked. Search the trial's Table 1 for "Weight (kg)". The mean starting weights were 90.8 kg with general advice, 89.6 kg with advice to eat breakfast and 91.9 kg with advice to skip it.

```
91.9 minus 89.6 = 2.3
```

The three means lie within 2.3 kg of each other. The authors' test of that gap gave P = 0.57: no difference beyond what chance gives. The groups started out close.

Now the result. Each person was told one of three things: general advice on healthy eating alone, the same plus "eat breakfast", or the same plus "skip breakfast". The authors also split them by their habit before the trial. Here are the mean weight changes over 16 weeks, in kg, for the 283 who completed it. A minus sign means weight went down. The authors adjusted each mean for starting weight, age, sex, site and race.

habit before the trial	general advice only	told to eat breakfast	told to skip breakfast
usually skipped breakfast	-0.71	-0.76	-0.61
usually ate breakfast	-0.53	-0.59	-0.71

Compare the two breakfast arms (the trial's groups) within each row. Among those who usually skipped breakfast, the group told to eat it lost a little more.

```
0.76 minus 0.61 = 0.15
```

Among those who usually ate breakfast, it went the other way: the group told to skip it lost a little more.

```
0.71 minus 0.59 = 0.12
```

Differences of 0.15 kg and 0.12 kg, in opposite directions. In both rows, those told to switch their habit lost a little more than the general-advice group, as the authors had predicted. But their test of all three arms together found no difference in weight larger than chance would explain (P = 0.77). Nor did the advice act differently on usual skippers and usual eaters beyond chance (P = 0.70).

Here premise 3 was not met. So by 2014 both legs were weak for this claim. The mechanism's key link had not been shown. The one large comparison of groups made by chance found no detectable difference in weight.

Figures quoted in this illustration are from [33].

WHERE THIS PICTURE BREAKS. Chance spreads other factors evenly on average, not in every trial. A small trial can be unbalanced by bad luck, which is why the authors' Table 1 is worth checking, as you just did.

These are adjusted mean changes among completers, in adults with overweight or obesity trying to lose weight, at five sites, one of them in Copenhagen, over 16 weeks. The trial did not report a range of plausible values for the difference between arms, so read 0.15 and 0.12 kg as "small", not as the true size of any effect.

Illustration 3. The opposite combination also happens: a strong difference between groups, with the mechanism unknown. Howick and colleagues give the classic case. Search their article (https://pmc.ncbi.nlm.nih.gov/articles/PMC2677430/) for "black box".

> *For example, Doll and Hill's famous study of the relation between the number of cigarettes smoked and the incidence of lung cancer38 did not refer in any way to what happens between inhalation of cigarette smoke and the development of tumours in the lung.*

(The "38" run into the text is a reference number. Incidence means new cases in a period.) The study compared groups that smoked different amounts. The death rate from lung cancer in smokers was, as Howick and colleagues report it from Hill, "over 9 times the rate for non-smokers". Later studies, which Howick and colleagues also cite, found that the risk fell when people stopped smoking: by half in those who stopped at age 50, and almost completely in those who stopped at age 30. That is a difference between groups too, run the other way.

Howick and colleagues count this comparison as good evidence of a cause, with no account of the steps inside the lung. That is the point of keeping the two legs apart. A missing mechanism does not cancel a strong difference. A strong mechanism does not supply a missing one.

The Hill of Doll and Hill is Austin Bradford Hill. In 1965 he set out a list of considerations for judging whether an association is a cause. Howick and colleagues regroup it into the kinds of evidence in this section. The list itself is taught at the next rung, the next level of this subject.

Figures quoted in this illustration are from [82].

WHERE THIS PICTURE BREAKS. Smokers chose to smoke, so this comparison was observational, and a third factor could in principle have produced it. Howick and colleagues argue that a large effect can have "swamped the combined effects of any plausible confounders", and that dose-response evidence is strongest "when the process is reversible". A third factor would have to be linked to both smoking and lung cancer, more strongly than smoking is linked to lung cancer. It would also have to fall away when people stop. A small difference between self-chosen groups, like the breakfast association, carries far less weight. Size and design both matter, and weighing them is taken further in the next section.

MUST KNOW POINTS FOR YOU.

- "There is a clear mechanism, so it must work" is the error. The chain can hold at every link you tested and still fail at the outcome, because people can eat more later, as they did at lunch in the second of Levitsky's experiments. Breakfast did lower lunch intake in some studies. No study showed lower intake over the whole day.
- Was a difference in the outcome measured between groups that differ in the exposure, and who decided who was in which group? Is there evidence for each step of how it would work? Then say which leg is weak.
- A study that measures one link, such as how much people eat at lunch, is not a study of the outcome, weight.

- Do not advise a patient to change what they do on a mechanism alone. Say it is a hypothesis that has not been tested on weight.
- "We know how it works" is not "we know it works".
- The two-leg check sorts evidence by kind. It does not weigh it.
- When a resident cites Bradford Hill's list as a checklist that proves a cause, stop them. Hill himself offered his points as viewpoints to consider, and the list is taught at the next rung. Here, ask only the two questions: is there a measured difference, and is there a shown mechanism?

COMMON MISREADING. "The breakfast trial found no effect, so the mechanism was wrong." Not necessarily. One link can be true, as the drop in lunch intake was in some studies. The outcome can still stay the same, because intake elsewhere in the day does not fall to match. The trial measured the outcome, not the links. A null comparison tells you the chain did not deliver a detectable change in weight under those conditions. It does not tell you which link failed.

REPORTING IT. The evidence that eating breakfast lowers weight is weak on both legs. The large trial that assigned breakfast advice by chance found no clear difference in weight. And no mechanism linking breakfast to a lower total daily intake has been consistently shown.

EXERCISE 1 (retrieval). From memory: what are the two kinds of support for a causal claim? Which of the two can a claim not do without, if it is to be more than a hypothesis? And what is the third kind that Howick and colleagues add?

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (critique). A slide in a postgraduate seminar on meal timing reads (made up for this exercise):

> *Late dinners cause obesity. The body handles food differently at night, so calories eaten late are more likely to be stored as fat.*

You do not need to know whether the physiology is right. Say which leg of support the slide offers, which it leaves out, and what the slide may honestly claim.

EXERCISE 3 (design). Take the late-dinner claim from the last exercise. Specify the comparison that would test the difference-between-groups leg. Say who is compared, who decides which group each person is in, what is measured, and for how long. Then name one way your design could still mislead.

EXERCISE 4 (teaching). For a first-year resident, ten minutes with a whiteboard. He says: "Breakfast reduces hunger at lunch, so of course it helps weight loss." Use the breakfast evidence to show him the two legs and where each one stood by 2014.

6 · Why evidence comes in strengths

DEFINITION. Evidence supports a conclusion by inductive reasoning: it goes from particular observations to a general conclusion. An inductive conclusion can be supported more strongly or less strongly. No amount of evidence makes it certain, because what is known can change with further information.

A body of evidence is all the studies that address one question for one outcome. How far a body of evidence can be trusted is judged for each outcome separately.

Six features decide how far a body of evidence for one outcome can be trusted. The last five are named by the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group, whose scheme is taught in the next section. In GRADE, design sets where the rating starts; the other five can lower it.

feature	the question to ask
design	Were the groups made by chance (a randomised controlled trial) or observed as they came (an observational study)?
risk of bias	Do flaws in how the studies were run push the estimate one way, such as many people lost before the end?
inconsistency	Do the studies give widely different results that nobody can explain?
indirectness	Do the studies answer a nearby question: other people, another version of the treatment, another comparison, or a stand-in outcome?
imprecision	Did the studies include few people or few events, so that the range of values the data allow is wide?
publication bias	Are studies missing because their results were never published?

A surrogate outcome is a measurement that stands in for the outcome that matters to patients. A confidence interval is the range of values the data still allow, at a stated level such as 95 per cent.

Some features can raise trust in observational evidence. Two are taught here. One is a large effect, which bias is unlikely to explain in full. The other is a dose-response gradient: more of the exposure goes with more of the effect. GRADE adds a third, taught in the next section.

Expert opinion is not a kind of evidence. [83, 84, 85, 86, 75, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 105, 106, 107, 108, 109, 109, 109, 110, 111]

IN PLAIN TERMS. A study shows you something that happened in some people. A conclusion says it will happen in people in general. That step from some to all is never certain. So evidence makes a conclusion stronger or weaker, never proven.

How strong? Take all the studies on one question, for one result you care about, and ask six things.

The name of the journal, the fame of the study and the seniority of the speaker are not on the list.



Balshem and colleagues' worked example: a number needed to treat of 6, with a 95 per cent confidence interval from 4 to 13. The upper end is 3.25 times the lower: one person helped in every 4 treated at one end, one in every 13 at the other.

Illustration 1. Start with two beliefs that doctors held strongly and that turned out wrong. Both are told by Guyatt and colleagues, in the first of a series of papers in the BMJ (the British Medical Journal) explaining GRADE, in 2008.

For a decade, organisations told clinicians to encourage women past the menopause to take hormone replacement therapy (HRT). The belief behind the advice was that HRT substantially lowered the risk of heart and blood-vessel disease. Then randomised controlled trials (a way to test a treatment in which chance decides who gets it, and then the groups are compared) were done. They showed that HRT fails to reduce that risk and may even increase it.

Take the six questions to the evidence as it stood before the trials.

Design. The studies were observational. The women chose whether to take HRT. Whatever led a woman to choose it could also affect her heart, which is the third factor of the section on one finding with several explanations. That question fails.

Inconsistency. Guyatt and colleagues say the observational results were inconsistent. That question fails too.

Their verdict: had the evidence been rated at the time, it would have been rated very low, the lowest of GRADE's four levels, taught in the next section. A strong belief stood on weak evidence, and nobody had rated the evidence by a system like GRADE.

Now the second case. The United States Food and Drug Administration licensed two heart drugs, encainide and flecainide. It did so because they reduced an irregular heart rhythm that is linked with sudden death. Ask the directness question. What did the studies measure? The rhythm. What mattered to patients? Death. The rhythm was a surrogate outcome. A later randomised trial showed that the two drugs increase the risk of sudden death.

So the two failures broke on different questions: HRT on design and consistency, the heart drugs on directness. Neither would have been caught by asking whether the studies were published in good journals.

WHERE THIS PICTURE BREAKS. Both cases are read with hindsight, by authors who were arguing for GRADE. The paper's own competing-interests statement says all its authors help spread GRADE and gain from its success. Weak evidence is not always wrong. Evidence that fails these questions has not been shown false; it has not been shown true either. A belief can rest on weak evidence and still turn out right.

Guyatt and colleagues also say that attention to the low quality of the evidence on the heart drugs would have saved thousands of lives. That is their estimate, and they give no count.

Illustration 2. The six questions are not an arbitrary checklist. You can rebuild them yourself. Balshem and colleagues, writing for the GRADE Working Group in 2011, set up the test.

Suppose you are told that a Cochrane review (a systematic review, a study of studies that collects and combines the results already published on one question, in a planned and careful way) has a result for people with long-lasting pain. A salicylate cream rubbed on the skin, against a dummy cream, gives a number needed to treat of 6. The number needed to treat (NNT) is how many people must be treated for one more of them to do well. Its 95 per cent confidence interval runs from 4 to 13.

Before you read on, write down what you would want to know before you believe that 6.

Here are the questions Balshem and colleagues list. How many studies were pooled, and how many patients? How wide is the range around the 6? Were the studies randomised? Did they have flaws, such as no blinding or many people lost before the end? Blinding means that the people in the study, or those measuring them, do not know who got which cream.

Then more. Were other studies done that the reviewers could not get? Did the trials give similar results or very different ones? Was the outcome measured at a sensible time, or were the studies too short? On what part of the body was the cream used?

The first questions, they say, belong to imprecision and risk of bias. The later ones belong to publication bias, inconsistency and indirectness.

Now see why each belongs. Each is a way the 6 could be wrong. Chance could have produced it: that is the width of the range. Look at how far apart the two ends are.

```
13 divided by 4 = 3.25
```

At one end, one person in 4 is helped who would not otherwise be. At the other end it is one in 13, about a third as many. Bias in how the studies were run could have produced it: that is design, which sets the start, and risk of bias. Studies that found nothing may be missing: publication bias. The studies may disagree: inconsistency. And the 6 may be the right answer to a different question: indirectness.

So the list is what you get when you ask, for one estimate, every way it could be wrong. Ask it of every result you are about to repeat.

Figures quoted in this illustration are from [112].

WHERE THIS PICTURE BREAKS. Balshem and colleagues say the five named categories cover "nearly all" the issues, and that they were narrowed to five by discussion and agreement among

GRADE's members. So the list was agreed, not proved complete. The rebuilding shows why each question belongs. It does not show that nothing else could ever matter.

The example is theirs, told as "suppose you are told". Treat the 6 and its range as a worked example, not as a finding about salicylate creams.

MUST KNOW POINTS FOR YOU.

- "Strong evidence" belongs to a body of evidence for one outcome. When someone calls evidence strong, ask: for which outcome, and which of the six questions does it pass?
- Before you put a finding on a slide or into your thesis, answer the six questions in one line each. They are design, risk of bias, inconsistency, indirectness, imprecision and publication bias (missing studies).
- A surrogate outcome can move the right way while the outcome that matters moves the wrong way. Ask what was measured before you ask how much it changed.
- Do not say that evidence "proves" a claim, and correct a trainee who does. Say how strongly, and why.
- One set of studies can give strong evidence on one outcome and weak evidence on another. So say which outcome you mean whenever you say how strong the evidence is.
- Ask the indirectness question of every such body of evidence: were these the people, the food and the setting you are talking about? Say so in the sentence when they were not.
- The six questions tell you how far an estimate can be trusted. They do not tell you whether to act on it, which also needs the harms, the costs and what people value.

COMMON MISREADING. "I have twenty years of experience with these patients, and that is my evidence." Balshem and colleagues take this case apart with an endocrinologist on a ward round. He says patients feel better soon after he starts treatment, and that his patients whose blood sugar is kept close to normal (tight control) have little of the eye, kidney and nerve damage that diabetes causes (complications).

The opinion he forms is not evidence at all. The experience behind it is evidence, but of the weakest kind. It is an uncontrolled case series, and they rate it very low.

His memory may be wrong. He may forget the patients who did not improve. People may get better on their own, or say what they think the doctor wants. Respect the experience and ask what it rests on.

EXERCISE 1 (retrieval). From memory: name the six features that decide how far a body of evidence can be trusted, with the question each one asks. Then name the two features that can raise trust in observational evidence. Last, say in one sentence why no amount of evidence makes a conclusion certain.

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (critique). A made-up slide from a postgraduate lecture reads:

> *Skipping breakfast causes obesity. Strong evidence: 20 studies agree, published in leading journals.*

Find each thing wrong with the claim of strength. Then write what the slide would have to say before you could judge it.

EXERCISE 3 (teaching). A journalist rings you about a new study reporting that people who eat late at night gain more weight. She asks: “So is it proven?” You have three minutes and no jargon. Give the answer you would say, ending in one sentence she can quote.

7 · GRADE certainty: four levels, where you start, and moving down or up

Checked against the sources as they stood on 2026-10-10.

DEFINITION. GRADE (Grading of Recommendations Assessment, Development and Evaluation) is a scheme for rating a body of evidence.

GRADE calls its rating the certainty of the evidence. Older texts, including GRADE's own 2013 Handbook, call it the quality of evidence. In a systematic review, certainty is a judgement of how sure one can be that an estimate of an effect is close to the true effect. It is rated for each outcome, across all the studies that measured it.

GRADE has four levels.

level	printed as	GRADE's meaning, in short
high	⨁⨁⨁⨁	very confident that the true effect lies close to the estimate
moderate	⨁⨁⨁◯	the true effect is likely to be close to the estimate, but may be substantially different
low	⨁⨁◯◯	the true effect may be substantially different from the estimate
very low	⨁◯◯◯	the true effect is likely to be substantially different from the estimate

This book counts each level by its filled circles: high 4, moderate 3, low 2, very low 1.

A body of randomised trials starts at high. A body of observational studies starts at low. An uncontrolled case series is observational, so it starts at low. With a newer bias tool, ROBINS-I (Risk Of Bias In Non-randomised Studies of Interventions), non-randomised studies may start at high. They are then generally rated down two levels for having no randomisation, so they land at low as before.

Five reasons can lower the level: risk of bias, inconsistency, indirectness, imprecision and publication bias. A reason judged serious lowers it by one level; judged very serious, by two. Nothing goes below very low. Publication bias is suspected, for example, when the evidence is a number of small studies, all paid for by industry: studies that find no effect are less often reported.

A risk ratio is the risk (the share of a group who have the outcome over a set time) in one group divided by the risk in another; 1 means no difference. Three reasons can raise the level, mostly for observational evidence; the first, a large effect, comes in two strengths:

reason to rate up	how far
a large effect: a risk ratio above 2 or below 0.5	up 1
a very large effect: a risk ratio above 5 or below 0.2	up 2
a dose-response gradient	up 1

reason to rate up	how far
all plausible bias works against the result: it would shrink an effect that was found, or would have made one appear where none was found	up 1

If the estimate passes 2 but its confidence interval reaches appreciably below 2, Cochrane says some would hesitate to rate up.

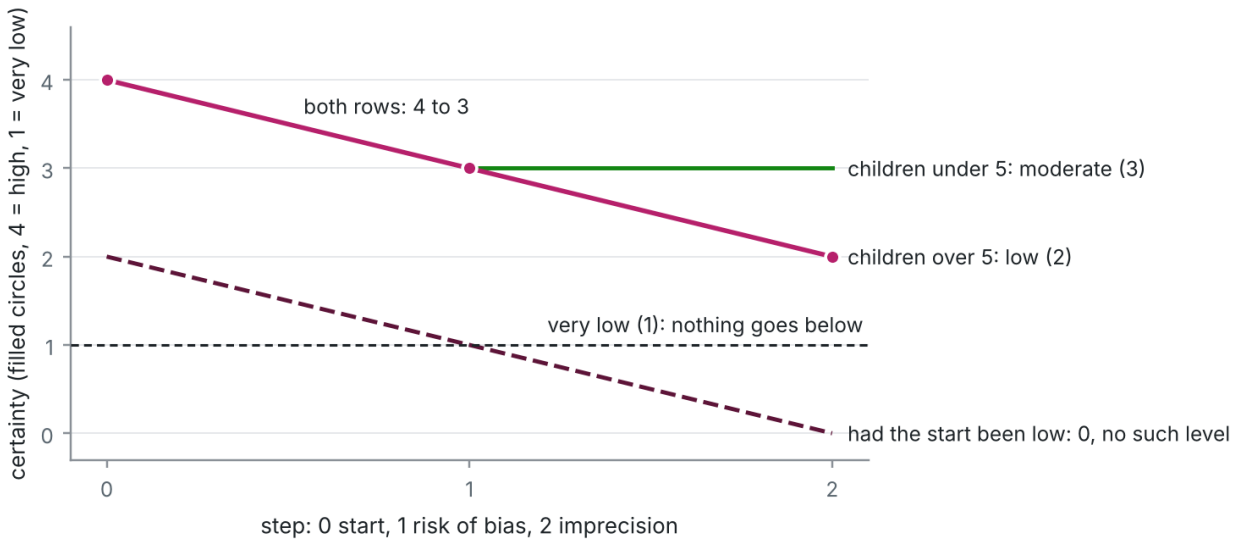
Two careful raters can end one level apart. So GRADE asks raters to explain every step down or up in a footnote. [113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147]

IN PLAIN TERMS. GRADE is a set of rules, agreed by a group of researchers, for saying how much to trust the evidence on one question. It gives one of four labels: high, moderate, low or very low.

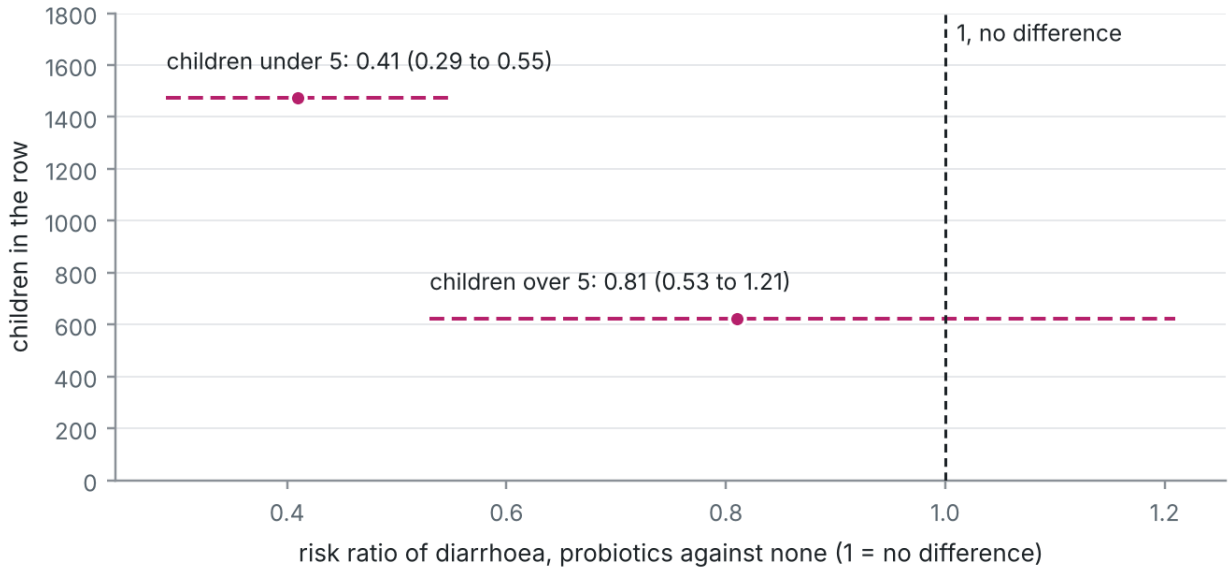
You start from how the studies were done. If chance decided who got the treatment, start at high. If the researchers only watched what people did, start at low.

Then you look for problems. Five kinds of problem can take any body of studies down. For watched groups, a few things can take you up.

Each step is a judgement, so the rater writes down why.



Counting the steps for the two rows of the Handbook's example table (Figure 14.1.b), with high as 4 and each step down taking 1 away. Both rows start at high (4) and lose one level for risk of bias (3). Only the row for children over 5 loses a second, for imprecision, and ends at low (2). Had the rows started at low, two steps would reach 0, below the floor of very low (1); so both started at high, as randomised trials do.



Why only the row for children over 5 lost a step for imprecision. Each dot is a risk ratio and each line its 95 per cent confidence interval, placed at the number of children in the row. Children under 5 (1474): 0.41, 0.29 to 0.55, wholly below 1. Children over 5 (624): 0.81, 0.53 to 1.21, which holds 1, no difference. The axis is plain; on the ratio (logarithmic) scale each interval sits about evenly around its dot. Numbers from the Cochrane Handbook's example table, Figure 14.1.b.

Illustration 1. Start with a mistake you will hear at journal clubs. The presenter points at a table of results and says: "These are randomised trials, so the evidence is high certainty." The presenter is made up; the table is real.

Open chapter 14 of the Cochrane Handbook for Systematic Reviews of Interventions, at https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14, and find Figure 14.1.b. It is an example 'Summary of findings' table. It compares probiotics with no probiotics in children given antibiotics. The outcome is how many children get diarrhoea.

Read the column headed "Certainty of the evidence (GRADE)". The row for children under 5 says moderate, due to risk of bias. The row for children over 5 says low, due to risk of bias and imprecision. The footnotes give the reasons: a high loss to follow-up (many children not followed to the end), then few events and a wide confidence interval. None says "very serious", so each reason cost one level.

Now count, with high as 4 and each step down taking 1 away.

```
children under 5: 4 minus 1 = 3, moderate
children over 5: 4 minus 1 minus 1 = 2, low
```

Check the start by trying the other one. If the rating had started at low, two steps down would give this.

```
2 minus 1 minus 1 = 0
```

There is no level 0. The rating would stop at very low, not low. So both rows started at high, the start GRADE gives to randomised trials.

step	reason considered	children under 5	children over 5
0	start	4	4
1	risk of bias	3	3
2	imprecision	3	2

Now look at why the second row lost a step for imprecision. Its risk ratio is 0.81, from 624 children in 4 studies. Its 95 per cent confidence interval (CI) runs from 0.53 to 1.21. That is from about half the risk to about a fifth more risk, and 1, no difference for a ratio, sits inside it. The first row's risk ratio is 0.41, from 1474 children in 7 studies, with an interval from 0.29 to 0.55. The whole of that range lies below 1.

So the presenter had the design right and the level wrong, for both rows. Randomised trials only start at high. And one table gave the same outcome two levels, in two groups of children. Certainty belongs to each row, not to the review.

Figures quoted in this illustration are from [148].

WHERE THIS PICTURE BREAKS. Counting back to the start works only when you know how many levels each reason cost. A footnote that says "very serious" costs two, and a footnote can be silent on it. When the footnotes do not say, you cannot work back.

The figure is an example the Handbook prints to show the format. Another review team, rating the same studies, could land a level apart. The counts 4 to 1 are this book's device for counting steps. GRADE adds no scores.

Illustration 2. Now a rating that goes up. It comes from the GRADE Handbook, and the Cochrane chapter repeats it.

A systematic review of observational studies covered 38 million patients. Death rates were higher in private for-profit hospitals than in private not-for-profit hospitals. The studies are observational, so the rating starts at low: 2.

Ask which way the plausible biases push. First, the patients in the not-for-profit hospitals were probably sicker. Sicker patients die more, so that bias makes the not-for-profit hospitals look worse. Second, well-insured patients bring money that can help a hospital's other patients too. For-profit hospitals probably admit more of them, so that bias favours the for-profit hospitals.

Both biases push against the result that was found. Remove them, and the gap in death rates would most likely grow, not shrink. That is GRADE's third reason to rate up.

```
2 plus 1 = 3
```

The Handbook says one might rate this evidence moderate rather than low.

Two rules come with rating up. Weigh every reason to rate down first: rating up should be rare when any of them is serious. And use it for observational evidence. GRADE's authors write that they have not yet found a convincing case for rating up randomised trials.

Take the same question to any observational result you read. Which way would each bias push? If every plausible one pushes against the finding, say so, and say that it raises your rating.

Figures quoted in this illustration are from [149].

WHERE THIS PICTURE BREAKS. "All plausible biases push against the result" is a judgement about direction, made by the raters. It is not a measurement. If one plausible bias pushes the other way, this reason does not apply.

The Handbook's words are "one might consider", not "must". A rater who doubted that the patients were sicker would leave the rating at low.

MUST KNOW POINTS FOR YOU.

- "These are randomised trials, so the evidence is high certainty" is wrong. Trials only start at high. Read the rating, never infer it from the design.
- When you read a 'Summary of findings' table, read the certainty column and its footnotes before the size of the effect. Count back from the start: the footnotes should account for every step.
- A large effect raises observational evidence by one level when the risk ratio is above 2 or below 0.5. Check the confidence interval as well as the estimate. If the interval reaches appreciably below 2, Cochrane says some would hesitate to rate up.
- "Non-randomised studies, starting at high" in a newer review is not a promotion. Such studies may start at high, but they are generally rated down two levels for having no randomisation. They usually land at low, as before.
- When you rate evidence yourself, write the reason for every step, and expect to defend each one.
- Certainty says how sure we can be of an estimate. It does not say whether to act.
- Teach "certainty" and "quality of evidence" as two names for one GRADE rating.

COMMON MISREADING. "Low certainty means the treatment does not work." Low certainty is about how sure anyone can be of the estimate. GRADE's meaning for low is that the true effect may be substantially different from the estimate. It may be larger or smaller. A low rating on a promising estimate says that the estimate cannot yet be trusted. It says nothing about which way it will move.

REPORTING IT. For diarrhoea in children over 5, the certainty of the evidence is low (GRADE). The randomised trials start at high and are rated down one level for risk of bias and one for imprecision.

EXERCISE 1 (retrieval). From memory: name GRADE's four levels in order and give each one's meaning in a few words. Say where a body of randomised trials starts and where a body of observational studies starts. Name the five reasons to rate down and the three to rate up, and say how many levels each can move the rating.

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (interpretation). The abstract of a made-up review reads:

> *Twelve randomised trials (3,400 adults) found that a meal-replacement programme lowered body weight at 8 weeks. Certainty of the evidence: low. Downgraded for risk of bias and for indirectness. All trials were in adults in high-income countries. None followed people beyond 8 weeks.*

Count the steps from the start to the rating. Then write one sentence for a district health officer that states the certainty in words and says why. Last, say what would have to change for the rating to rise.

EXERCISE 3 (teaching). For a first-year resident, ten minutes with a whiteboard. She has seen "⊕⊕◯◯ low" in a review and asks what it means and how anyone arrives at it. Teach her.

PRACTICE. Ten problems on this technique, easiest first. Work them on paper before turning to the answers, which are in the appendix at the back under this section's name.

1. A body of evidence starts at high. It is rated down for one serious reason and one very serious reason, and for nothing else. Give the final level.

2. A body of evidence starts at low. Give the final level in each case.

1. It is rated up one level for a dose-response gradient, and nothing else moves it.
2. It is rated down one level for imprecision, and nothing else moves it.
3. It is rated down two levels for very serious risk of bias and one level for inconsistency.

3. Work backwards.

1. A body of randomised trials is rated low, and nothing rated it up. How many levels was it rated down in all?
2. A body of observational studies is rated high, and nothing rated it down. How many levels was it rated up in all?
3. A body of randomised trials is rated very low. What is the smallest number of reasons that could have taken it there?

4. Metformin is a drug for diabetes. A related drug, phenformin, causes lactic acidosis, a build-up of lactic acid in the blood. So metformin was suspected of the same harm. Very large observational studies found no link between metformin and lactic acidosis. Because of the suspicion, clinicians were likely to be on the alert and to over-report the harm in patients on metformin.

Give the level this evidence starts at. Name the reason that could move it, and say which way. Then give the level it ends at if nothing else moves it.

5. A meta-analysis of observational studies found that bicycle helmets lower the risk of head injury in cyclists. The odds ratio was 0.31, with a 95 per cent confidence interval of 0.26 to 0.37. An odds ratio is the odds of an outcome in one group divided by the odds in another. Odds are the chance an outcome happens divided by the chance it does not.

Decide whether the effect counts as large, very large or neither on GRADE's thresholds, checking the interval as well as the estimate. Then give the final level if nothing else moves the rating.

6. In Figure 14.1.b of the Cochrane Handbook, chapter 14, the outcome "duration of diarrhoea" is rated low. Its footnotes report serious unexplained inconsistency and serious imprecision.

1. Work back to the starting level.
2. A second rater judges the imprecision very serious. Where does the row end?
3. Suppose the studies had been observational, with the same two serious reasons. Where would the row end?

7. A colleague rates a body of observational studies. Find the step that broke.

```
start: 4 (high)
the studies disagree and nobody can explain why: serious inconsistency, down 1
4 minus 1 = 3
final: moderate
```

8. A systematic review of observational studies looked at the drug rofecoxib and cardiovascular events (harms to the heart and blood vessels). A mg is a milligram, a thousandth of a gram. At doses under 25 mg a day, the risk ratio was 1.33 (95 per cent CI 1.00 to 1.79). At doses above 25 mg a day, it was 2.19 (95 per cent CI 1.64 to 2.91). Here is a worked rating. Find the step that broke.

```
start: observational studies, low = 2
higher dose: risk ratio 2.19, above 2, a large effect, up 1
2 plus 1 = 3
risk rises with the dose, a dose-response gradient, up 1
3 plus 1 = 4
final: high
```

9. A newspaper health column says: "A meta-analysis of 14 studies has shown that flavonoid tablets shrink haemorrhoids (piles). A meta-analysis is the highest level of evidence, so the question is settled." The Cochrane Handbook describes the same 14 studies. They show what look like large benefits. They took in only 1432 patients in all. The companies paying for them were closely involved in most of them. Neither the column nor the Handbook says whether the studies were trials. Decide what to compute, compute it, and say what the claim does not establish.

10. On a ward round a senior colleague says: "I have put hundreds of patients with obesity on a high-protein breakfast plan. Nearly all of them lose weight. The evidence is clear." Decide what to rate, rate it, and say what your rating does not establish.

8 · Certainty of the evidence is not the strength of a recommendation, and certainty has words

Checked against the sources as they stood on 2026-10-10.

DEFINITION. In the GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach of the previous section, the certainty of the evidence answers one question about one outcome. How sure can we be that the estimate of the effect is close to the true effect?

A recommendation answers a different question: what should be done? GRADE offers two grades, strong and weak; some panels write "conditional" or "discretionary" for weak.

Certainty is one input to strength, and only one. Guyatt and colleagues (2008) list three more: the balance between desirable and undesirable effects, the values and preferences of the people affected, and whether the action is a wise use of resources. High certainty does not by itself make a recommendation strong, and a strong recommendation can rest on low or even very low certainty.

Each level of certainty also has its words. At high certainty the statement carries no hedge: the treatment reduces, or increases, the outcome. At moderate certainty it adds "probably" or "likely". At low certainty it uses "may", or opens with "The evidence suggests". At very low certainty it reads "The evidence is very uncertain about the effect of X on outcome", with the treatment in place of X.

The size has its own words: "large"; none for a moderate effect; "slightly" for a small effect that matters; and "little to no difference" for one too small to matter. [150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166]

IN PLAIN TERMS. Two different questions get asked about any treatment or programme. How sure are we about what it does? And should we do it?

A strong recommendation says: do this, nearly everyone would choose it. A weak one says: most people would choose it, but many would not, so decide with the person.

The two answers can point different ways.

Keep the word when you repeat a finding, and your listener hears how sure the evidence is.

Illustration 1. A slide in a postgraduate teaching session reads: "Strong recommendation means high-quality evidence. Weak recommendation means low-quality evidence." It looks tidy. Test it against four cases that GRADE's own authors use to teach.

Draw a grid of two rows and two columns. Across the top, write the certainty of the evidence: high, or low. Down the side, write the strength of the recommendation: strong, or weak. Then place each case in its cell.

Case 1. A short course of steroid tablets in a flare of asthma. Guyatt and colleagues use it as the plain case: large, rigorous trials show consistent benefits, with few side effects and little inconvenience or cost. Clinicians can offer it to almost all their patients "with little or no hesitation". That is evidence GRADE would rate high: high certainty, strong recommendation. Here the slide works.

Case 2. Warfarin, a medicine that slows blood clotting, after a first clot in a deep vein with no obvious cause. Trials of high quality show that taking it long term lowers the chance of another clot. It also raises the chance of bleeding, and it is a nuisance to take. Patients weigh these differently. So GRADE's Handbook says the panel should give a weak recommendation, despite the high-quality evidence. High certainty, weak recommendation.

Case 3. Antibiotics given fast in severe infection, or sepsis. No randomised trial has compared giving them fast with giving them slowly. The evidence comes from observational studies and is rated low. Yet the Handbook says a panel would very likely recommend fast antibiotics strongly, because the benefits clearly outweigh the downsides in most patients. Low certainty, strong recommendation.

Case 4. Fever in a child with chickenpox: aspirin or paracetamol? Observational studies link aspirin to an illness called Reye's syndrome, and that evidence is of low quality. But the two medicines relieve fever and pain about equally well. Guyatt and colleagues say the low-quality evidence does not stop a strong recommendation for paracetamol. Low certainty, strong recommendation.

Now fill in the grid.

	high certainty	low certainty
strong recommendation	case 1, steroids in asthma	case 3, fast antibiotics; case 4, paracetamol
weak recommendation	case 2, long-term warfarin	

Two of the three filled cells break the slide. Ask what moved cases 2, 3 and 4. It was not the evidence. Case 2 moved on harm, and on how differently patients value bleeding against a second clot. Cases 3 and 4 moved on a clear balance: much to gain, little or nothing to lose.

Guyatt and colleagues list four things that set the strength of a recommendation. The certainty of the evidence is one. The others are the balance of benefits against harms, the values and preferences of the people affected, and whether the action is a wise use of resources. So rewrite the slide: "Strength depends on certainty, on the balance of benefit and harm, on what people value, and on cost. Certainty is one input of four."

WHERE THIS PICTURE BREAKS. These are teaching cases, chosen by GRADE's authors in 2008 and 2013 to make the point. They may not be current clinical advice, so check a current guideline before you act on any of them.

The grid has only two levels of certainty. Real evidence also sits at moderate and very low, and a panel's reasons for a strength are a judgement it should publish. Two panels can grade the same evidence differently.

Illustration 2. Now use the words. Sievert and colleagues (2019) pooled the randomised trials that compared eating breakfast with skipping it in adults. Their review is free to read at https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/. Search its Results for "favouring participants who skipped breakfast". The difference in weight favoured those who skipped breakfast, by a mean of 0.44 kg. The trials followed people for seven weeks on average. Then find the heading "Quality of evidence". The authors judge its quality low, for risk of bias and short follow-up. They did not use GRADE, so treat "low" as their own judgement, made on GRADE-like grounds. The word for it is still "may".

A newspaper health column could turn this into: "Breakfast makes you gain weight, a major review proves." Hold that sentence against the review. It has dropped the hedge, so it reads as high certainty when the review's authors judged the evidence low. And "proves" is a word no level of certainty earns.

Write the sentence yourself, in three steps.

1. Take the certainty word. The evidence is low, so the word is "may".
2. Decide the size. Is a 0.44 kg difference over a few weeks a small effect that matters, or too small to matter? That is your judgement, and Cochrane's statements differ for the two. If you judge it small but important, write "eating breakfast may increase weight slightly". If you judge it too small to matter, write "eating breakfast may make little to no difference to weight". Either way, "may" carries the low certainty.
3. Say who and how long: adults in trials in high-income countries, over weeks, not years.

Here is one version: "In short trials of adults in high-income countries, eating breakfast may increase weight slightly compared with skipping it."

Notice what the sentence does not say. A district health officer who reads it may ask whether the district's nutrition talks should stop telling people to eat breakfast to control weight. That is a recommendation question. The review answers how sure we can be. The officer still needs the other three inputs: what else the advice does, what people value, and what it costs to change the message.

Figures quoted in this illustration are from [167].

WHERE THIS PICTURE BREAKS. The words carry the level only to a reader who knows the convention. They are Cochrane's suggested statements for a review's conclusions. Other bodies use other words, and a reader who has never met the convention may read "may" as "might or might not", with no sense of how sure. So in writing for readers outside the field, name the level as well: "low-certainty evidence suggests".

The breakfast trials were run in high-income countries. None was in India, so the sentence is not a statement about Indian adults.

MUST KNOW POINTS FOR YOU.

- "A strong recommendation means strong evidence" is the error.
- Read every recommendation for two labels: its strength, and the certainty of the evidence behind it. If a guideline prints only one, ask for the other.

- A weak, or conditional, recommendation does not mean weak evidence, and it does not mean "optional". In clinic it means you make the choice with the patient, and it takes more time.
- Drop the "may" from a low-certainty finding, on a slide or to a reporter, and you have raised it to high without saying so.
- A systematic review tells you how sure; it does not tell you what to do. You still need the harms, the costs and what the people affected value.
- When you adapt a guideline written for another country, do not copy its strength. The certainty may carry over if the trials fit your people. The strength also rests on cost, on what can be done, and on what people value. These can be different in an Indian district.
- In anything written for readers outside the field, name the level beside the word: "low-certainty evidence suggests that …".

COMMON MISREADING. It is tempting to read the certainty word as a size: "may" for a small effect, no hedge for a big one. That reading is wrong. In Cochrane's suggested statements the certainty word and the size word are separate parts of the sentence. A large effect known with moderate certainty takes "probably" and "large". A slight effect known with high certainty takes no hedge and "slightly". So read the sentence twice: once for how sure, once for how big.

REPORTING IT. Randomised trials have compared eating breakfast with skipping it in adults in high-income countries, over seven weeks on average. The difference in weight favoured skipping breakfast, by a mean of 0.44 kg (Sievert and colleagues, 2019). The authors judged the quality of this evidence low, so eating breakfast may increase weight slightly compared with skipping it.

EXERCISE 1 (retrieval). From memory: what word or phrase does each of the four levels of certainty put into a conclusion, in Cochrane's suggested statements? And what four things set the strength of a recommendation, in Guyatt and colleagues' list?

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (interpretation). Each sentence below is the conclusion of a review, written to Cochrane's convention. The treatments are left unnamed. Say what level of certainty each sentence signals, and what it says about the size of the effect.

(a) Treatment A probably reduces hospital admissions.

(b) Treatment B reduces pain slightly.

(c) The evidence is very uncertain about the effect of treatment C on falls.

(d) Treatment D may result in little to no difference in weight.

EXERCISE 3 (critique). A resident ends a journal club on a systematic review with this slide. The review rated the evidence for its main outcome as low certainty.

> *The review shows that the programme reduces readmissions. Every district should adopt it: this should be a strong recommendation.*

Find the faults, and rewrite the first sentence.

EXERCISE 4 (teaching). For a health secretary, one page. The secretary has read that a guideline makes a strong recommendation based on low-certainty evidence, and asks whether that is a mistake. Answer first.

9 · No evidence of an effect is not evidence of no effect

DEFINITION. A comparison that finds no clear difference between two groups has shown an absence of evidence of a difference. It has not shown evidence that there is no difference. Altman and Bland (1995) call these "quite different statements".

A study can miss a real difference because it is too small.

"Not significant", as earlier sections used it, is a statistical test's verdict that the difference is not too large to put down to chance. The verdict does not say how large a difference the data still allow.

The confidence interval does. It is the range of values the data still allow, at a stated level (section 6). For a 95 per cent confidence interval, if the study were repeated many times, 95 per cent of such ranges would contain the true value.

Read a confidence interval for a difference against two marks. One is zero, no difference. (For a ratio, such as a risk ratio, the no-difference mark is 1.) The other is the smallest difference that matters: the smallest difference that would change what someone does, decided before the result is seen.

	interval includes differences that matter	interval holds only differences too small to matter
interval excludes zero	evidence of a difference that may matter, or does	evidence of a real but unimportant difference
interval includes zero	absence of evidence: the study cannot tell	evidence of little or no difference

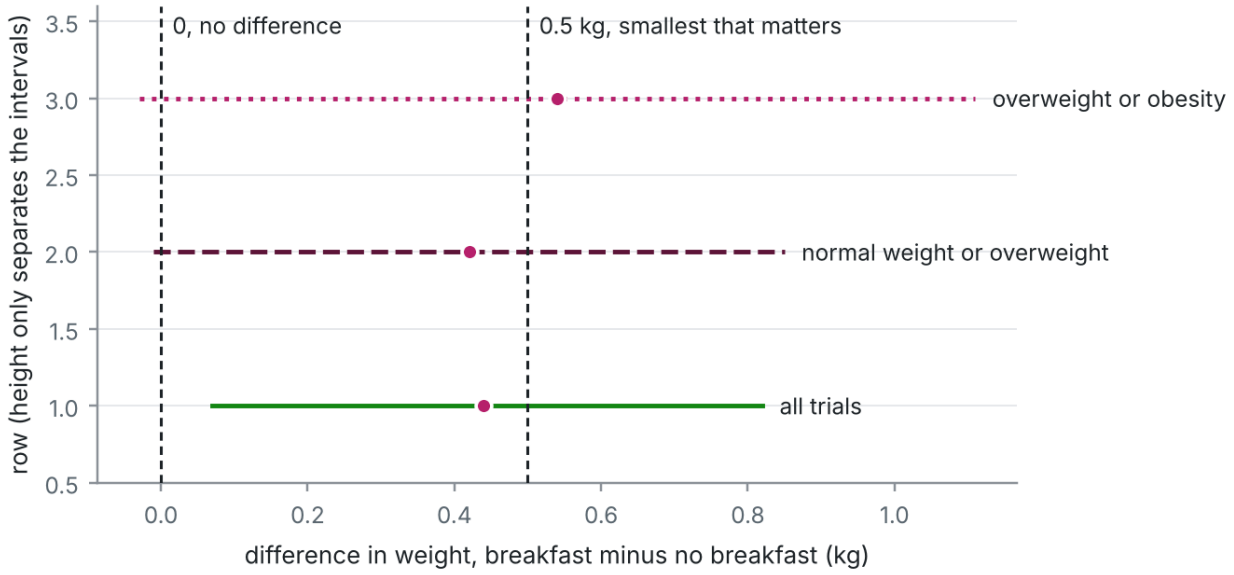
Only the bottom-right case supports saying there is little or no difference. [168, 168, 169, 170, 171, 172, 53, 173, 174, 175, 176, 177, 178, 179]

IN PLAIN TERMS. A study that finds no difference has not shown that there is no difference. It may simply have been too small to see one.

Picture two groups whose true averages differ a little. Measure ten people in each, and chance can swamp that small gap. Measure thousands, and the gap shows.

To learn more, look at the confidence interval: the range of differences the data still allow. If that range runs from a real benefit to a real harm, the study cannot tell you which.

The reverse holds too. "There is no evidence that it harms" can mean nobody has looked.



Nearly the same estimate, three different verdicts by zero alone. Each dot is a mean difference in weight, breakfast minus no breakfast (kg), and each line its 95 per cent confidence interval, one row each; height only separates the rows. All trials: 0.44, 0.07 to 0.82, 0.75 wide, clear of zero. Normal weight or overweight: 0.42, −0.01 to 0.85, 0.86 wide. Overweight or obesity: 0.54, −0.03 to 1.11, 1.14 wide. The wider intervals reach across zero; the estimates barely move. All three hold 0.5 kg, the smallest difference that matters chosen in this section's example. Sievert and colleagues (2019).

Illustration 1. Start with the breakfast trial, read in its authors' own words. Dhurandhar and colleagues (2014) randomly assigned 309 adults who were trying to lose weight to advice to eat breakfast, advice to skip it, or general advice alone, for 16 weeks. The paper is at https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/.

Read the Conclusions of the abstract. The advice "had no discernable effect on weight loss". Now search the Discussion for "only suggest". The same authors write that their findings "only suggest that in a free-living less controlled setting, breakfast eating habits do not influence weight loss". The same sentence keeps open an effect in a more controlled clinical setting.

Put the two sentences side by side. The first says the trial could not see an effect. The second, softened by "only suggest", still says there is no effect. That second claim needs more evidence than the first. It needs a confidence interval for the difference, narrow enough to rule out any difference that would matter.

So look for one. Search the Results for "P = 0.77": the advice had no effect on weight loss that the test could detect. Now search the whole paper for "confidence interval". The paper gives none for the difference in weight between the groups.

What it does give is each group's mean change in weight, which you met in the section on two kinds of support. Among people who usually skipped breakfast, those told to eat it lost 0.15 kg more than those told to skip it. Among people who usually ate breakfast, the gap was 0.12 kg the other way. Those are the differences that were observed. They are not a confidence interval, so they do not tell you how large a true difference the data still allow. The authors themselves note that 16 weeks "may not have been long enough to detect an effect on body weight" if the change in energy intake is very small.

So here is what you may say: “In a 16-week trial of 309 adults trying to lose weight, advice to eat or skip breakfast made no detectable difference to weight. The paper reports no confidence interval for the difference, so it does not show that breakfast has no effect.”

Figures quoted in this illustration are from [33].

WHERE THIS PICTURE BREAKS. The P value of 0.77 is from the 283 people who finished. The authors also analysed all 309, estimating the missing weights from the data that were recorded (multiple imputation), and found comparable results.

The trial tested advice, not breakfast itself, in adults trying to lose weight, at several sites, one of them in Copenhagen. It says nothing about children, or about Indian adults.

Illustration 2. Now read confidence intervals for yourself. Sievert and colleagues (2019) pooled the randomised trials of eating against skipping breakfast in adults (https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/). The breakfast trial above is one of them. Their pooled answer and two smaller answers are below. Each is the mean difference in weight, the mean of the breakfast groups minus the mean of the skipping groups, in kg. A positive number means the breakfast eaters ended heavier.

trials pooled	mean difference, kg	95% confidence interval, kg
all trials of weight	0.44	0.07 to 0.82
trials in people with normal weight or overweight	0.42	−0.01 to 0.85
trials in people with overweight or obesity	0.54	−0.03 to 1.11

The two smaller answers group trials by the people they enrolled, so people with overweight appear in both.

The sign − (minus) marks a difference in favour of eating breakfast. So −0.01 means the breakfast eaters ended 0.01 kg lighter.

First, decide your smallest difference that matters, before you read any interval. Suppose you decide that 0.5 kg over a few weeks is the smallest gap that would change your advice. That is your judgement, not the review’s, and another reader may set it elsewhere.

Now sort each row into the four cases.

All trials: 0.07 to 0.82. Zero is outside, so there is evidence of a difference, favouring skipping. The interval holds values below 0.5 and above it, so the difference may or may not matter.

Normal weight or overweight: −0.01 to 0.85. Zero is inside, and so is 0.5 and more. This set of trials cannot tell: absence of evidence.

Overweight or obesity: −0.03 to 1.11. The same case.

Here is the trap. Sort the rows only by whether zero is inside, and the table seems to say: a difference overall, and none in either group. A careless reader then writes: “Breakfast makes no difference to weight in people of normal weight or overweight, or in people with obesity.” Look at the estimates: 0.42 and 0.54 kg, close to the pooled 0.44. The estimates barely moved. Each smaller answer pools only some of the trials, so its interval is wider.

```
all trials: 0.82 minus 0.07 = 0.75 kg wide
normal weight or overweight: 0.85 minus -0.01 = 0.86 kg wide
overweight or obesity: 1.11 minus -0.03 = 1.14 kg wide
```

Wider intervals reach across zero more easily. Sievert and colleagues report that the effect did not differ between the two groups of trials. The article text does not print their test.

Notice too that the breakfast trial, which on its own detected no effect, is part of the pooled answer that did find a difference. One trial's absence of evidence did not settle the question.

Figures quoted in this illustration are from [167].

WHERE THIS PICTURE BREAKS. The 0.5 kg is a choice made for this exercise, not a published threshold. Move it and the sorting of the first row changes.

These intervals measure only the scatter of chance. They say nothing about bias in the trials, which the review's authors judged mostly of low quality, and they hold for adults in short trials in high-income countries.

MUST KNOW POINTS FOR YOU.

- "Not statistically significant, so no effect" is the error. Say "no clear difference", and give the estimate with its confidence interval.
- Before you claim that something has little or no effect, fix the smallest difference that matters, and then read the confidence interval.
- Small trials that each miss a real effect can long delay its recognition. Altman and Bland describe trials of a clot-dissolving treatment after a heart attack. Pooled together, they showed that deaths fell by 22 per cent. Yet only five of the 24 trials had shown a significant effect on their own.
- Significant in one group and not significant in another does not mean the effect differs between the groups. Compare the estimates, not the verdicts.
- Watch for one-sided reading: a wide interval called "promising" when the estimate favours the treatment, and "no effect" when it points to harm. Test your own conclusion by reversing the direction of the result and asking whether you would still write it the same way.
- "There is no evidence that it harms" may mean nobody has looked.
- A confidence interval measures only the scatter of chance. So "evidence of little or no effect" also needs a study whose design you trust.
- To a reporter, say "this trial could not detect a difference", not "breakfast does not matter".

COMMON MISREADING. It is tempting to read a large P value as strong evidence of no effect. P = 0.77 is far above 0.05, so surely the answer is "no effect"? That reading is wrong. A P value, large or small, does not tell you how big a difference the data still allow. Only the estimate and its confidence interval do. Altman and Bland's advice is to look for the size of the association, not just a

P value. The breakfast trial printed P = 0.77 and no interval, so it cannot say how large an effect it rules out.

REPORTING IT. A 16-week trial randomly assigned 309 adults trying to lose weight to advice about breakfast (Dhurandhar and colleagues, 2014). Advice to eat or to skip breakfast made no detectable difference to weight. The trial reports no confidence interval for the difference, so it does not show that breakfast has no effect on weight.

EXERCISE 1 (retrieval). From memory: what is the difference between "no evidence of an effect" and "evidence of no effect"? Then name the two marks you read a confidence interval against, and the four cases they give.

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (interpretation). Five comparisons, with made-up numbers. Each gives a 95 per cent confidence interval for a difference between two groups. You have decided beforehand that the smallest difference that matters is 2, in either direction. Sort each into one of the four cases.

(a) 0.5 to 4.5

(b) −3.0 to 5.0

(c) −0.8 to 1.2

(d) 0.3 to 1.1

(e) 2.5 to 6.0

EXERCISE 3 (critique). A journal club slide reads:

> *Breakfast trial (309 adults, 16 weeks, P = 0.77): breakfast has no effect on weight. So we should stop telling patients that breakfast matters for weight.*

Rewrite the first sentence so that it says what the trial showed, and say what is wrong with the second.

EXERCISE 4 (teaching). For a journalist, three minutes. A reporter calls: "That big breakfast trial found nothing. So breakfast doesn't matter for weight, right?" Answer with one sentence she can quote, and say how sure you are without becoming useless to her.

10 · Confidence as a probability: putting a number on how sure you are of one claim

DEFINITION. In this book, a probability can also be a degree of belief: a number from 0 to 1 that states how strongly one person holds one claim to be true, given what that person knows.

A degree of belief of 0 says the person is sure the claim is false. A degree of belief of 1 says they are sure it is true. Between them, it obeys the same arithmetic as Book 0's probability. The degree of belief that the claim is false is its complement: 1 minus the degree of belief that it is true.

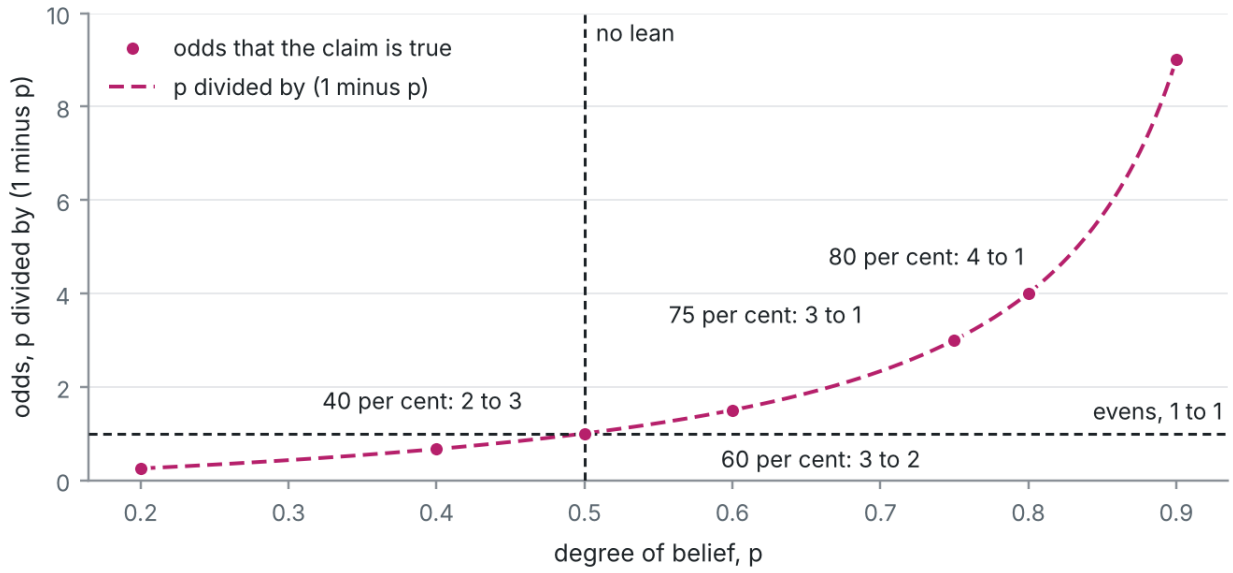
The odds that a claim is true are the degree of belief, p, divided by its complement: p divided by (1 minus p), as in Book 0. This book writes odds as two whole numbers joined by "to", with the claim's side first. A degree of belief of 0.8 gives 0.8 divided by 0.2, which is 4, written "4 to 1". Odds "against" put the other side first: 4 to 1 for is 1 to 4 against. Odds of a to b, with a and b standing for any two counts, turn back into a degree of belief as a divided by (a plus b).

A degree of belief of 0.5 (odds of 1 to 1) states no lean either way. It is not a statement of moderate confidence. [180, 181, 182, 183, 184, 185]

IN PLAIN TERMS. Here you use the same kind of number for something else: how sure you are that one claim is true.

Say it as a per cent. "I am 80 per cent sure." Then you are 20 per cent sure it is false.

Fifty per cent means you lean neither way. It does not mean "fairly sure".



The same degree of belief as a probability and as odds: odds are p divided by (1 minus p). At 50 per cent the odds are 1 to 1, evens, the one point with no lean. At 40 per cent they are about 0.667 (2 to 3); at 60 per cent, 1.5 (3 to 2). At 20 per cent, 0.25, or 1 to 4. At 75 per cent, 3 to 1; at 80 per cent, 4 to 1; at 90 per cent, 9 to 1. The odds climb faster and faster as p nears 1: from 0.8 to 0.9 they more than double.

Illustration 1. This section stands on the first one in this book, where a belief is what a person holds true, held with some strength. Here you put a number on that strength.

Start at a journal club. A colleague says: “Eating breakfast helps people lose weight. I am fairly sure of it.” Ask her: “How sure? Give me a number.” She thinks, and says “80 per cent.”

Now try to read her number the way Book 0 reads a probability (Book 0, D1). There, 0.8 would mean that a chance process, repeated many times, gives this outcome 8 times in 10. What is the process here? The claim is either true or false. Nobody can rerun the world a thousand times and count the worlds in which breakfast helps. So her 80 per cent is not a frequency. It is a fact about her: how strongly she holds the claim, given what she has read.

That second use of the word is old. In 1847 De Morgan wrote, as the *Stanford Encyclopedia of Philosophy* reports: “By degree of probability, we really mean, or ought to mean, degree of belief”. Its entry on probability calls this the subjective reading, in which probabilities are “degrees of confidence, or credences, or partial beliefs”. This book calls it a degree of belief, or your confidence in a claim.

The arithmetic does not change. Write her number three ways, as Book 0 did for percentages (Book 0, A4).

```
80 divided by 100 = 0.8
```

And 0.8 is 8 in 10, which is 4 in 5.

Next, the complement. If she is 0.8 sure the claim is true, she is 1 minus 0.8 sure it is false.

```
1 minus 0.8 = 0.2
```

She is 20 per cent sure it is false. Say both halves out loud. “80 per cent sure it is true” sounds like a firm view. “One chance in five that I am wrong” is the same number, and it is the half that a listener forgets.

Now the odds. Book 0 defined them as p divided by (1 minus p), where p stands for the probability. Here p is her degree of belief, 0.8.

```
0.8 divided by 0.2 = 4
```

The odds are 4. This book writes that as “4 to 1”: for every 4 chances that the claim is true, 1 that it is false. The claim’s side always comes first.

Go back the other way. Odds of a to b, where a and b stand for any two counts, give a degree of belief of a divided by (a plus b). Here a is 4 and b is 1, so 4 + 1 = 5 chances in all.

```
4 divided by 5 = 0.8
```

You are back at 80 per cent. Run that check every time you convert. It catches the commonest slip, which is to swap the two sides.

WHERE THIS PICTURE BREAKS. The number belongs to the person, not to the claim. A second colleague who has read different papers may say 40 per cent about the same sentence, and neither number is a mistake in arithmetic. The number also says nothing about how large any effect of breakfast is. It is confidence that the claim is true, not a size of weight change.

Her 80 per cent cannot be shown right or wrong by this one claim. If breakfast turns out not to help, she had said there was a 1 in 5 chance of that. A degree of belief is checked only across many claims held at the same level, which a later section on calibration does.

Illustration 2. Here is a real case of one judgement written three ways, on purpose. Sherman Kent sat on the US Board of National Estimates, which oversaw the government’s intelligence forecasts. In a 1964 article he drafted a note addressed to the readers of those forecasts. Open his article *Words of Estimative Probability* at https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf and search for “Important note to consumers”.

> *when we say .such and such is unlikely” we mean that the chances of its NOT happening are in our judgment about three to one. Another, and to you critically important, way of saying the same thing is that the chances of its HAPPENING are about one in four.*

(The stray full stop before “such” is in the printed copy.) In the example that follows, he writes the same judgement as “around a 25-percent chance”.

Check that his three forms agree. “Three to one” that it does NOT happen puts the not-happening side first. Turned round, with the happening side first, that is 1 to 3.

```
1 divided by (1 plus 3) = 0.25
0.25 times 100 = 25
```

Odds of 1 to 3 that it happens give 0.25, which is 1 in 4, which is 25 per cent. Kent's three forms are one number.

Why did he write it three ways? Because a reader who hears "three to one" without its side can take it either way round. One reading gives a 25 per cent chance of the event; the other gives 75 per cent. When you quote odds, say whose side comes first, or add "against".

Now a trap that even careful writers fall into. A 2024 review by Hashim of how clinicians and patients read probability words lists formats it advises against, among them "odds (one in three)". By the definitions you are using, "one in three" is a chance, not odds.

```
1 divided by 3 is about 0.333
0.333 divided by (1 minus 0.333) = 0.5
```

A chance of 1 in 3 is odds of 1 to 2: one chance for, two against. "In" counts the whole; "to" sets one side against the other. Read which word the writer used before you convert.

Figures quoted in this illustration are from [15, 186].

WHERE THIS PICTURE BREAKS. Kent's note fixes what his own team meant by "unlikely". It does not tell you what any other writer means by the word. How different readers turn such words into numbers is the subject of the next section.

The 2024 review's point is about which format patients understand, and it recommends "X in 100" for that reason. Its use of "odds" is loose; its advice may still be sound.

Illustration 3. Now the number that is easiest to misread. A district health officer asks you whether a planned school nutrition talk will lower the share of children with overweight in her district. You reply, "Fifty-fifty." The case is made up.

Convert it first.

```
0.5 divided by (1 minus 0.5) = 1
```

Odds of 1 to 1, or "evens". Fifty per cent is the one point at which you lean neither way. Move a little either side and you have a lean.

```
0.6 divided by 0.4 = 1.5
0.4 divided by 0.6 is about 0.667
```

Sixty per cent is odds of 1.5 to 1, which is 3 to 2 in whole numbers. Forty per cent is 2 to 3.

So "fifty-fifty" is not "moderately sure". A person who is fairly sure leans clearly one way, so any number they have in mind is above 50. Fifty per cent is no lean at all, not a middling lean.

Now ask what you meant. There are two very different reasons to say 50. One is that you know a great deal, and the evidence points both ways about equally. The other is that you know nothing at all. The number is the same, and the officer cannot tell which you meant.

Bodies that write probabilities for a living keep the two apart. The guidance note for authors of the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) says its phrase for a middle probability, "about as likely as not", "should not be used to express a lack of

knowledge". If you mean "I don't know", say that, and say why: "I have seen no study of a talk like this one." A later section shows how to state your grounds beside your number.

WHERE THIS PICTURE BREAKS. "Fifty-fifty" has a real use: two outcomes that the evidence shows to be about equally likely. The point is not to avoid 50 per cent. It is to say whether it comes from evidence or from its absence.

How far different people's readings of words such as "fairly sure" spread is measured in the next section.

MUST KNOW POINTS FOR YOU.

- "Fifty per cent" is not "fairly sure". It is the one point at which you lean neither way, odds of 1 to 1.
- Give the complement with the number: "80 per cent sure it is true, so a 1 in 5 chance I am wrong."
- "1 in 4" and "1 to 4" are different numbers. The first is a 25 per cent chance; the second is a 20 per cent chance.
- Do not say 0 or 100 per cent about a claim that rests on evidence. Evidence never puts a claim beyond correction.
- One outcome cannot show a single degree of belief wrong. To learn whether your numbers mean anything, write them down and count how often claims held at each level turn out true.
- Your number describes you and what you have read, not the patient or the world. Two honest colleagues can hold 80 and 40 per cent on the same claim.

COMMON MISREADING. "A probability is how often something happens over many tries. So a probability that this one claim is true means nothing." That is Book 0's frequency reading, and for a one-off claim it has no long run to count.

The degree-of-belief reading is a second use of the same number. It is at least as old as De Morgan's remark of 1847, and it obeys the same arithmetic. Bodies such as the IPCC use it for a single event. So say which you mean. "About 40 in 100 children like these" is a frequency. "I am 40 per cent sure the scheme works" is a degree of belief.

REPORTING IT. I am about 80 per cent confident that this claim is true, which is odds of about 4 to 1. That is my judgement on the evidence I have read, not a measured rate.

EXERCISE 1 (retrieval). From memory: say what a degree of belief is, and how it differs from the probability defined in Book 0. Then write 80 per cent as a decimal, as a chance "in 5", and as odds. Last, say what a stated 50 per cent does and does not tell a listener.

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (critique). A senior colleague speaks at a journal club. "Book 0 is right that a probability describes a group, not the one person in front of you. So when you say you are 70 per cent

sure that this trial's result will hold up, you are talking nonsense." Find what is right in this, and the specific flaw.

EXERCISE 3 (teaching). A first-year resident asks you a question on a ward round. "If I am 75 per cent sure of a diagnosis, what are the odds? And why does it matter which way round I say them?" You have ten minutes and a whiteboard. Give your answer.

PRACTICE. Eleven problems on this technique, easiest first. Work them on paper before turning to the answers, which are in the appendix at the back under this section's name.

1. Write each degree of belief as a decimal. Then write it as a chance "in" a whole number, in lowest terms. The four are 90 per cent, 25 per cent, 60 per cent and 5 per cent.

2. For each degree of belief that a claim is true, give the degree of belief that it is false: 0.85, 0.3, 0.5, 0.99.

3. Give the odds that the claim is true, in whole numbers with the claim's side first, for each degree of belief: 0.75, 0.2, 0.9, 0.6.

4. Each pair of odds is for a claim being true. Turn each into a degree of belief, as a per cent: 3 to 2, 1 to 9, 7 to 3, 1 to 1.

5. In 1951 Sherman Kent was asked what his board had meant by calling an attack "a serious possibility". His answer was that "the odds were around 65 to 35 in favor of an attack." Give his degree of belief in an attack as a per cent. Give his degree of belief that there would be none. Then give his odds in the smallest whole numbers.

6. A 2024 review proposes that clinicians use "unlikely" to mean a 20 per cent chance, written as 20 in 100. Give the chance that the event does not happen, the odds that it happens, and the odds against it.

7. Kent asked the members of his board what odds they had had in mind for that "serious possibility". He wrote that "the low man was thinking of about 20 to 80, the high of 80 to 20." Give each member's degree of belief as a per cent, and his odds in the smallest whole numbers. Then say how many times larger the high member's odds were than the low member's.

8. Here is a worked answer. Find the step that broke, and say why it looks reasonable.

> *"I am 30 per cent sure the claim is false. So my odds that it is true are 3 to 7, and I am 30 per cent sure it is true."*

```
0.3 divided by 0.7 is about 0.43, so 3 to 7
```

9. Here is a worked answer to "A patient information leaflet says the chance of a side effect is 1 in 4. What are the odds?" Find the step that broke.

```
1 in 4 means odds of 1 to 4
1 divided by (1 plus 4) = 0.2
so the chance of the side effect is 20 per cent
```

10. A made-up meeting is discussing a new weight-management clinic. A consultant says: "Whether it will cut referrals to the hospital is fifty-fifty, so it is a coin toss either way." The administrator concludes: "Then the evidence must be evenly balanced." Decide what to compute, compute it, and say what the statement does not establish.

11. A made-up newspaper health column says: "Experts put the odds at 3 to 1 that skipping breakfast makes you gain weight. So skippers will put on three times as much weight." Decide what to compute, compute it, and say what the claim does not establish.

11 · The same word read as different numbers

Checked against the sources as they stood on 2026-10-10.

DEFINITION. A verbal probability is a word or phrase that states a degree of belief without a number: “likely”, “unlikely”, “possible”, “a serious possibility”. Different people turn the same word into different numbers. The writer of a sentence and its reader may do so too.

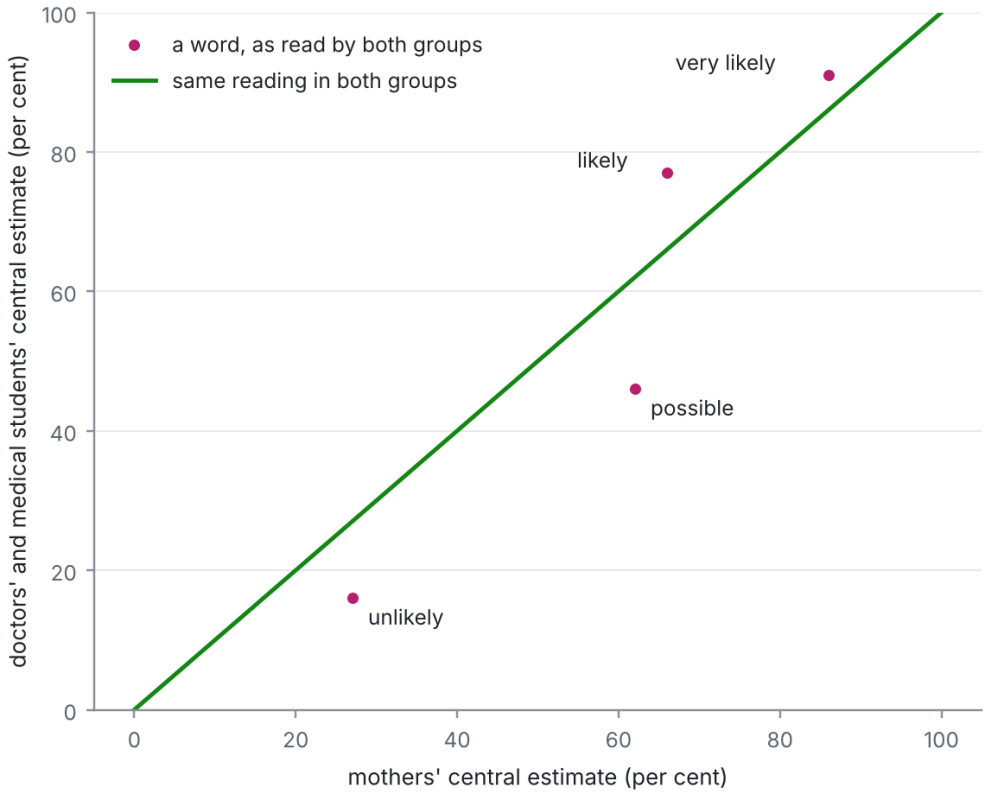
This is measured, not assumed. In an online experiment with 924 people, Wintle and colleagues (2019) asked readers for a range of numbers for each word. The spread between readers was larger for “unlikely” than for “likely”. Readings also move with context, including how severe the outcome would be.

The cost falls on the reader. A reader who acts on a word acts on the number they read into it, which may not be the number the writer meant.

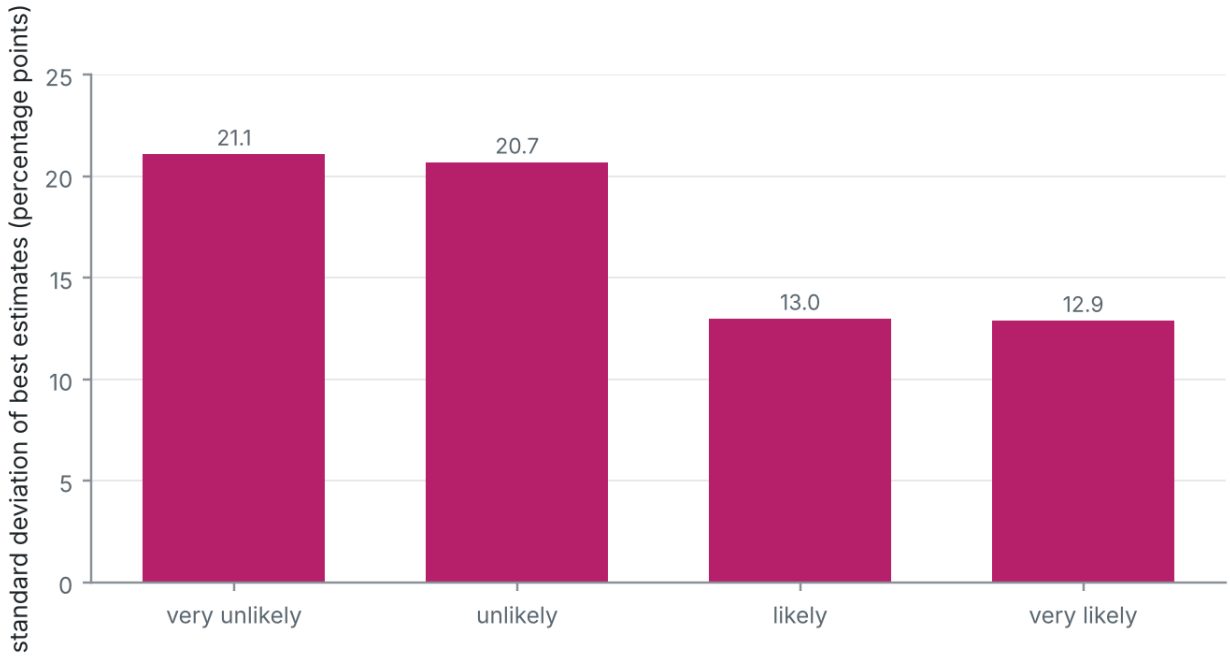
None of the studies here was done in India. [187, 188, 189, 190, 191, 192, 193, 194, 195]

IN PLAIN TERMS. You say “unlikely”. In one study, doctors read that word as about 16 in 100, and mothers as about 27 in 100. Neither side can see the gap from the word alone.

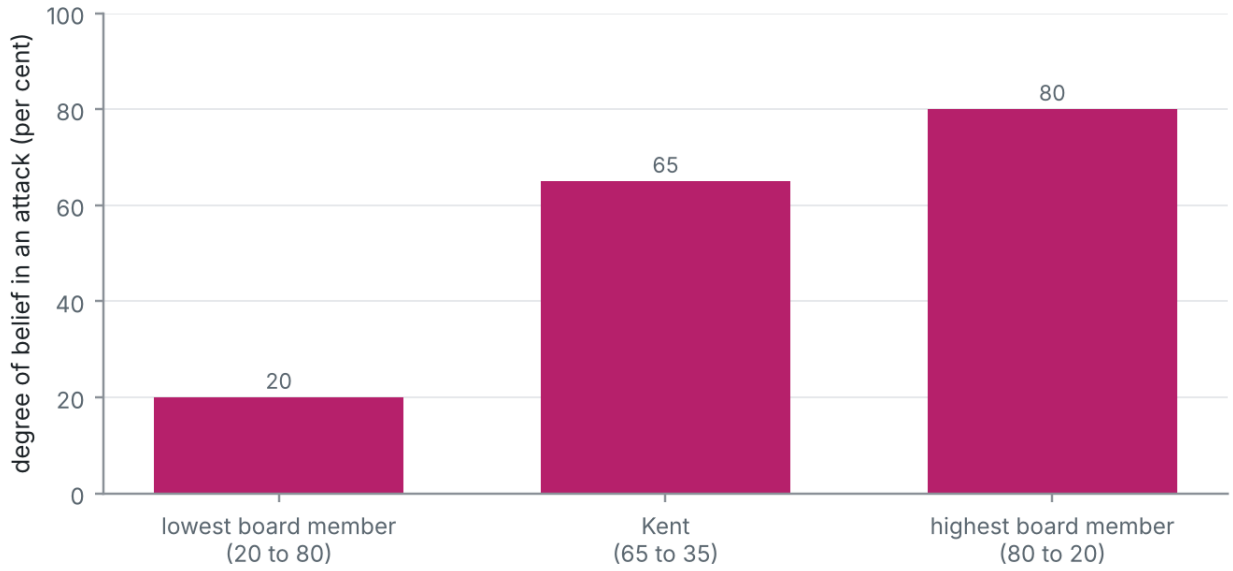
People agree on the order of the words: “likely” is more than “possible”. They do not agree on the numbers.



Four probability words, as read by 100 mothers and 50 doctors and medical students in one UK study (Shaw and Dear 1990, reported in Hashim 2024, Table 2): central estimates in per cent, the mothers' along the bottom and the doctors' up the side. On the diagonal line both groups would read a word alike. "Very likely" and "likely" sit above it: the doctors read them higher, "likely" by 11 points. "Possible" and "unlikely" sit below: the doctors read them lower, by 16 and 11 points. Both groups keep the words in the same order.



How far readers' best numbers for one word spread, in Wintle and colleagues' online experiment (2019), across all four groups, some of whom had been shown the standard's numbers: the standard deviation of the best estimates, in percentage points. The two negative words spread more, 21.1 for "very unlikely" and 20.7 for "unlikely", than "likely" (13.0) and "very likely" (12.9).



One phrase, "a serious possibility", signed by one board in 1951, held as three different degrees of belief that an attack would come: 20 per cent for the lowest member, 65 for Kent and 80 for the highest. The words recorded none of this spread. Kent (1964).

Illustration 1. The last section turned a degree of belief into a per cent and into odds. Here is what happens when nobody writes the number down.

In March 1951 the US intelligence services issued a forecast, a National Intelligence Estimate, on whether the Soviet Union and its allies would attack Yugoslavia that year. Its key sentence ended: an attack "should be considered a serious possibility." Sherman Kent, a member of the board that oversaw its drafting, tells what happened next in *Words of Estimative Probability* (1964), at https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf. Search for "serious possibility".

A few days later, the chairman of a State Department planning staff asked him:

> *"By the way, what did you people mean by the expression `serious possibility'? What kind of odds did you have in mind?"*

Kent's own answer was "the odds were around 65 to 35 in favor of an attack." The chairman "was somewhat jolted by this". He and his colleagues "had read"serious possibility" to mean odds very considerably lower."

So Kent asked his fellow board members what odds they had held when they agreed to the phrase.

> *each Board member had had somewhat different odds in mind and the low man was thinking of about 20 to 80, the high of 80 to 20.*

Turn those odds into degrees of belief, as in the last section.

```
20 divided by (20 plus 80) = 0.2
65 divided by (65 plus 35) = 0.65
80 divided by (80 plus 20) = 0.8
```

The men who signed one sentence held anything from 20 to 80 per cent. The phrase did not record a disagreement that big. It hid it. And the readers of the sentence read a number lower than Kent's own.

Notice what went wrong. Nobody misused arithmetic. Each person turned the phrase into a number in his own head, and nobody said his number out loud.

Figures quoted in this illustration are from [15].

WHERE THIS PICTURE BREAKS. This is one phrase, one board and one planning staff, told by one of the people involved. It shows that the spread can be very wide among experts who worked on the text together. It does not show how wide it usually is. That needs the counted studies in the next two illustrations.

"Serious possibility" is also an unusual phrase. Kent's own diagnosis was that "possible" should mark something whose odds cannot be estimated at all. Adding "serious" to it was, in his words, the team's "particular error".

Illustration 2. Now the clinic. A 2024 systematic review by Hashim collected studies that asked people to put a number on probability words. Open it at https://pmc.ncbi.nlm.nih.gov/articles/PMC11067312/ and find Table 2.

One of the six studies, by Shaw and Dear in the UK in 1990, asked 100 mothers and 50 doctors and medical students the same thing. Here are the central estimates the review reports for each group, for four words. Read the table as Book 0 taught (Book 0, F1): one row is one word, and each column is one group's central number for it, in per cent.

word	mothers (per cent)	doctors and medical students (per cent)
very likely	86	91
likely	66	77
possible	62	46
unlikely	27	16

Work out each gap in percentage points, as in Book 0, A4.

```
91 minus 86 = 5
77 minus 66 = 11
62 minus 46 = 16
27 minus 16 = 11
```

The doctors put "very likely" and "likely" higher than the mothers did. They put "possible" and "unlikely" lower. Both groups kept the four words in the same order.

Now picture it on a ward. The case is made up. You tell a mother that a complication after her child's minor operation is "unlikely". You have in mind something like the doctors' figure, 16 per

cent. If she reads the word as the mothers in this study did, she hears 27 per cent. For "possible", the gap runs the other way, and she may hear more than you meant.

The review also prints a lower and an upper figure for each group, without saying how they were made. For "likely" they were 49 and 83 for the mothers and 64 and 90 for the doctors, evenly either side of the central figure.

Figures quoted in this illustration are from [186].

WHERE THIS PICTURE BREAKS. Shaw and Dear's is one small study from 1990, and the review calls the quality of its six studies "generally low". It does not tell you what a mother in an Indian district hospital hears, in English or in any other language. No study in the review was done in India.

The review itself reads the evidence more hopefully than this section does. It stresses that the order of the words is stable, and proposes fixing a number to each of five words. That is a scale, and scales are the subject of the next section.

Illustration 3. One board and one small study could be flukes. So look at a large count. Wintle and colleagues reported an online experiment with 924 people in 2019. Open it at https://pmc.ncbi.nlm.nih.gov/articles/PMC6469752/.

Each person read sentences taken from intelligence reports, each holding one word such as "likely". For each sentence they gave their lowest, best and highest number for the word. The team compared each person's range with the range that a US government standard fixes for that word, Intelligence Community Directive 203 (ICD 203). The section after this one sets out that standard. The score was the overlap: the stretch of numbers both ranges share, as a per cent of the whole stretch the two cover together. Ranges that do not meet at all score below zero.

One group saw the sentences with no numbers anywhere. Search for "control group".

> *Participants in the control group specified ranges with, on average, 32% overlap (29, 35) (95% confidence interval) with the ranges specified in the guidelines.*

On average, the readers' ranges and the intended ranges overlapped by 32 per cent. Two ranges that matched exactly would score 100.

Now search for "standard deviations". The standard deviation is the typical distance from the average, as built in Book 0, D5. Here it measures how far people's best numbers for one word spread out, across all four groups, some of whom had been shown the standard's numbers.

> *with the standard deviations for participants' best estimates being: Very Unlikely (21.1), Unlikely (20.7), for Likely (13.0), and for Very Likely (12.9).*

The negative words spread more. Readers disagreed about "unlikely" more than about "likely".

Last, the same word in two studies. Search for "mean best estimates".

> *Our participants' mean best estimates of … 73 and 75 for Likely … were substantially more extreme than those reported in Budescu et al.: … 54 for Likely*

Budescu's study used sentences from climate reports. Wintle's 73 was with sentences from intelligence reports, 75 with the word on its own. Many readers in both studies had been shown a standard's numbers. The average reading of "likely" was 54 in one and 73 to 75 in the other.

```
73 minus 54 = 19
75 minus 54 = 21
```

About 20 percentage points apart, for one word read in two studies. Wintle's team offers several reasons, including the different settings and the way the numbers were shown. They also report, from earlier studies, that readings move with how severe the outcome is.

Figures quoted in this illustration are from [196].

WHERE THIS PICTURE BREAKS. Wintle's readers were members of the public who had signed up to a research project, mostly in Australia. Only 2 per cent were in India. Their numbers describe that group reading English sentences from intelligence reports. They are not a measure of how a district health officer or a patient in India reads a medical report.

The 32 per cent is an average overlap with one particular standard. A different standard, with different ranges for each word, would give a different score. What survives the change is the finding that the same word carried very different numbers for different readers.

MUST KNOW POINTS FOR YOU.

- Agreeing on a word is not agreeing on a number. Before you sign a jointly written sentence that says "likely" or "a serious possibility", ask each author for the number they have in mind.
- "A word is safer than a number because it does not over-commit" is the error. The word still lands as a number, but the reader chooses it.
- When you tell a family a complication is "unlikely", they may hear a larger chance than you mean. In one study, mothers read it as about 27 in 100, and doctors as 16. If the number matters to their decision, say the number.
- People agree on the order of the words far better than on their numbers.
- Negative words spread most.
- None of these studies was done in India, and none tested Hindi or any other Indian language.

COMMON MISREADING. "Only patients misread these words; doctors all mean the same thing by 'likely'." The doctors and medical students in Shaw and Dear's study gave lower and upper fig-

ures of 64 and 90 per cent for "likely". Kent's board were experts who had worked on the very sentence, and they spread from 20 to 80 per cent.

EXERCISE 1 (retrieval). From memory: what happened to the phrase "a serious possibility" in 1951, and what range of degrees of belief did Kent's board hold? Then give one measured finding about how readers turn probability words into numbers, and one thing readers do agree on.

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (interpretation). The 2024 review's Table 2 gives estimates for the word "possible" from Shaw and Dear's 1990 study. All are in per cent.

group	lower	central	upper
mothers	45	62	79
doctors and medical students	27	46	65

Say what the rows show, and what they do not show.

EXERCISE 3 (critique). Here is a made-up line from a district report on a school screening programme.

> *There is a serious possibility that the programme will reduce overweight among the children screened.*

The district health officer will use it to decide whether to fund a second year. Find the flaw. Then rewrite the line so the officer reads what the writer meant.

EXERCISE 4 (teaching). A journalist asks you: "Doctors say a side effect is 'unlikely'. Isn't that clear enough?" You have three minutes and no jargon. Give your answer, ending with one sentence she can quote.

12 · Calibrated scales: fixing what the words mean, and putting the number beside the word

Checked against the sources as they stood on 2026-10-10.

DEFINITION. A calibrated scale of probability words is a published table that fixes, for each word, the range of degrees of belief it stands for. It is not a fact about the words, and two bodies can fix the same word differently.

Sherman Kent, of the US Board of National Estimates, set out a five-band chart (Kent 1964). Each band is a centre with a margin, "give or take", in per cent. The band runs from the centre minus the margin to the centre plus the margin.

Kent's word	centre (%)	give or take (%)	band (%)
almost certain	93	6	87 to 99
probable	75	12	63 to 87
chances about even	50	10	40 to 60
probably not	30	10	20 to 40
almost certainly not	7	5	2 to 12

The Intergovernmental Panel on Climate Change (IPCC) gave the authors of its Fifth Assessment Report a likelihood scale (Mastrandrea and colleagues, 2010, Table 1). Its ranges overlap by design: each range above or below the middle one reaches to the end of the scale on its side. Its note calls the boundaries "fuzzy".

IPCC term	likelihood of the outcome (%)
virtually certain	99 to 100
very likely	90 to 100
likely	66 to 100
about as likely as not	33 to 66
unlikely	0 to 33
very unlikely	0 to 10
exceptionally unlikely	0 to 1

The note allows three more terms in limited cases; this book uses the seven.

The IPCC keeps a second, separate scale, which it calls confidence. It has five words, from "very low" to "very high", and authors choose one from the evidence ("limited", "medium" or "robust") and from how far the studies or experts agree ("low", "medium" or "high"). The note says this confidence should not be read as a probability. It is a judgement of the grounds, close to GRADE's certainty. It is not a degree of belief, which is what this book calls confidence.

US Intelligence Community Directive 203 (ICD 203), signed in January 2015 and amended in January 2022, fixes seven bands that run end to end, from 1 to 99 per cent. It gives no word for below 1 or above 99. Each band starts where the one below it stops, so the bands share only their end points.

ICD 203 word	band (%)
almost certain	95 to 99
very likely	80 to 95
likely	55 to 80
roughly even chance	45 to 55
unlikely	20 to 45
very unlikely	5 to 20
almost no chance	1 to 5

ICD 203 also allows a second set of words for the same bands, such as “probable” or “probably” for 55 to 80 per cent. It asks you not to mix the two sets.

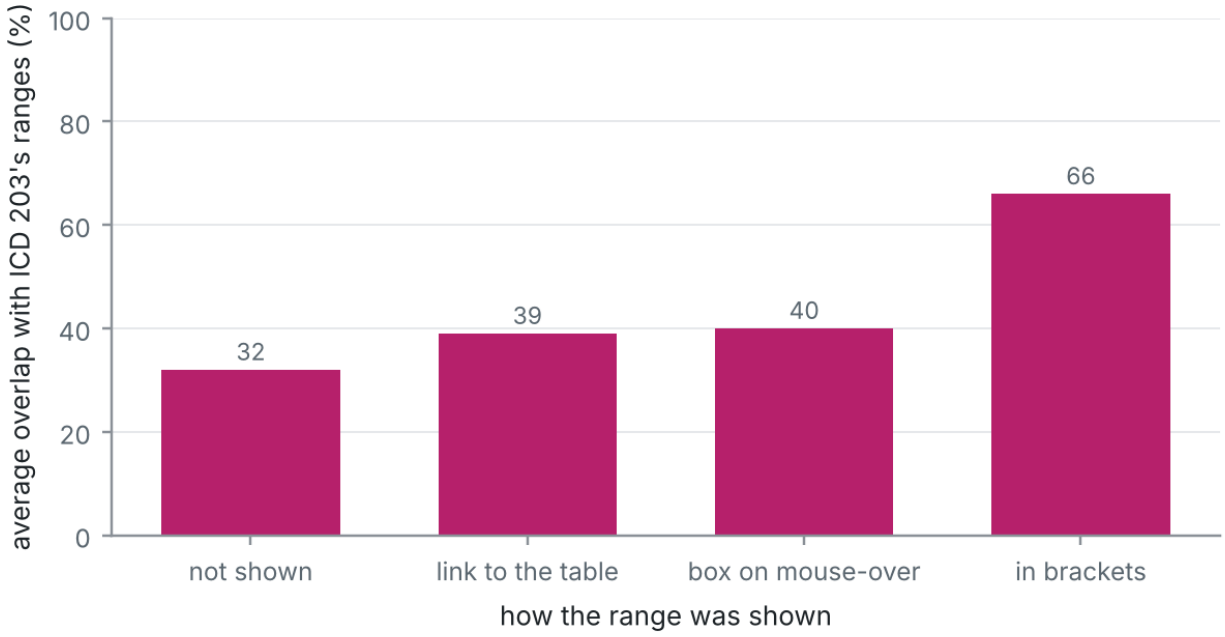
ICD 203 also has a confidence level, such as “high confidence”: a judgement of the logic and evidence behind an assessment, like the IPCC’s confidence. ICD 203 forbids putting a confidence level and a likelihood in the same sentence.

A scale fixes a word only for a reader who sees it. [197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 189, 190, 226, 227, 228]

IN PLAIN TERMS. A word like “likely” has no fixed number. So some bodies publish a table that fixes one.

Different bodies fixed it differently. For the US intelligence services, “likely” means 55 to 80 per cent. For the climate scientists of the IPCC, it means 66 to 100 per cent.

So pick one table, say which one, and put the number in brackets every time.



Wintle's experiment: how far 924 readers' ranges overlapped ICD 203's, by how the range was shown. A link to the table gained 7 points over showing nothing, a mouse-over box 8, and the range in brackets beside the word 34: about twice the overlap. Even in brackets, a third of the overlap was missing.

Illustration 1. The last section showed that readers turn one word into many numbers. A scale is the fix that bodies have tried. Build three of them, then watch one number change its word.

Start with Kent's chart. Open *Words of Estimative Probability* (1964), at https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf, and find the chart under "An Odds Table". Each row gives a centre and a margin. The band is the centre minus the margin, up to the centre plus the margin.

```
93 minus 6 = 87 and 93 plus 6 = 99
75 minus 12 = 63 and 75 plus 12 = 87
50 minus 10 = 40 and 50 plus 10 = 60
30 minus 10 = 20 and 30 plus 10 = 40
7 minus 5 = 2 and 7 plus 5 = 12
```

Look at what the chart leaves out. No word covers 12 to 20 per cent, 60 to 63, below 2 or above 99. Kent fixed five words, not every number.

Now the IPCC scale, in Table 1 of the guidance note at https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf. Its "likely" runs from 66 to 100 per cent. Its "very likely" runs from 90 to 100. So every number that is "very likely" is also "likely". The writer is meant to pick the narrower word when it fits.

Last, ICD 203, at https://www.dni.gov/files/documents/ICD/ICD-203.pdf. Search for "01-05%". The row of bands sits under the words. Its "likely" runs from 55 to 80 per cent, and "very likely" starts where it stops.

Now take one judgement. A colleague tells you she is 70 per cent sure that a new weight-management clinic will cut the hospital's referrals. The colleague is made up. Find the band that holds 70 on each scale.

scale	band that holds 70 (%)	word
Kent	63 to 87	probable
IPCC	66 to 100	likely
ICD 203	55 to 80	likely

So far the word travels well. Now she becomes less sure, and says 58 per cent.

scale	band that holds 58 (%)	word
Kent	40 to 60	chances about even
IPCC	33 to 66	about as likely as not
ICD 203	55 to 80	likely

Here is the trap. Suppose she wrote "likely" in a note that follows ICD 203. Someone copies that word into a report written to the IPCC scale. The reader of that report now hears 66 per cent or more. Her 58 per cent has been raised by the copying, and no number was ever written down to show it.

The safe route between two scales always goes through the number. Turn the word back into its number on the first scale. Then find that number's word on the second.

For the clinic there is a fourth scale. Hashim's 2024 review, which the last section met, proposes five words for talking with patients, each with a number "per 100 cases". It gives "likely" as 70 in 100, with a range of 60 to 80. That is one author's proposal from the studies he reviewed. No body has adopted it.

Figures quoted in this illustration are from [15, 229, 230, 186].

WHERE THIS PICTURE BREAKS. The bands are conventions with soft edges. Kent writes "give or take", and the IPCC calls its boundaries fuzzy. ICD 203 does not say which word a number exactly on a shared edge takes, such as 80. So the table above is a lookup, not a measurement. A judgement of 81 per cent and one of 79 per cent are not really different kinds of thing.

None of the three scales was written for medicine. Kent and ICD 203 were written for intelligence analysts, and the IPCC scale for climate assessments. Borrowing one for a district report is a choice you make, and the reader can follow it only if you name it.

Illustration 2. The IPCC published its scale, and readers still misread it. Here is how badly, and what fixed most of it.

Open Wintle and colleagues' 2019 paper at https://pmc.ncbi.nlm.nih.gov/articles/PMC6469752/ and search for "1 in 8". It reports an earlier online survey by Budescu and colleagues. Readers met sentences from IPCC reports, with a link to the IPCC table. They clicked the link only 1 in 8 times. Now search for "18.5%".

> *They reported an improvement in mean percentage overlap with prescribed ranges in the IPCC conversion table from 18.5% in their equivalent to our Table condition to 33.6% in their equivalent to our Brackets condition.*

With the table a click away, readers' ranges overlapped the IPCC's by 18.5 per cent on average. Even with the range in brackets, they reached only 33.6 per cent. Wintle's team suggests the overlapping bands may be part of the reason. Search for "not intuitive".

> *E.g., is the range for 'Likely' supposed to mean 67–90%, or 67–100%? The latter is the intended meaning, but this is not intuitive.*

Readers expect each word to stop where the next begins.

Wintle's own experiment used ICD 203, whose bands do stop where the next begins. Its 924 readers were split into four groups. Each group saw the same sentences, shown in one way. Put the four results side by side, in per cent of overlap with ICD 203's ranges. Wintle's overlap is the shared part of the two ranges as a share of the whole stretch they cover together: 1 to 15 against 5 to 20 shares 10 points of 19, about 53 per cent.

how the range was shown	average overlap (%)
not at all	32
a link to the full table	39
a small box that appears when the mouse is over the word	40
in brackets beside the word	66

Work out what each format gained over showing nothing, in percentage points.

```
39 minus 32 = 7
40 minus 32 = 8
66 minus 32 = 34
66 divided by 32 is about 2.06
```

The link and the box gained 7 and 8 points. The brackets gained 34, about twice the overlap of showing nothing. The table existed in every case. What changed was whether the number sat where the reader's eye already was.

Now use it. Here is a made-up line from a district report.

> *It is likely that the school canteen scheme will reach its target this year.*

Choose a scale, name it once, and bracket the range.

> *We judge it "likely" (55 to 80 per cent, on the scale of US Intelligence Community Directive 203) that the school canteen scheme will reach its target this year.*

If you mean a single number, give that instead: "likely (about 70 per cent)".

Figures quoted in this illustration are from [196].

WHERE THIS PICTURE BREAKS. Wintle's readers were members of the public, mostly in Australia, reading sentences from intelligence reports online. The 66 per cent measures how far their ranges overlapped ICD 203's, not whether they then made better decisions. It is the best evidence here that brackets help, and it was not measured on patients, doctors or officers in India.

Even with brackets, a third of the overlap was missing. The number beside the word narrows the readings; it does not make every reader read the same.

MUST KNOW POINTS FOR YOU.

- "A scale makes the word precise" is the error. A scale fixes a word only for a reader who sees it. Put the range in brackets beside the word, every time.
- Never move a probability word from one scale to another as it stands. Turn the word back into its number, then find that number's word on the new scale.
- Name your scale the first time you use one of its words, and keep to that one scale through the whole document.
- "High confidence" in an IPCC report is not a probability. It judges the evidence and the agreement behind a finding.
- On the IPCC scale "likely" runs from 66 to 100 per cent, so it covers everything "very likely" covers.
- A scale fixes what your word means. It cannot make your judgement right.
- The evidence that brackets help comes from members of the public, mostly in Australia, reading intelligence sentences. Nobody has measured it in Indian clinics or district offices.

COMMON MISREADING. "Official reports use official words, so 'likely' means the same in all of them." It does not. Likely is 55 to 80 per cent in a report written to ICD 203. It is 66 to 100 per cent in one written to the IPCC scale. A report that names no scale fixes nothing at all. Read the report's own scale, or its bracketed number, before you repeat its word.

REPORTING IT. We judge it "likely" (55 to 80 per cent, on the scale of US Intelligence Community Directive 203) that the scheme will reach its target this year.

EXERCISE 1 (retrieval). From memory: what range does "likely" cover on the IPCC scale, and on ICD 203? Which of the two scales has overlapping ranges? What did Wintle and colleagues

find when the range was printed in brackets beside the word? Last, give the three things to do when you write a probability word.

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (critique). Here is a made-up sentence from the discussion of a thesis on a school nutrition programme.

> *It is very likely, with high confidence, that the programme would work equally well in other districts.*

Find what is wrong with how it states its probability. Then rewrite it.

EXERCISE 3 (teaching). The health secretary of your state asks you for one page. Her department's reports use words like "likely" and "unlikely" with no numbers, and two districts have read the same report in opposite ways. She wants a rule for all reports. Write the page, answer first.

PRACTICE. Ten problems on this technique, easiest first. Work them on paper before turning to the answers, which are in the appendix at the back under this section's name.

1. Each band is given as a centre, give or take a margin. Write each as a range from its lowest number to its highest.

centre	give or take
60	15
85	5
10	8
45	20

2. A scale has seven bands, in per cent.

letter	band
A	1 to 5
B	5 to 20
C	20 to 45
D	45 to 55
E	55 to 80
F	80 to 95
G	95 to 99

Give the letter for each number: 12, 50, 72, 88, 97, 30.

3. For each pair of ranges, give the numbers both share. Then say how many percentage points wide that shared part is. The ranges are in per cent.

first range	second range
55 to 80	66 to 100
20 to 45	0 to 33
40 to 60	45 to 55
63 to 87	80 to 95
2 to 12	20 to 40

4. A colleague gives three degrees of belief: 70 per cent, 85 per cent and 58 per cent. For each, give the word on Kent's chart, on the IPCC likelihood scale and on ICD 203.

5. "Likely" is 55 to 80 per cent on ICD 203 and 66 to 100 per cent on the IPCC scale. Which degrees of belief are "likely" on both? What share of ICD 203's "likely" band does that shared part make up, and what share of the IPCC's?

6. In Wintle and colleagues' experiment, about half the readers given the link or the box never used it. Take only the readers who did use it. Their average overlap with ICD 203's ranges was 44 per cent with the link to the table. With the box under the mouse it was 49. With nothing shown it was 32, and with the range in brackets 66.

Give the link's and the box's gain over nothing for these readers, in percentage points. Compare each with the gain from brackets. What does this say about brackets?

7. Here is a worked answer. Find the step that broke, and say why it looks reasonable.

> *An IPCC report calls a change "very likely". A colleague needs it in the words of ICD 203.*

```
IPCC "very likely" = ICD 203 "very likely"
ICD 203 "very likely" = 80 to 95 per cent
so the IPCC report means 80 to 95 per cent
```

8. Here is a worked answer. Find the step that broke.

> *A reviewer is 95 per cent sure of a finding and must use the IPCC scale.*

```
IPCC "likely" starts at 66 and "very likely" starts at 90
so "likely" runs from 66 to 90
95 is above 90, so the finding cannot be called likely
so the only word left is "virtually certain"
```

9. A made-up district report says: "It is very likely that the school canteen scheme will halve sugary drink sales this year." It names no scale. Decide what to compute, compute it, and say what the sentence does not establish.

10. A newspaper health column says: "A study of 924 people found that putting a number beside words like 'likely' doubles how well people understand risk." Decide what to compute, compute it, and say what the claim does not establish.

13 · Stating how confident you are, and saying why

DEFINITION. A confidence statement says how sure you are of one claim, and why, in a form that someone else can check. It has four parts.

part	what it says
the claim	one claim, worded so that it could later be found true or false: in whom, what, compared with what, over what time
the confidence	your degree of belief in the claim, as a word from a named scale with its number
the grounds	the kind of evidence the belief rests on and, where the evidence has been rated, its certainty
what would change it	a finding that could turn up and would move your number, and which way it would move it

The confidence belongs to the person who states it. Two people with the same grounds can state different numbers. The grounds are what let a reader see why, and argue with the step from grounds to number.

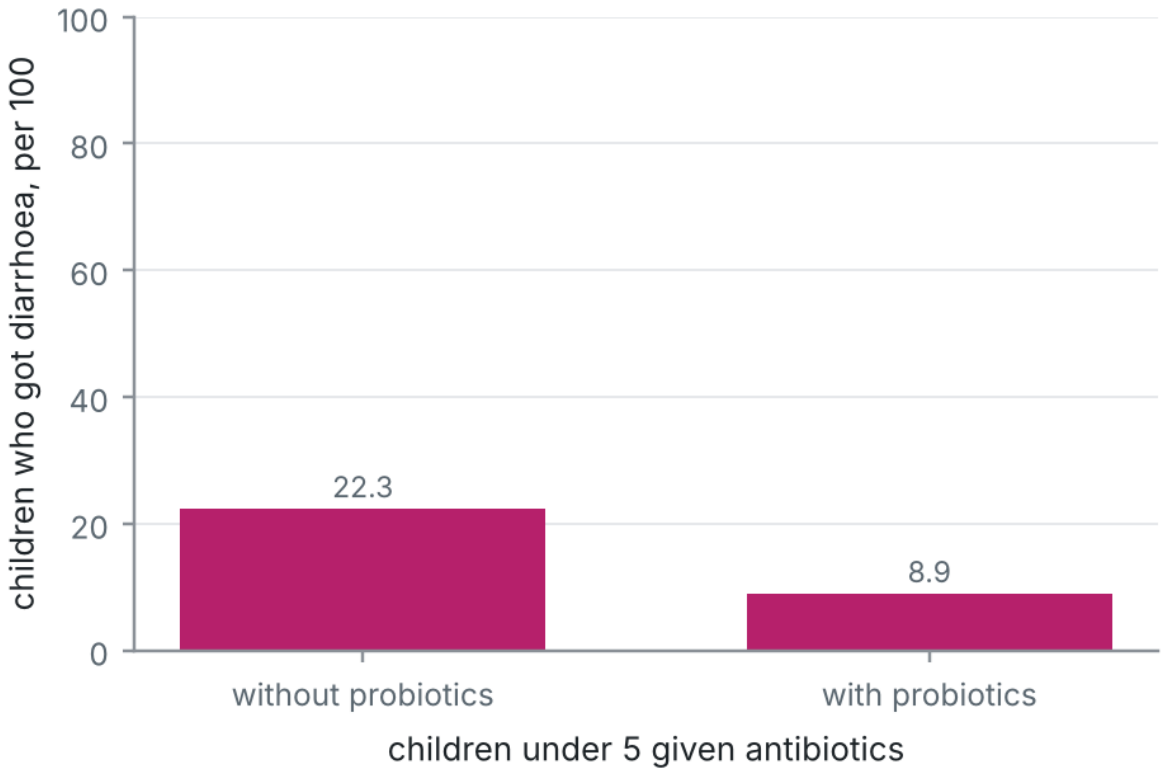
The certainty of the evidence, rated with GRADE (Grading of Recommendations Assessment, Development and Evaluation), belongs among them. It is not the confidence. GRADE gives no number for its levels, so the step from a level to a degree of belief is the writer's judgement.

The Intergovernmental Panel on Climate Change (IPCC) and Intelligence Community Directive 203 (ICD 203) both use the word confidence for the strength of the grounds. This book calls that the grounds, and keeps confidence for the degree of belief. The IPCC asks for a "traceable account": a description of the evidence and agreement behind each finding.

A claim about an effect says what something does on average in a group of people: "probiotics given with antibiotics lower the share of young children who get diarrhoea". A claim about a single event says what will happen once: "this child will not get diarrhoea". Even an effect you are very confident of leaves a single event uncertain, because people in a group do not all respond alike. [180, 231, 232, 233, 234, 220, 134, 142, 235, 139, 236, 237, 202, 238, 239]

IN PLAIN TERMS. The number is yours. Someone else could look at the same evidence and pick another. That is why you give the reasons: they let the listener see how you got there.

The listener decides the form. A resident wants the reasons in full. A journalist wants one clear sentence. A health secretary wants the answer first.



Children under 5 on antibiotics in the trials behind the Cochrane example row (numbers from the Cochrane Handbook's example table): 22.3 in 100 got diarrhoea without probiotics and 8.9 in 100 with them, 13.4 fewer. You can be about 75 per cent sure of the effect on the group (the worked choice below) and still not promise one child anything: most stay free either way (77.7 and 91.1 in 100), and some get it even with the probiotic.

Illustration 1. Here is a sentence from a journal club. The speaker is made up; the evidence behind it is real.

> *Probiotics probably prevent diarrhoea in kids on antibiotics.*

Ask four questions of it. Which claim, exactly? Which children, compared with what? How sure, as a number, and on what scale? On what grounds? What would change the speaker's mind? The sentence answers none of them.

Worse, "probably" can be read two ways. It is Cochrane's word for moderate certainty of the evidence, from the section on certainty words. It is also an everyday word for a degree of belief. And ICD 203 allows "probably" for its 55 to 80 per cent band. The listener cannot tell which the speaker meant.

Now rebuild it. Go back to Figure 14.1.b in chapter 14 of the Cochrane Handbook, at https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14, the table you met in the section on GRADE. Take the row for children under 5.

The claim. Write it so that a later study could show it false.

> *In children under 5 who are given antibiotics, giving probiotics as well lowers the share who get diarrhoea.*

This claims a direction, a fall, and not a size. A claim about the size, "by about 13 in 100", would need more confidence in the estimate than a claim about its direction.

The grounds. Randomised trials, pooled in the table. Certainty moderate (GRADE), down one level for risk of bias. Footnote b says why: many children were lost to follow-up.

The confidence. This is your judgement, and nothing in the table makes it for you. The whole range of the risk ratio lies below 1, so the trials point one way. But many children were lost, and the missing ones could change the picture. Say you settle on 75 per cent. On ICD 203 that is "likely", 55 to 80 per cent.

What would change it. A large trial in such children, with few lost, that found no difference would lower your number. Another such trial that found the same fall would raise it.

Now write it, in two sentences so the certainty and the likelihood do not share one.

> *I judge it "likely" (about 75 per cent, ICD 203 scale) that, in children under 5 given antibiotics, adding probiotics lowers the share who get diarrhoea. This rests on randomised trials rated moderate certainty (GRADE), down for children lost to follow-up; a large trial with few losses and no difference would lower my number.*

Every part can now be checked. A colleague who reads the same table and says 60 per cent can see exactly where you differ: in the step from the grounds to the number.

Figures quoted in this illustration are from [148].

WHERE THIS PICTURE BREAKS. The 75 per cent is a worked choice, not the right answer. GRADE gives no number for moderate certainty, and a careful reader of the same row could defend 65 or 85. The statement does not make your number correct. It makes it open to argument.

The table is an example in a handbook chapter, not a full review. Its row tells you about the children in those trials, given those probiotics, followed for 10 days to 3 months. A claim about other children or other products needs grounds of its own.

Illustration 2. Same row, a different question. On the children's ward, a mother whose son is under 5 and has just started antibiotics asks you: "If he takes the probiotic, will he be all right?" She is made up.

Your 75 per cent does not answer her. It is your degree of belief that probiotics lower the share of such children who get diarrhoea. She is asking about one event: her son.

Look at the table's middle columns. Without probiotics, 22.3 in 100 such children got diarrhoea. With them, 8.9 in 100. Work out how many stayed free of it.

```
100 minus 22.3 = 77.7
100 minus 8.9 = 91.1
22.3 minus 8.9 = 13.4
```

So about 78 in 100 children like her son stayed free of diarrhoea without probiotics, and about 91 in 100 with them. The difference is about 13 children in 100. Those are the table's estimates from the trials: the risk in the control groups, and that risk with the pooled effect applied. They are the best starting point for one child. A later section takes up how a frequency in a group becomes a degree of belief about one case.

Now answer her in four parts too.

> *Most children his age on antibiotics do not get diarrhoea. With the probiotic, about 9 in 100 do; without it, about 22 in 100. The trials behind those numbers had problems, so the true numbers could be somewhat different. If he gets loose stools anyway, it does not mean the probiotic failed.*

The effect claim got a likelihood word. Her question got frequencies "in 100", which she can picture. Both are honest. They answer different questions.

Figures quoted in this illustration are from [148].

WHERE THIS PICTURE BREAKS. The frequencies come from the trials' children. If her son differs from them in a way that matters, such as being very ill or on a different antibiotic, the numbers fit him less well. And they rest on moderate-certainty evidence, so they are estimates, not counts from her hospital.

Some single events have no group to borrow from. Then the degree of belief rests on judgement alone, and the grounds should say so.

Illustration 3. One statement, three readers. Take the under-5 row once more, and write it for each of the three people this course keeps returning to.

A first-year resident, ten minutes and a whiteboard. Give all four parts, in full, and show the step from grounds to number. Draw the table's row. Show the risk ratio's range lying wholly below 1, and the footnote that cost one level. Say: "Moderate certainty, so I land at about 75 per cent, 'likely' on ICD 203. You might land at 65. Tell me why, and we can argue about the step." Then say what would change it.

A journalist, three minutes, no jargon. Lead with the frequencies, then your confidence as chances, then why it is not higher.

> *On the trials' estimates, giving young children probiotics with their antibiotics cut diarrhoea from about 22 in 100 children to about 9 in 100. I'd put the chance that the benefit is real at about 3 in 4. Not higher, because many children dropped out of the trials.*

That last sentence sets how sure she should be. It tells her the evidence is decent, not settled, and still gives her something to print.

A health secretary, one page, answer first, no methods. She is asking a different question: should the state do it? The section on recommendations showed that certainty of the evidence is one input of four.

The same table has a row for side effects: 1.8 in 100 children without probiotics and 2.3 in 100 with them, rated low certainty, with Cochrane's comment that there may be little or no difference.

> *Answer: it is likely (about 75 per cent, ICD 203 scale) that probiotics given with antibiotics reduce diarrhoea in children under 5: about 13 fewer cases in every 100 children treated. The evidence is moderate certainty (GRADE). Side effects: there may be little or no difference (low certainty). The cost of the product is not in this evidence and is needed before deciding. A large trial with few children lost would settle the benefit either way.*

The order changed three times. The four parts did not. Drop the grounds for the journalist, and she cannot tell your 3 in 4 from a guess. Drop the number for the secretary, and she cannot weigh the benefit against a cost.

Figures quoted in this illustration are from [148].

WHERE THIS PICTURE BREAKS. The three forms are trimmed to their readers, and trimming has a floor. If a journalist's space allows only one sentence, keep the number and one reason, and drop the rest. If a secretary's page cannot hold the grounds, give them in an attached note. A statement that has lost its grounds has become a bare assertion.

The secretary's page reports evidence. It is not a recommendation. A recommendation also needs the cost, what can be delivered in her districts, and what families value.

MUST KNOW POINTS FOR YOU.

- A number with no grounds cannot be argued with, and a word with no number cannot be checked.
- Keep the certainty of the evidence and your confidence in separate sentences, each named. Write "moderate certainty (GRADE)" for the grounds and "likely (about 75 per cent, ICD 203)" for the confidence.
- "Moderate certainty means about 60 per cent" is the error. GRADE puts no number on its levels. Choose the number, and show the step.
- For one person, give the frequency in people like them, as a number in 100.
- Before a meeting that will agree a judgement, write your own number down. The IPCC's guidance warns that groups tend to converge on one view and become overconfident in it.
- "What would change my mind" must name a finding that could actually turn up. "If new evidence emerges" names nothing.
- Change the order and the words for each reader, never the parts. Give a journalist chances "in 100" and say plainly why your number is not higher.

COMMON MISREADING. "Giving a number makes me sound more certain than saying 'probably'." It does the opposite. "75 per cent" says, out loud, that you think there is a 1 in 4 chance you are wrong. "Probably" hides how unsure you are, and lets each listener choose their own number for it. A number with its grounds is the most modest way to state a belief, because it can be checked.

REPORTING IT. I judge it "likely" (about 75 per cent, ICD 203 scale) that, in children under 5 given antibiotics, adding probiotics lowers the share who get diarrhoea. This rests on randomised trials rated moderate certainty (GRADE), down for children lost to follow-up. A large trial with few losses and no difference would lower my number.

EXERCISE 1 (retrieval). From memory: name the four parts of a confidence statement. Say what the IPCC calls the written description of the grounds. Then say why your degree of belief in an effect is not the chance that one patient will benefit.

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (critique). At a made-up district meeting on a community yoga programme, an officer says:

> *We are 100 per cent confident the yoga programme will reduce obesity in the district. Everyone in the team agrees.*

Find what is missing or wrong. Then rewrite the statement, leaving a blank wherever you would need a fact the officer has not given.

EXERCISE 3 (teaching). A journalist rings you about the same Cochrane example table, the row for children over 5. In those trials, diarrhoea went from about 11 in 100 children without probiotics to about 9 in 100 with them. The difference could be anywhere from about 5 fewer to about 2 more in 100. The certainty is low, for risk of bias and imprecision. She asks: "So do probiotics work for older kids?" You have three minutes and no jargon. Give your answer, ending with one sentence she can quote.

14 · The base rate: how often claims like this one hold up

Checked against the sources as they stood on 2026-10-10.

DEFINITION. A base rate is how often claims of one kind have turned out true when they were tested. The kind is the reference class: the group of past claims, alike in what matters, among which you count the present claim. The evidence particular to that claim then moves you from there.

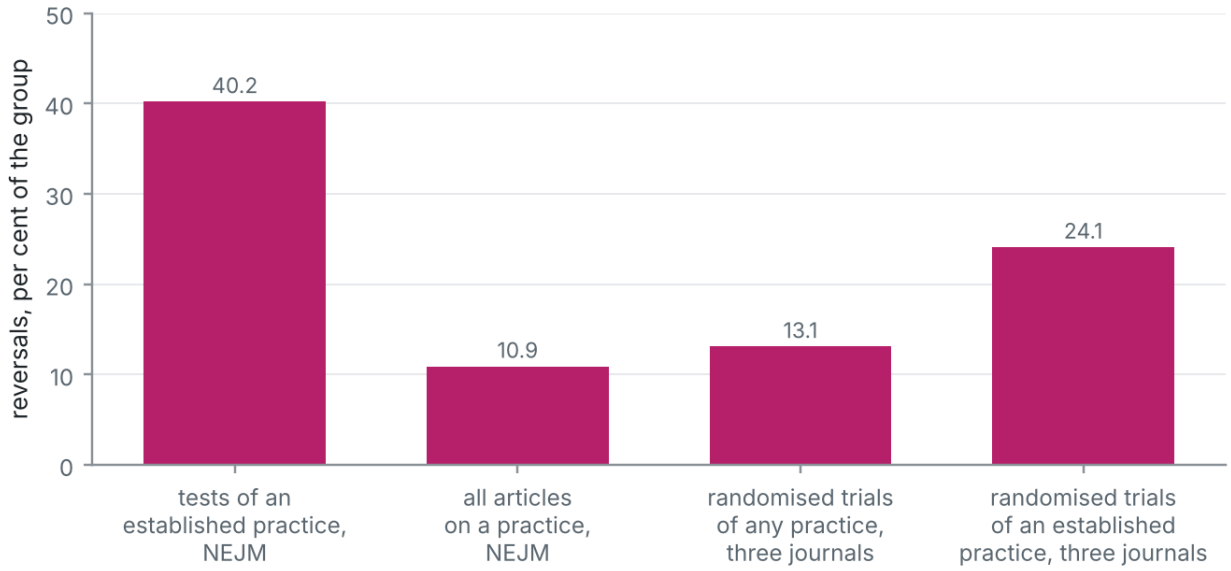
A medical reversal is an established practice found by a later study to be no better than a lesser or earlier standard of care.

Prasad and colleagues (2013) read every original article in the New England Journal of Medicine (NEJM) from 2001 to 2010: 2044 articles, of which 1344 concerned a medical practice. Of these, 363 tested an established practice: 146 (40.2 per cent) reversed it, 138 upheld it and 79 reached no clear answer.

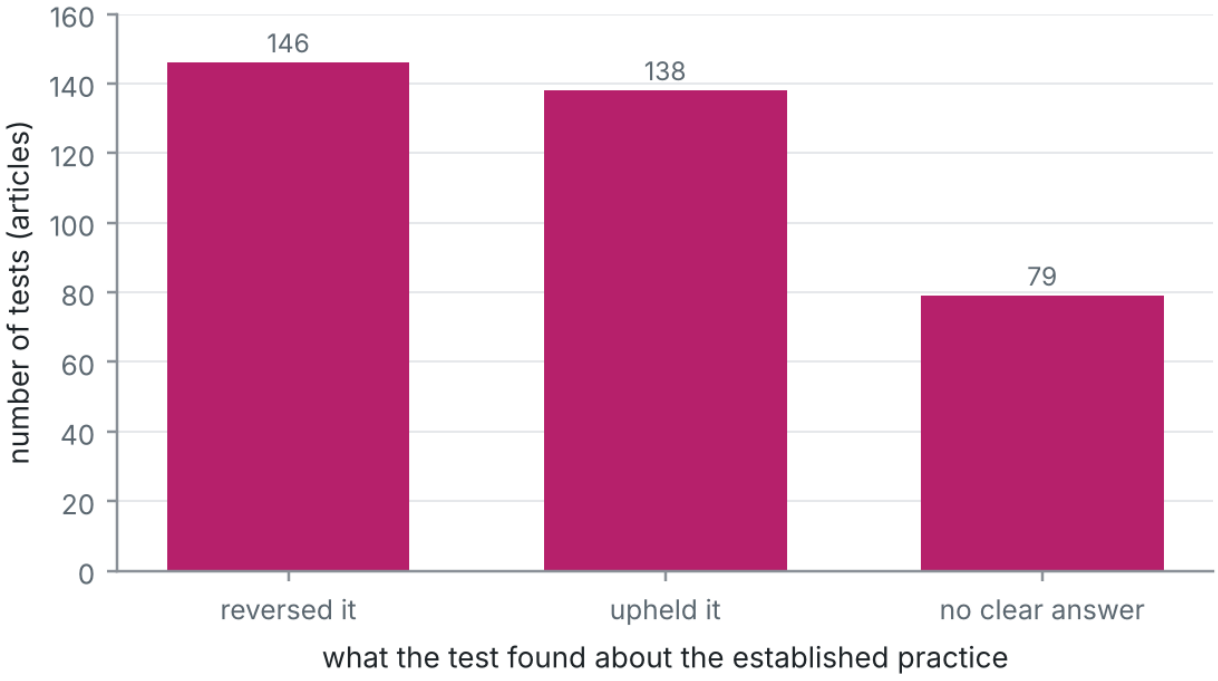
Herrera-Perez and colleagues (2019) read the randomised trials of a medical practice in three journals. Of 3017 such trials, 396 were reversals, 13 per cent of all the trials.

A rate measured in one class does not carry over to another. [240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257]

IN PLAIN TERMS. Which claims count as "like it" is a choice, and the number changes with the choice. Pick the group that matches your claim in what matters, and say which group you picked.



One question, four answers. "How often is a medical practice reversed?" gives 40.2 per cent among tests of an established practice in the New England Journal of Medicine, 2001 to 2010; 10.9 per cent among all its articles on a practice; 13.1 per cent among randomised trials of any practice in three journals; and 24.1 per cent among those trials that tested an established practice. Each is right for its own group. The base rate depends on the reference class you choose.



What became of the 363 tests of an established practice in the New England Journal of Medicine, 2001 to 2010: 146 reversed it (about 40 per cent), 138 upheld it (about 38 per cent) and 79 gave no clear answer (about 22 per cent). The three add back to 363. Of the 284 clear answers, about 49 per cent upheld the practice.

Illustration 1. This section stands on the last one, where you put a number on how sure you are of one claim. Here you find where that number should start.

Go back to the time before the 2014 trial's result. "Eat breakfast to lose weight" was standard advice, as the first section of this book showed. A large trial of the advice was under way. Before its result came out, how sure should you have been that it would support the advice?

Here is the tempting answer. The advice was everywhere, and there was a story of how it might work. It felt like 90 per cent. Nothing in that feeling was counted.

Now count. Ask what kind of claim this was: an established practice, about to be tested again in a large study. That group of past claims is the reference class. Prasad and colleagues counted it in ten years of one journal, the New England Journal of Medicine (NEJM), from 2001 to 2010. Their abstract is at https://pubmed.ncbi.nlm.nih.gov/23871230/. Search it for 363.

what the study found about the established practice	articles
no better than a lesser or earlier standard: a reversal	146
upheld it	138
no clear answer	79
all tests of an established practice	363

Turn each count into a share of the 363, as Book 0 did with any proportion (Book 0, A5). Then check that the three counts add back to the total.

```
146 divided by 363 is about 0.4022
138 divided by 363 is about 0.3802
79 divided by 363 is about 0.2176
146 plus 138 plus 79 = 363
```

About 40 per cent of these tests reversed the practice, 38 per cent upheld it, and 22 per cent settled nothing. Now leave out the tests with no clear answer.

```
146 plus 138 = 284
138 divided by 284 is about 0.4859
```

Of the tests that gave a clear answer, about 49 per cent upheld the practice. That is close to evens: odds of 138 to 146, or 69 to 73, that a clear answer would uphold it.

A test with no clear answer does not support the advice either, so count all 363: 138 of 363. Among clear answers only it would be 49; say which you used.

So a fair starting confidence that the trial would support the advice was about 38 per cent, not 90. That number is a base rate: a frequency counted over a group of past claims, used as your degree of belief in one new claim. A base rate is a share, not a rate in Book 0's sense; the name is the usual one. It is where you start, not where you finish. The advice also had evidence of its own: studies that watched people and found that skippers more often had overweight. The next section shows how evidence for one claim moves you from the base rate. It moves you from 38 per cent, not from 90.

When the trial came out in 2014, it found that the advice made no detectable difference to weight, as an earlier section described.

Figures quoted in this illustration are from [258].

WHERE THIS PICTURE BREAKS. The class is a loose fit. Prasad's articles are mostly drugs, procedures and tests in one leading medical journal. The breakfast trial tested advice about eating, and appeared in a nutrition journal. Nobody in the sources this book holds has counted how often advice about food is reversed when it is tested.

The class is also chosen by someone else's doubt. A practice is tested again mostly when someone doubts it, and Prasad and colleagues write that "testing of standard of care is rarely done". So 40 per cent is a rate among practices doubted enough to be tested. For a practice nobody has questioned, it may be too high a start.

Illustration 2. Here is the trap with base rates. Ask "how often is medicine wrong?" and you can get any of several numbers. Each one is right for the group it counted. Work them out.

Prasad and colleagues give two. Their 146 reversals were out of 363 tests of an established practice. They were also out of all 1344 articles on any practice, old or new.

Herrera-Perez and colleagues read the randomised trials of a medical practice in three leading journals (open their paper at https://pmc.ncbi.nlm.nih.gov/articles/PMC6559784/). They found 396 reversals among 3017 trials. They also say that 1373 of the 3017 tested a new practice, used only inside research. A new practice cannot be a reversal, because a reversal is an established practice found no better. So take those trials out.

```
146 divided by 363 is about 0.4022
146 divided by 1344 is about 0.1086
396 divided by 3017 is about 0.1313
3017 minus 1373 = 1644
396 divided by 1644 is about 0.2409
```

the group counted	reversals	out of	share, %
tests of an established practice, NEJM 2001-2010	146	363	40.2
all articles on a practice, NEJM 2001-2010	146	1344	10.9
randomised trials of any practice, three journals	396	3017	13.1
randomised trials of an established practice, three journals	396	1644	24.1

The same question gives 10.9 per cent, 13.1, 24.1 or 40.2, from two careful studies. The 13 per cent that Herrera-Perez and colleagues print (13.1 to one decimal) is right for its group. It is the wrong base rate for a trial of an established practice, because almost half its denominator is trials that could never have been reversals.

So choose the class the way you choose a denominator (Book 0, A5). Ask what your claim is, and count only past claims of that kind. For "this established practice will hold up when a large trial tests it", the closest counts give a reversal rate of about 24 to 40 per cent.

These counts include practices close to this course. Among Herrera-Perez's examples is a weight-loss programme with a wearable activity tracker added. After 24 months, the group without the tracker had lost significantly more weight: 5.9 kg against 3.5 kg (difference 2.4 kg, 95 per cent interval 1.0 to 3.7).

Figures quoted in this illustration are from [258, 259].

WHERE THIS PICTURE BREAKS. The 24.1 per cent is this book's arithmetic, not a number the authors print. Here is where the 1644 trials went.

```
3017 minus 1373 = 1644
1229 plus 415 = 1644
415 minus 19 = 396
```

So the 1644 holds the 396 reversals, 19 trials first called reversals and then dropped after a search of systematic reviews, and 1229 trials that upheld the practice or settled nothing. The 24.1 counts the 19 as tests that did not reverse. The two studies also define their groups differently: Prasad counted every design, while Herrera-Perez counted randomised trials only.

None of these classes was measured in India. Of Herrera-Perez's 396 reversals, 92 per cent came from studies in high-income countries. No source this book holds gives a reversal rate for practice in India.

MUST KNOW POINTS FOR YOU.

- Before you judge one claim, name its reference class out loud. Then look for a count of how often claims of that class held up. Start your confidence there.
- Of 363 established practices tested again in one leading journal, 138 were upheld: about 38 per cent.
- "Forty per cent of medicine is wrong" is the error. Quote the class with the number every time.
- A count of reversals by type is not a rate.
- Check the denominator before you use a base rate. The 13 per cent of trials that were reversals counts trials of new practices, which could never be reversals.
- These counts come from leading journals, and 92 per cent of Herrera-Perez's reversal trials were in high-income countries. When you start from them for a practice in an Indian district, say that you are borrowing the number, and from where.
- A small class swings. Prefer the pooled count, and distrust a base rate built on a few dozen claims.
- A base rate is where you start, not where you end.

COMMON MISREADING. "A decade of research found that 40 per cent of medicine is wrong." The 40.2 per cent is not a share of medicine. It is a share of the 363 established practices that someone chose to test again and that one leading journal published.

Practices get tested again mostly when someone doubts them, and the authors say such testing is rare. Of all 1344 articles on a practice in the same ten years, 146 were reversals: 10.9 per cent. Neither number tells you what share of everyday practice would fail a test, because most of it has never had one.

REPORTING IT. Before this trial, my starting confidence that it would uphold the practice was about 38 per cent. In one count, 138 of 363 tests of an established practice in a leading journal upheld it, and 146 reversed it [ref]. That count comes from one leading American journal, and a later count found most reversal trials were in high-income countries. I have adjusted from it for the evidence below.

EXERCISE 1 (retrieval). From memory: say what a base rate is and what a reference class is. Then give the count Prasad and colleagues found for tests of an established practice, and the group it was counted in.

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (critique). At a district review meeting, an officer says: "This practice is in the national guidelines and every doctor here uses it. So I am 99 per cent sure it works." Find the specific flaw, and say what you would ask for.

EXERCISE 3 (teaching). A journalist rings you: "Is it true that medicine keeps changing its mind?" You have three minutes. Give one quotable sentence, a number with its group, and say what the number does not show.

EXERCISE 4 (design). Pick one plan of your own. It might be a thesis finishing on time, a paper accepted at the first journal you send it to, or a programme you are planning. Write down its reference class, where you would find a count for it, and what you would do if no count exists.

PRACTICE. Ten problems on this technique, easiest first. Work them on paper before turning to the answers, which are in the appendix at the back under this section's name.

1. Of 48 tests, 14 gave result A, 20 gave result B and 14 gave result C. Give each as a per cent of the 48, to one decimal place. Check that the three counts add to 48.

2. In a class of 80 past claims, 20 held up when tested. Give the base rate as a per cent, the chance that a claim of this class fails, and the odds that one holds up.

3. Of 50 tests, 18 reversed a practice, 22 upheld it and 10 gave no clear answer. Give the share reversed of all 50, and of the tests with a clear answer. Say which is larger, and why.

4. Prasad and colleagues counted the articles in the New England Journal of Medicine, 2001 to 2010, that tested an established practice. There were 363. 146 reversed it, 138 upheld it and 79 gave no clear answer. Give each as a per cent. Then give the odds, among the tests with a clear answer, that a test upheld the practice.

5. Herrera-Perez and colleagues found 396 reversals among 3017 trials in three journals. Of the 3017, 1373 tested a new practice. Give the reversals as a share of all the trials. Then give them as

a share of the trials of an old, established practice. Which share would you start from, for a trial that tests an established practice?

6. Prasad and colleagues give their count year by year. In 2006, 12 of 20 tests of an established practice reversed it. In 2001, 14 of 48 did. In 2009, 16 of 35 did. Over the whole decade, 146 of 363 did. A colleague takes 2006 alone as her base rate. Give all four as per cents and say which you would use.

7. Here is a worked answer. Find the step that broke, and say why it looks reasonable.

> *"A large trial has just found a new drug better than the standard treatment. Prasad found that 40.2 per cent of practices are reversed. So I start at 59.8 per cent that this benefit will hold up."*

```
1 minus 0.402 = 0.598
```

8. Here is a worked answer. Find the step that broke.

> *"Herrera-Perez and colleagues found that 53 of their 396 reversals were vitamins, supplements or food. So 13.4 per cent of supplement trials end in reversal."*

```
53 divided by 396 is about 0.1338
```

9. A newspaper column says: "Most medical advice is later overturned, a study of a decade of research has found." The study is Prasad and colleagues' count in the New England Journal of Medicine. Decide what to compute, compute it, and say what the claim does not establish.

10. A fellow resident plans a thesis that will test a new school nutrition programme in Raipur. She says: "Prasad found that 756 of 981 new practices tested were better than the standard. So I am 77 per cent sure my programme will work." Decide what to compute, compute it, and say what her number does not establish.

15 · Updating: changing your confidence by as much as the evidence warrants

DEFINITION. Updating is changing your confidence in a claim when new evidence arrives, by as much as that evidence warrants and no more.

How far one piece of evidence should move you depends on two rates: how often that evidence turns up when the claim is true, and how often it turns up when the claim is false. Weak evidence turns up about as often either way.

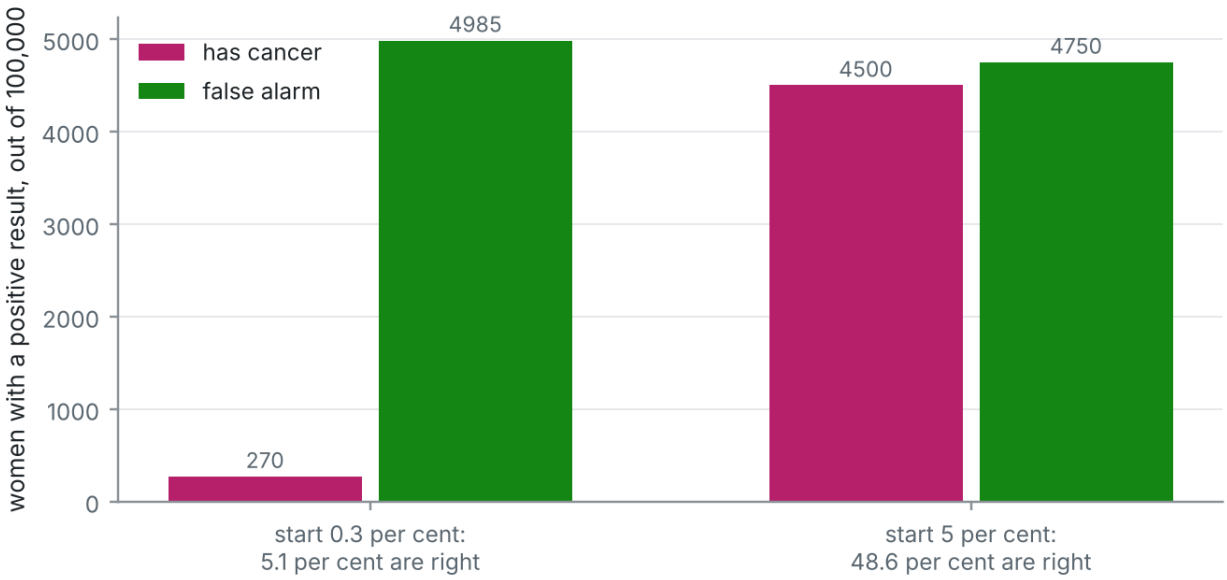
Work it in Book 0's two-way table (Book 0, D3), filled with natural frequencies: whole counts of imagined cases, out of 1,000. The columns sort the 1,000 by whether the claim is true, using the base rate. The rows sort them by whether the evidence turns up, using its two rates. Take the row of cases in which the evidence turned up. Your confidence after the evidence is the count in that row where the claim is true, divided by the row's total. When the base rate is rare, such as 1 in 1,000, imagine 100,000 cases instead, so every count stays whole. The figures here do.

For a diagnostic test, the base rate is the prevalence. The first rate is the sensitivity, and the second is 1 minus the specificity.

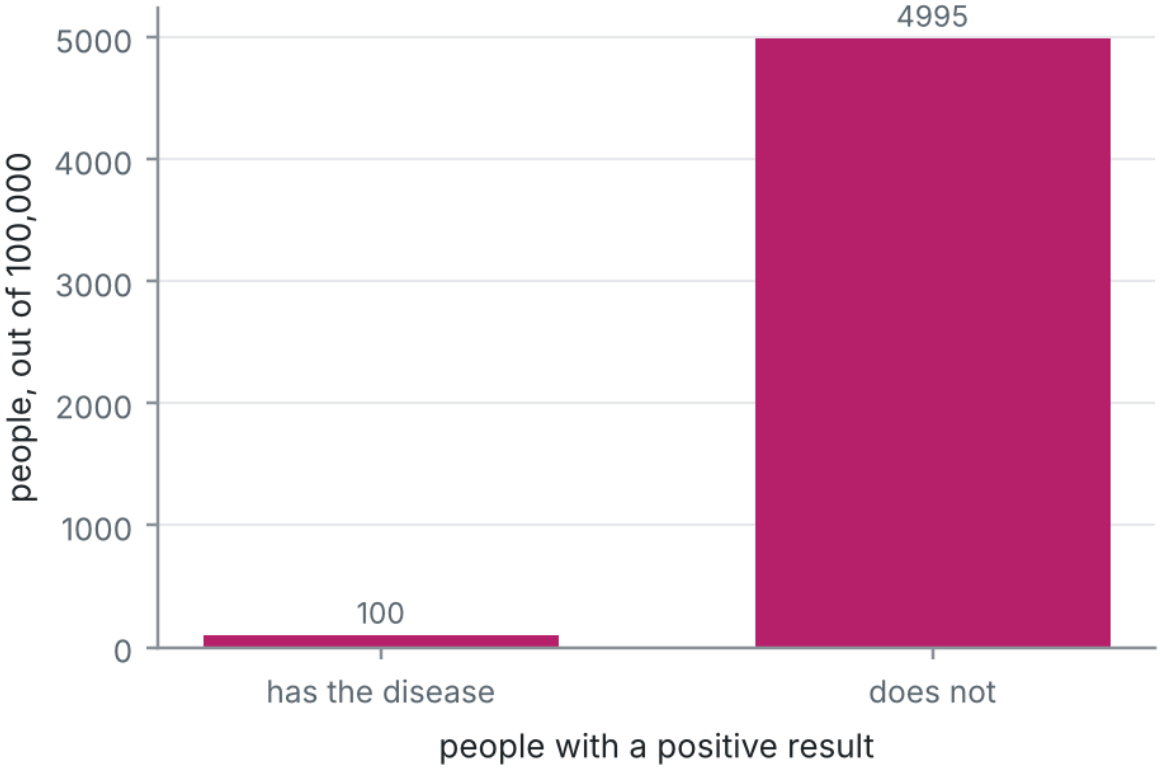
A second piece of evidence starts from the confidence the first one left you with: not from the base rate, and not from zero. [260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271]

IN PLAIN TERMS. Start from the base rate. Imagine 1,000 cases like this one (100,000 if the claim is rare).

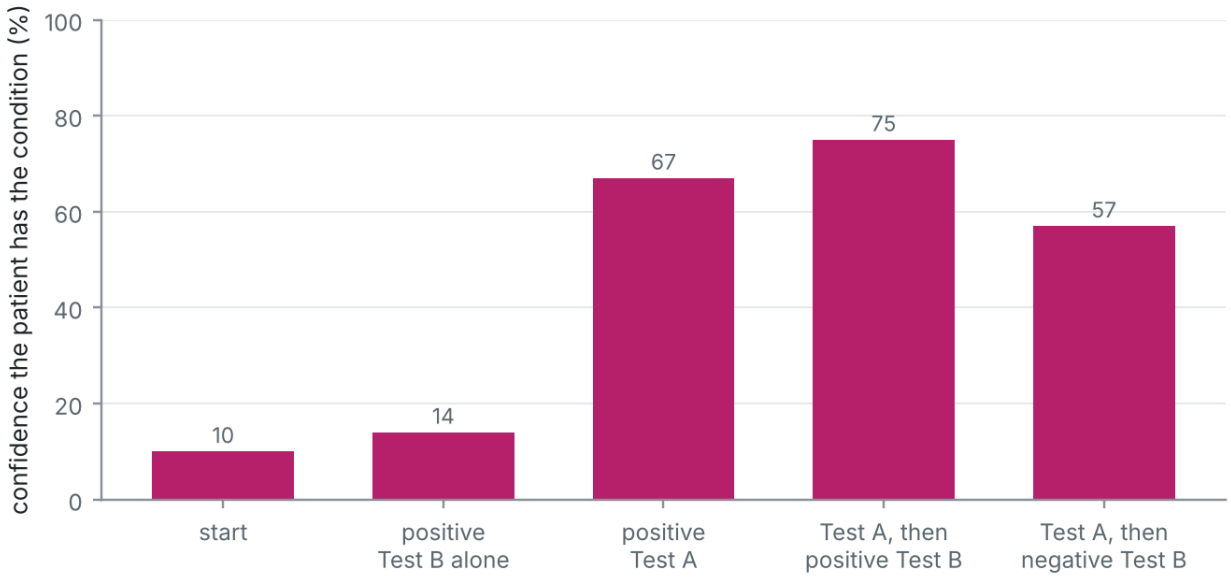
Strong evidence moves you far. Weak evidence moves you a little.



The same made-up test, run on 100,000 women from two starting points (illustration 3). From a start of 0.3 per cent, a positive result is right for 270 women and a false alarm for 4,985: about 5.1 per cent. From a start of 5 per cent, it is right for 4,500 and a false alarm for 4,750: about 48.6 per cent. Same test, same arithmetic; the start decides the end.



Morgan's question worked on 100,000 people (illustration 1). The positive row holds 100 people with the disease and 4,995 false alarms: 100 of 5,095 is about 2 per cent. The practitioners' median answer, 95 per cent, was the specificity. The false alarms win because people without the disease outnumber those with it almost a thousand to one.



Updating with made-up round numbers (illustration 2). From a start of 10 per cent, a positive weak Test B moves you only to 14 per cent; a positive strong Test A moves you to 67. From 67, a second, weak result moves you a little either way: to 75 if Test B is positive, to 57 if it is negative. It does not undo the strong one.

Illustration 1. This section stands on the last one, where the base rate gave you a place to start. Here you move from it when evidence arrives. Start with a question that most practitioners got wrong when it was put to them.

Morgan and colleagues put it to 553 practitioners in primary care in the United States. Their paper is at https://pmc.ncbi.nlm.nih.gov/articles/PMC8022260/. Search it for "1 out of 1000".

> *A test to detect a disease for which prevalence is 1 out of 1000 has a sensitivity of 100% and specificity of 95%. What is the chance that a person found to have a positive result actually has the disease?*

Write your own answer down before you read on.

The median answer was 95 per cent, and at least three in four answered 95 or more. The correct answer is about 2 per cent. Here is how to see it without any formula.

Imagine people, not percentages. The disease is rare, so imagine 100,000 people rather than 1,000; then every count stays a whole person. One in 1,000 have the disease.

```
100000 divided by 1000 = 100
100000 minus 100 = 99900
```

So 100 people have the disease and 99,900 do not. These are the two columns of Book 0's two-way table (Book 0, D3). Now the rows. A sensitivity of 100 per cent means the test catches all 100 people with the disease. A specificity of 95 per cent means it clears 95 in 100 of the people without it, so it wrongly flags the other 5 in 100.

```
100 times 1 = 100
99900 times 0.05 = 4995
99900 minus 4995 = 94905
100 plus 4995 = 5095
```

	has the disease	does not	total
positive result	100	4,995	5,095
negative result	0	94,905	94,905
total	100	99,900	100,000

Now the question asked. Take the row of positive results, 5,095 people. Of them, 100 have the disease.

```
100 divided by 5095 is about 0.0196
```

About 2 per cent. Of every 51 people with a positive result, about 1 has the disease and about 50 are false alarms. The false alarms win because the people without the disease outnumber the people with it by almost a thousand to one. Even a small false-alarm rate, applied to that many people, swamps the few true cases.

Where did 95 per cent come from? It is the specificity, or 100 minus the false-alarm rate. It answers a different question: of people without the disease, how many test negative? The answer reached for a number in the question. It skipped the base rate.

Figures quoted in this illustration are from [272].

WHERE THIS PICTURE BREAKS. Here the test's two rates were given exactly. For a real test, they are measured in some group, and carrying them to your patients is an assumption, as Book 0 warned. For a study's finding about a claim, the two rates are rarely known at all. The next illustration shows what to do with them when you have them.

The practitioners in this survey worked in the United States. Nobody has asked Indian doctors this question in the sources this book holds.

Illustration 2. Now see how far a result should move you. The numbers in this illustration are made up and round, so that the arithmetic shows. Your starting confidence that a patient has a condition is 10 per cent. Two tests are on offer.

- Test A turns positive in 90 of 100 people with the condition, and in 5 of 100 without it.
- Test B turns positive in 60 of 100 people with the condition, and in 40 of 100 without it.

Imagine 1,000 patients like this one. At 10 per cent, 100 have the condition and 900 do not. Now take each test's two rates.

```
100 times 0.9 = 90
900 times 0.05 = 45
90 divided by (90 plus 45) is about 0.6667
100 times 0.6 = 60
900 times 0.4 = 360
60 divided by (60 plus 360) is about 0.1429
```

A positive Test A moves you from 10 per cent to 67 per cent. A positive Test B moves you to 14 per cent. The difference lies in how much more often each test turns positive with the condition than without it.

```
90 divided by 5 = 18
60 divided by 40 = 1.5
```

Test A turns positive 18 times as often in people with the condition. Test B does so only 1.5 times as often. That ratio is the strength of the evidence. Ask for both rates before you react to any result.

Now a second result. Suppose Test A came back positive, so you stand at 67 per cent. You do not go back to 1,000 patients. You keep only the 135 with a positive Test A: 90 with the condition, 45 without. Then you run Test B on them.

```
90 times 0.6 = 54
45 times 0.4 = 18
54 divided by (54 plus 18) = 0.75
90 times 0.4 = 36
45 times 0.6 = 27
36 divided by (36 plus 27) is about 0.5714
```

A positive Test B moves you from 67 to 75 per cent. A negative Test B moves you down to 57 per cent. The weak result moves you a little, either way. It does not undo the strong one. Updating is not swinging to the newest result: the newest result counts for as much as its strength, and no more.

Run it backwards too. At 67 per cent you hold 90 with the condition for 45 without, 2 to 1. To end at 75 per cent you need 3 to 1.

```
90 divided by 45 = 2
100 minus 75 = 25
75 divided by 25 = 3
3 divided by 2 = 1.5
```

To turn 2 to 1 into 3 to 1, the second result must turn up 1.5 times as often with the condition as without it. That is Test B.

WHERE THIS PICTURE BREAKS. The second step treats Test B's rates as the same among the 135 people as in everyone. That holds only if the two tests do not share a cause of error. Two results

from the same sample, the same laboratory or the same flawed method are closer to one piece of evidence than two. Updating twice on them moves you too far.

For a study about a claim, you seldom know the two rates. Then judge them, keep them round, and say that they are judgements: the table works the same way, and a reader can change your rates and redo it. Or judge in words whether the result would turn up far more often if the claim were true, and move by a little or a lot. Do not print a precise rate you cannot back.

Illustration 3. Morgan and colleagues measured two faults in practitioners. They started too high, and they moved too far. Here is the first fault in one of their four cases.

The scenario: a 45-year-old woman comes for an annual visit with no risk factors for breast cancer. The study's expert panel put her chance of breast cancer before a mammogram at 0.2 to 0.3 per cent. The practitioners' median starting estimate was 5 per cent. After a positive mammogram, the panel's range was 3 to 9 per cent; the practitioners' median was 50 per cent.

To see what the starting point alone does, take one made-up test and run it from both starts. It catches 90 of 100 cancers and wrongly flags 5 of 100 women without cancer. These round numbers are chosen for the arithmetic; they are not mammography's measured figures. Imagine 100,000 women.

```
100000 times 0.003 = 300
300 times 0.9 = 270
100000 minus 300 = 99700
99700 times 0.05 = 4985
270 divided by (270 plus 4985) is about 0.0514
100000 times 0.05 = 5000
5000 times 0.9 = 4500
100000 minus 5000 = 95000
95000 times 0.05 = 4750
4500 divided by (4500 plus 4750) is about 0.4865
```

From a start of 0.3 per cent, a positive result takes you to about 5.1 per cent, inside the panel's range. From a start of 5 per cent, the same result takes you to about 48.6 per cent, close to the practitioners' 50. Same test, same arithmetic, and a tenfold difference at the end. It all came from the start.

The paper found both faults. The overestimates were "mostly associated with overestimates of pretest probability", their name for the starting confidence. For a positive exercise stress test (a heart tracing taken while the patient walks or runs on a treadmill) and a positive or negative mammogram, their answers after the test were also far too high. For the positive results, the paper's own measure shows they moved too far. So write your starting number down before you see the result, and say where it came from.

Figures quoted in this illustration are from [272].

WHERE THIS PICTURE BREAKS. The test here is made up, so the 5.1 and 48.6 per cent show the effect of the start, not what a real mammogram does. The panel's ranges are its own judgement from the literature, for one described patient. They are not a measured rate in any Indian clinic.

MUST KNOW POINTS FOR YOU.

- "The test catches 90 in 100 cases, so a positive result means 90 per cent" is the error. Fill in the two-way table before you say a number.
- Morgan and colleagues asked 553 primary care practitioners in the United States about a disease in 1 of 1,000 people. Their median answer for a positive result was 95 per cent. The correct answer is about 2 per cent.
- Write your starting confidence down before you see a result, and say where it came from.
- Evidence that turns up almost as often either way should barely move you, however striking it looks.
- A new result moves you from where you stood. It does not replace where you stood.
- The table needs the evidence's two rates. Do not print an exact number you cannot back.
- Two results that share a sample, a laboratory or a flaw are closer to one piece of evidence than two.
- Counts may teach this better than formulas (Morgan and colleagues suggest it; they did not test it): "imagine 1,000 patients like her".

COMMON MISREADING. "Updating means going with the latest study." That is swinging, not updating. A new result moves you from where you stood, by as much as its strength allows. A weak result against a well-founded view moves you a little. A strong result moves you far, but still from your starting point.

The opposite error is holding the starting number fixed whatever comes in. Both errors skip the same step. Ask how much more often this evidence turns up when the claim is true than when it is false.

REPORTING IT. Before this result I put the chance at about 10 per cent, from [the base rate and its source]. This kind of result turns up about 90 times in 100 when the condition is present, and 5 times in 100 when it is not. So after a positive result I put it at about two in three, not at 100 per cent.

EXERCISE 1 (retrieval). From memory: set out the steps for updating a confidence with a two-way table out of 1,000. Then say what makes one piece of evidence strong and another weak.

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (critique). A senior resident says at a journal club: "We used to think this practice worked. A trial last month found no benefit, so now we know it does not. Science updates." Find what is right in this, and the specific flaw.

EXERCISE 3 (teaching). A first-year resident asks you: "If a test is 95 per cent specific, why can most of its positive results be wrong?" You have ten minutes and a whiteboard. Give your answer.

PRACTICE. Ten problems on this technique, easiest first. Work them on paper before turning to the answers, which are in the appendix at the back under this section's name.

1. Out of 1,000 cases, the base rate is 20 per cent. The evidence turns up in 80 of 100 cases where the claim is true and in 10 of 100 where it is false. Fill in the two-way table: claim true or false across the top, evidence or no evidence down the side, with every total.

2. From the table of the last problem, give your confidence that the claim is true after the evidence turns up, and after it does not.

3. Here is a two-way table of 1,000 cases.

	claim true	claim false
evidence present	45	95
evidence absent	5	855

Give the base rate. Give the share of true claims that show the evidence. Then give the share of cases with the evidence in which the claim is true.

4. Morgan and colleagues set practitioners this question.

> *A test to detect a disease for which prevalence is 1 out of 1000 has a sensitivity of 100% and specificity of 95%.*

What is the chance that a person with a positive result has the disease? And a person with a negative result? Work it out of 100,000 people. Then compare with the practitioners' median answers, 95 per cent and 5 per cent.

5. In Morgan and colleagues' heart case, a 43-year-old woman has atypical chest pain and no risk factors. Cardiac ischaemia is a reduced blood supply to the heart muscle. The study's expert panel put her chance of it at 1 to 4.4 per cent before an exercise stress test. The practitioners' median was 10 per cent.

Suppose the test turns positive in 70 of 100 people with ischaemia and in 30 of 100 without; these are round numbers for the arithmetic. Work out of 1,000. Find the chance of ischaemia after a positive test from a start of 4 per cent. Then find it from a start of 10 per cent. Compare with the practitioners' median after a positive test, 70 per cent, and the panel's range, 2 to 11 per cent.

6. Your starting confidence in a claim is 30 per cent. A first piece of evidence turns up. It appears in 80 of 100 cases where such a claim is true, and 20 of 100 where it is false. Then a second, separ-

ate piece of evidence fails to turn up. It would have appeared in 60 of 100 true cases and 30 of 100 false ones. Work out of 1,000. Give your confidence after each step.

7. Here is a worked answer. Find the step that broke, and say why it looks reasonable.

> *"Two per cent of patients like this one have the condition. The test turns positive in 90 of 100 people with the condition and in 10 of 100 without it. She tested positive. The test catches 90 per cent, so she is 90 per cent likely to have it."*

8. The question: "The base rate is 30 per cent. The test turns positive in 90 of 100 people with the condition and in 10 of 100 without. What is the chance of the condition after a positive result?" Here is a worked answer. Find the step that broke.

```
1000 times 0.3 = 300
300 times 0.9 = 270
1000 times 0.1 = 100
270 divided by (270 plus 100) is about 0.7297
```

9. On a ward round, a colleague talks about a 45-year-old woman. She has no risk factors and a positive screening mammogram. He says: "The mammogram is positive, so it is certain she has cancer." Morgan and colleagues' expert panel put the chance for such a woman at 0.2 to 0.3 per cent before testing. After a positive result, they put it at 3 to 9 per cent. Decide what to compute, compute it, and say what your answer does not establish.

10. At a journal club a colleague says: "I used to be about 30 per cent sure that skipping breakfast causes weight gain. Then I read an observational study that found skippers more often had overweight. Now I am 90 per cent sure." Decide what to compute, compute it, and say what your answer does not establish.

16 · Confidence is not accuracy: what being calibrated means

DEFINITION. A person is calibrated at a level of confidence when, of all the claims they have stated at that level, the proportion that turn out true is about that level. Of all the claims they gave 80 per cent, about 80 in 100 turn out true. Calibration is a property of many claims checked over time, never of one claim, because one outcome cannot show a single degree of belief wrong (section 10). The word has another use: a "calibrated scale" (section 12) is a table that fixes each word's numbers. That is a property of a scale; this section's sense is a property of a person.

Calibration is checked with a tally. Each claim is written before its outcome is known, with its date and a degree of belief, on the side the person leans to. A claim held at 30 per cent is written as its opposite held at 70 per cent, the complement (section 10), so every stated level is 50 per cent or more. A claim at exactly 50 per cent leans neither way: it is written as first phrased. When the outcomes are known, the claims are grouped by stated level. For each level, the number of claims and the number that turned out true are counted, and the second is divided by the first: the proportion true. That proportion is compared with the level.

At a level above 50 per cent, a proportion true below the level is over-confidence: the person was right less often than they said. A proportion true above the level is under-confidence. At 50 per cent, any proportion true far from 50, high or low, is under-confidence: the person had a lean and did not state it.

Confidence and accuracy are measured separately. How sure a person feels is evidence about the person, not about the claim.

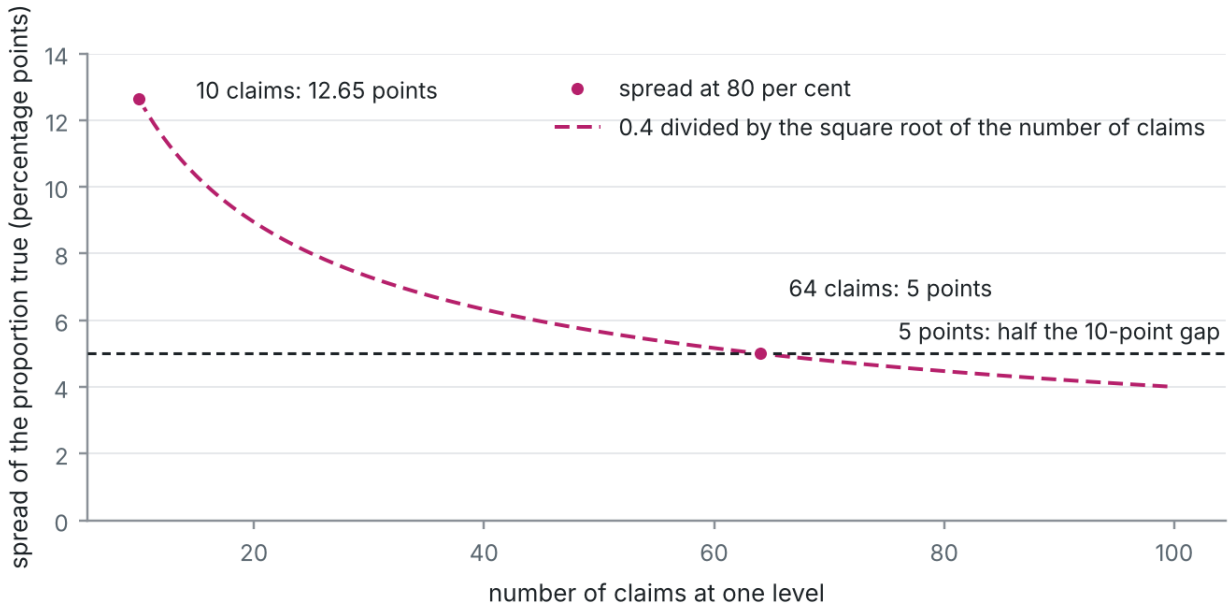
A tally is itself a sample, and chance moves it (Book 0, D6). Take a person calibrated at one level. Over a number of claims at that level, their proportion true has a spread. To find it, divide the square root of the level times (1 minus the level) by the square root of the number of claims. Ten claims at 80 per cent scatter by about 13 percentage points, so they cannot tell 80 per cent from 70. [273, 274, 275, 276, 277, 278, 279, 280, 281]

IN PLAIN TERMS. Every time you say "I am 80 per cent sure", you take a reading. One reading cannot be checked. So write them down, with the date, before you know the answer. Later, gather all the claims you gave 80 per cent and count how many came true. If about 8 in 10 did, your 80 means what it says. If only 6 in 10 did, you are more sure than you have reason to be.

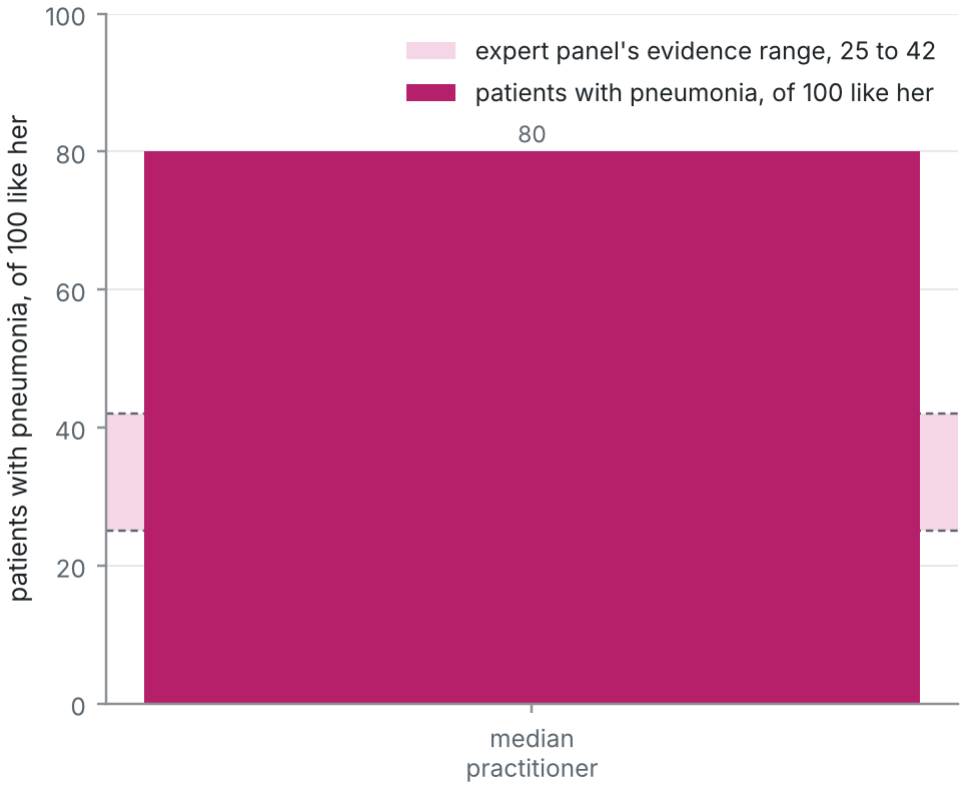
Feeling sure is not the same as being right.



The made-up resident's tally from illustration 1, below. Each point is one level she wrote, against the proportion of those claims that turned out true. The straight line is perfect calibration: proportion true equals stated level. At 60 per cent she sits near the line, at 62.5. At 90 per cent she is far below it, at 60. Her proportion true is about 60 per cent whatever level she wrote.



How far chance alone moves the tally of a person calibrated at 80 per cent, against the number of claims at that level. The spread is 0.4 divided by the square root of the number of claims. At 10 claims it is 12.65 percentage points, more than the whole 10-point gap between 80 and 70 per cent. It comes down to 5 points, half that gap, only at 64 claims.



Morgan's pneumonia scenario (illustration 3), read as a tally of 100 patients like her. Before any test, the median practitioner expected 80 to have pneumonia; the expert panel, from the published evidence, put it at 25 to 42. The estimate was 38 to 55 percentage points too high.

Illustration 1. Here is the trap with feeling sure. A second-year resident says she is "usually right" on ward rounds. She is made up for this section, and so are all her numbers. She means it, and she may be right. But she cannot say how often, because she has never written anything down.

So she starts. For three months, before each test result comes back, she writes one line in a notebook: the date, the claim, and how sure she is as a per cent.

> *12 March. Bed 7's urine culture will grow bacteria. 80%.*

When the result comes back, she ticks the line or crosses it out.

Some lines lean the other way.

> *14 March. Bed 3's chest X-ray will show pneumonia. 30%.*

Before counting, she rewrites each of these on the side she leans to, using the complement from the section on putting a number on one claim: "Bed 3's chest X-ray will not show pneumonia. 70%." Now every line is 50 per cent or more. A tick means she was right the way she leaned.

After three months she has 40 lines. She sorts them by the level she wrote, and counts.

stated level (%)	claims	turned out true	proportion true (%)
60	8	5	62.5
70	10	6	60
80	12	7	58.3
90	10	6	60

Work one row yourself. At 80 per cent she wrote 12 claims and 7 came true. Divide, then turn the fraction into a per cent (Book 0, A4).

```
7 divided by 12 = 0.583333333
0.5833 times 100 = 58.33
58.33 minus 80 = -21.67
```

She was right 58.3 per cent of the time when she said 80: about 22 percentage points below.

Now read the whole table. The proportion true sits near 60 per cent at every level. When she wrote 60, she was right about 6 times in 10. When she wrote 90, she was still right about 6 times in 10. Her 90 told her nothing more than her 60.

Now the overall count, and the average level she stated.

```
5 plus 6 plus 7 plus 6 = 24
24 divided by 40 = 0.6
8 times 60 = 480
10 times 70 = 700
12 times 80 = 960
10 times 90 = 900
480 plus 700 plus 960 plus 900 = 3040
3040 divided by 40 = 76
```

“Usually right” was true: 24 of 40, or 60 per cent. But the levels she wrote averaged 76 per cent. At her higher levels she is over-confident. Her feeling of being sure did not show her this. Only the notebook did.

Keep one doubt for the next illustration. Ten or twelve claims at one level is a small tally. Is a gap of 10 points at 70 per cent real?

WHERE THIS PICTURE BREAKS. The resident and her 40 claims are made up. A real notebook is rarely this tidy: some results never come back, and those lines can be neither ticked nor crossed. Leave them out and say how many you left out.

Her claims are all about test results on one ward. A tally of one kind of claim tells her about that kind only. It says nothing about how well calibrated she is about prognosis, or about which trial will show an effect.

Illustration 2. Take the smallest case. Suppose you are calibrated at 80 per cent: over a long run, 8 in 10 of your 80 per cent claims come true. You write ten such claims. How many come true?

Not always 8. Seven is no surprise, nor is nine. Two things make ten claims a weak test.

First, each claim is a big step. With ten claims, one claim moves the proportion true by 10 percentage points.

```
1 divided by 10 = 0.1
```

The whole gap between 80 and 70 per cent is one claim.

Second, chance scatters a tally the way it scatters the average of a sample. Book 0's square-root law gives the size (Book 0, D6). Score each claim 1 if it came true and 0 if it did not. The proportion true is then the mean of those scores.

Find the spread of one score first. Over a long run at a calibrated 80 per cent, 8 in 10 scores are 1 and 2 in 10 are 0, so take ten scores in that ratio. Their mean is 0.8. Work out the standard deviation as Book 0 did (Book 0, D5), dividing by the count, because these ten stand for the whole long run.

```
1 minus 0.8 = 0.2
0 minus 0.8 = -0.8
0.2 times 0.2 = 0.04
-0.8 times -0.8 = 0.64
8 times 0.04 = 0.32
2 times 0.64 = 1.28
0.32 plus 1.28 = 1.6
1.6 divided by 10 = 0.16
0.4 times 0.4 = 0.16
```

The variance is 0.16, so the standard deviation is its square root, 0.4. Notice that 0.16 is 0.8 times 0.2: the level times (1 minus the level). The same steps at any level give the same pattern. At 90 per cent, nine scores of 1 and one of 0 give 0.09, which is 0.9 times 0.1, and a standard deviation of 0.3.

Now the square-root law. The spread of the proportion true over a number of claims is the spread of one score divided by the square root of the number of claims.

```
3.162 times 3.162 = 9.998
0.4 divided by 3.162 = 0.1265
```

That is 12.65 percentage points, about 13. So a calibrated 80 per cent person's tally of ten claims commonly lands anywhere from about 67 to 93 per cent. Seven out of ten is ordinary for them. It is just as ordinary for a person whose true level is 70. Ten claims cannot tell the two apart.

How many could? Use a rough rule: the spread should be no more than half the gap you want to see. Chance alone often moves a tally by one spread and seldom by two; exactly how seldom is beyond this book. To see a 10-point gap, the spread must come down to 5 points, which is 0.05. Use the spread at the level you stated, because the question is whether your tally fits that level.

```
0.4 divided by 0.05 = 8
8 times 8 = 64
```

The square root of the number of claims must be 8, so the number of claims must be about 64, all at that one level. That is many months of a notebook.

Go back to the resident. At 90 per cent she wrote 10 claims.

```
0.3 divided by 3.162 = 0.0949
30 divided by 9.49 = 3.16
```

The spread is about 9.5 points. Her gap there, 30 points, is more than three spreads. Chance alone seldom does that: her 90 really is too high. Now her row at 70 per cent.

```
0.7 times 0.3 = 0.21
0.4583 times 0.4583 = 0.21
0.4583 divided by 3.162 = 0.1449
```

The spread is about 14.5 points, larger than her gap of 10. That row tells her nothing yet.

WHERE THIS PICTURE BREAKS. The square-root law treats each claim as a separate chance, like a member of a random sample. Claims that share a cause rise and fall together: ten predictions about one trial, or ten culture results from one laboratory with one faulty method. Then the tally scatters more than the rule says, and you need more claims than it says.

“Half the gap” is a rough rule for deciding whether to look harder, not a test. Scoring a whole list of claims in one number is beyond this book.

Illustration 3. A related error can be measured in groups: estimates set against the published evidence rather than against outcomes. It is not calibration as defined above, but it shows the same direction of error. Morgan and colleagues (2021) gave written scenarios to 723 practitioners at primary care clinics in eight US states, and 553 answered every question. Their paper is at https://pmc.ncbi.nlm.nih.gov/articles/PMC8022260/.

One scenario describes a previously healthy 35-year-old woman who smokes. She has had five days of fatigue and a productive cough, a temperature of 38.9 degrees Celsius, and fewer breath sounds at the lower right of the chest. How likely is pneumonia, before any test? The median answer was 80 per cent. An expert panel, working from the published evidence, put it at 25 to 42 per cent.

Read it as a tally. Of 100 patients like her, the median practitioner expects 80 to have pneumonia. The evidence says 25 to 42.

```
80 minus 42 = 38
80 minus 25 = 55
```

The median estimate was 38 to 55 percentage points too high. Before testing, estimates were too high in every scenario. The authors sum it up as 2 to 10 times the evidence, before and after testing. One participant told them that estimating the probability of disease “isn’t how you do medicine.”

That remark is the whole problem of this section. If you never write the number, nobody can check it, including you.

Figures quoted in this illustration are from [272].

WHERE THIS PICTURE BREAKS. These were written scenarios, not the practitioners' own patients. The evidence range was set by the study's expert panel from the literature, and the authors say the survey's reliability was not assessed. More than half the respondents were residents, in US clinics. No study held for this book has measured how Indian clinicians' estimates compare with the evidence.

The study shows over-estimates of disease, in one direction, for one kind of claim. It does not show that every clinician is over-confident about everything.

MUST KNOW POINTS FOR YOU.

- "I am usually right, so my confidence is about right" is the error. Being right often and being calibrated are different things.
- Write the number before the answer is known, with the date, on the side you lean to. A number remembered after the result is not a reading.
- One overall proportion hides miscalibration. Tally level by level.
- Before you act on your sense of how likely a diagnosis is, ask how common it is in patients like this one.
- A small tally says little. Ten claims at 80 per cent scatter by about 13 percentage points by chance alone, so they cannot tell 80 from 70. To see a gap of 10 points you need roughly 64. Below that, report the counts and say they cannot yet tell.
- Calibration is not the whole of good judgement. Someone who says 50 per cent on every claim can be calibrated and tell you nothing.
- When a trainee says "I knew it would be positive", ask what number they wrote down before the result. If they wrote none, the claim cannot be checked.
- An expert in a committee says they are "90 per cent sure the programme will meet its target". Ask whether they have ever counted how often their 90s came true.

COMMON MISREADING. "Being calibrated means being right." It does not. A calibrated person is wrong often, and by the amount they said. At 70 per cent, they are wrong about 3 times in 10. When one of your 70 per cent claims turns out false, that is what calibration predicts, not a sign that it failed.

The reverse error is as common. A person who is right 9 times in 10 can still be badly calibrated, if they said 60 per cent each time. Calibration asks whether the numbers match the record. It does not ask how often you were right.

REPORTING IT. Between [first date] and [last date] I wrote [number] claims at [level] per cent, and [count] turned out true ([proportion] per cent). With that many claims, chance alone moves

the proportion by about [spread] percentage points. So the gap of [gap] points [does / does not yet] show that I am over-confident at this level.

EXERCISE 1 (retrieval). From memory: say what it means to be calibrated at 80 per cent. Describe the tally in four steps, including what you do with a claim stated at 30 per cent. Then say why ten claims cannot show that you are over-confident at 80 per cent.

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (interpretation). A colleague hands you her tally, made up for this exercise, and asks what it shows.

stated level (%)	claims	turned out true
60	20	15
75	16	12
95	20	15

Say, level by level, what the tally does and does not show.

EXERCISE 3 (teaching). For a first-year resident, ten minutes with a whiteboard. She says: "When I am sure, I am right. I can feel the difference." Teach her the difference between confidence and accuracy, and leave her with something to do.

EXERCISE 4 (design). Design the tally sheet you will use for your own dated claims. Say what each column holds and the rule for claims under 50 per cent. Then say when you will tally, and how many claims you need at a level before you act on it.

PRACTICE. Ten problems on this technique, easiest first. Work them on paper before turning to the answers, which are in the appendix at the back under this section's name.

1. At one level, 20 claims were stated at 70 per cent and 13 turned out true. At another, 10 claims were stated at 90 per cent and 7 turned out true. For each level, give the proportion true as a per cent and the gap from the stated level in percentage points. Say whether each gap points to over-confidence or under-confidence.

2. Here are four claims, each with its stated degree of belief. Claim A is at 30 per cent, B at 10 per cent, C at 55 per cent and D at 50 per cent. A turned out false, B true, C true and D false. Rewrite each claim on the side leaned to, give its level, and say whether the rewritten claim turned out true.

3. For a person calibrated at 90 per cent, find the spread of the proportion true over 25 claims and over 100 claims, in percentage points.

4. A 2021 survey gave US primary care practitioners a written case of a 35-year-old woman with cough and fever. Before any test, the median answer from resident doctors put her chance of pneumonia at 80 per cent. The expert panel's range from the published evidence was 25 to 42 per cent. For 50 patients like her, how many patients would the resident expect to have pneumonia, and how many does the evidence range give? Give the gap in percentage points.

5. You keep a register: a dated list of your claims and their outcomes. It holds your claims at journal club, and many of them are at 90 per cent. Roughly how many claims at that level do you

need? The tally should be able to tell a true 90 per cent from a true 80 per cent. Give the answer as a number of claims.

6. A colleague's list of dated claims is made up for this problem. It has 10 claims at 50 per cent, of which 8 turned out true. It has 10 claims at 90 per cent, of which 6 turned out true. She writes: "20 claims, 14 true, 70 per cent. My average stated level was 70 per cent, so I am calibrated." Work out the tally level by level, give the spread at each level, and say what the tally shows.

7. Here is a worked answer. Find the step that broke, and say why it looks reasonable.

```
I wrote 10 claims at 30 per cent; 4 came true
4 divided by 10 = 0.4, which is 40 per cent
40 is above my 30, so at this level I am under-confident
so I can afford to be bolder with my low numbers: say 20 per cent instead of 30
```

8. Here is a worked answer. Find the step that broke.

```
I wrote 10 claims at 80 per cent; 7 came true
7 divided by 10 = 0.7, which is 70 per cent
70 minus 80 = -10, so I am over-confident by 10 points
from now on I will write 70 where I used to write 80
```

9. On a ward round, a senior colleague says: "When I tell a family I am 90 per cent sure, I am right. In thirty years I can hardly remember being wrong." Decide what to compute, compute it, and say what your answer does not establish.

10. A newspaper health column says: "A study found doctors overestimate disease 2 to 10 times. So when your doctor gives you a chance, halve it." Decide what to compute, compute it, and say what your answer does not establish.

17 · Changing a stated position out loud

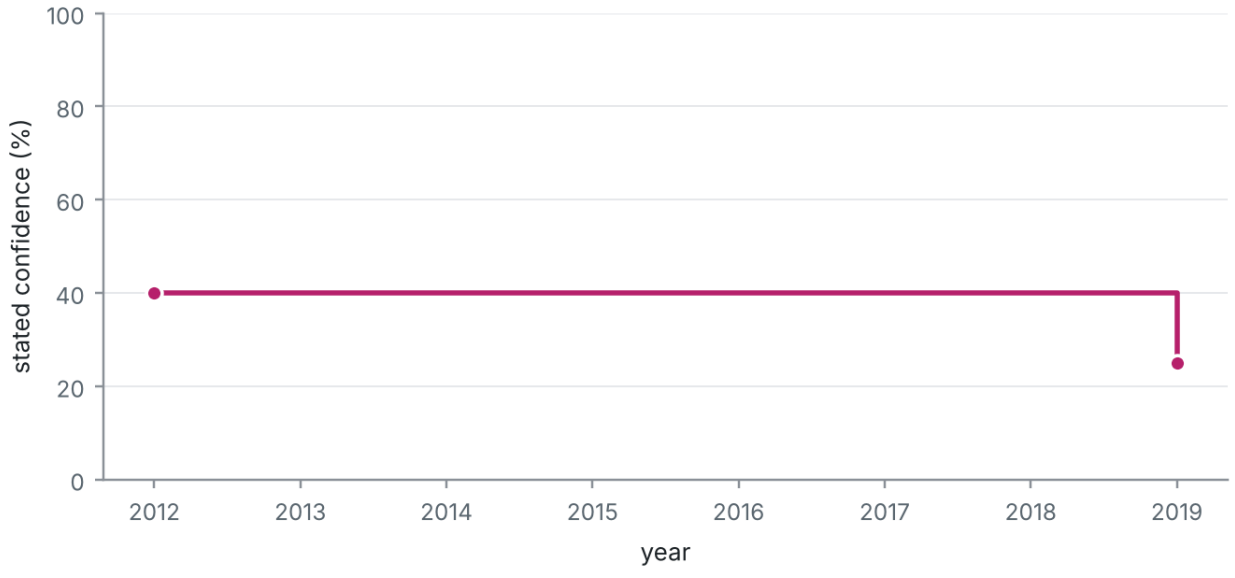
DEFINITION. Changing a stated position out loud means telling the people who heard your old position that you now hold a different one, and why. It takes four sentences, in this order.

1. What I said, to whom, and when.
2. What new evidence has appeared since.
3. What I now think, how confident I am, and on what grounds.
4. What I will do differently.

Two rules go with them. First, say whether the old position was reasonable on the evidence of its time, and name separately any part of it that was not reasonable even then. Second, move as far as the new evidence warrants (section 15) and no further.

The four sentences are an argument laid out in time (Book 0, F4). Sentence 1 gives the old conclusion. Sentence 2 adds a premise the old argument did not have. Sentence 3 gives the new conclusion, with its grounds, so the listener can check the inference. Sentence 4 gives the action that follows from it. A position changed in silence leaves the old conclusion with everyone who heard it. [282, 283, 284, 285, 286, 3, 287, 53, 288, 289, 290, 291, 291]

IN PLAIN TERMS. Suppose you taught something, and the evidence has since moved. Changing your slides is not enough. The people you taught still carry the old version.



The made-up lecturer's confidence that eating breakfast helps adults lose weight, as she would have written it at each date. Her feeling in 2012 might have been 90 per cent, but a fair start was about 40: the 38 per cent of tested practices that trials upheld, rounded. After the 2019 review, which pooled the 2014 trial with the others, about 25 per cent, which is 75 per cent that it does not help. One update, dated, and neither end reaches 0 or 100.

Illustration 1. Here is a case of the kind this book has followed from its early sections. A lecturer in community medicine taught her postgraduate batch in 2012. She is made up for this section, and so are her slides and her numbers. Her slide said:

> *Skipping breakfast leads to weight gain. Advise every patient who is trying to lose weight to eat breakfast.*

In 2019 she reads the review by Sievert and colleagues and decides the slide is wrong. Here are three ways she could respond, all of them common. Read each and say what is wrong with it before you read on.

> *Version A. (She deletes the slide and says nothing.)*

> *Version B. "I owe you an apology. I taught you nonsense about breakfast, and I should never have done it."*

> *Version C. "As I always said, breakfast is no magic bullet. In fact the 2019 review shows that skipping breakfast is better for weight loss."*

Version A leaves the old advice with every batch that heard it. They will repeat it to patients for years.

Version B is self-abasement. It gives the batch no new fact, no new position and no reason. It also calls the old view nonsense, which it was not.

Version C does two things wrong. "As I always said" rewrites what she said. "Skipping is better" swings to a new claim that the evidence does not carry, as you will see below.

Now build the statement she should make, one sentence at a time.

1. What I said, and when. "In my nutrition lecture to the 2012 batch, I said that skipping breakfast leads to weight gain. I told you to advise every patient trying to lose weight to eat breakfast, and I said it as a fact."
2. What new evidence has appeared. "Since then, two kinds of evidence have come out. In 2014, a trial in the United States and Denmark randomly assigned 309 adults who were trying to lose weight to advice to eat breakfast, advice to skip it, or general advice. Over 16 weeks the advice changed what they ate in the morning and made no detectable difference to their weight. In 2019, a review in the BMJ (British Medical Journal) pooled seven trials that measured weight, that one among them. People assigned to eat breakfast ended up 0.44 kg heavier on average (95 per cent confidence interval 0.07 to 0.82 kg) than people assigned to skip it. The trials were short, all in high-income countries, and the reviewers judged their quality mostly low."

3. What I now think, how confident, and why. “I think it”likely” (about 75 per cent, on the ICD 203 scale) that eating breakfast does not help adults lose weight. ICD 203 is the US Intelligence Community Directive 203; on it, “likely” runs from 55 to 80 per cent. My grounds are trials that point the same way, but low-certainty ones: few, short, at risk of bias, and none done in India.”
4. What I will do differently. “I have taken the advice off my slides. When a patient trying to lose weight asks about breakfast, I now say that eating it or skipping it can follow what suits them. I will look again when longer trials report.”

Then the first rule: was the old view reasonable? Partly. In 2012 many observational studies agreed that people who skip breakfast more often have overweight or obesity. Public health authorities were commonly recommending breakfast to reduce obesity, as the 2014 trial’s authors put it, and no trial had yet reported: the one registered trial was published in 2014. Teaching that association, and the advice that went with it, was reasonable on the evidence of its time. Teaching it as a cause was not, even then, because observational studies show an association only. Brown and colleagues’ review of those same studies, published in 2013, called the causal claim “only presumed true” (the section on belief beyond the evidence). So she adds one sentence:

> *The studies I relied on showed an association, and I taught it as a cause. That part I should have taught differently even in 2012.*

Now the second rule: how far to move. The review’s 0.44 kg favours skipping. Does she now teach “skip breakfast to lose weight”? No. The evidence is low certainty, the trials were a few weeks long, and the 2014 trial found no detectable difference. Moving from “eat breakfast to lose weight” to “skip breakfast to lose weight” swaps one belief beyond the evidence for another. Low-certainty evidence earns “may not help”, not a new rule.

Put numbers on her path. Had she written down, at each date, how sure she was that eating breakfast helps adults lose weight, her feeling in 2012 might have read 90 per cent. That is a feeling, not a count. A fair start was about 40 per cent (38, rounded): the share of tested established practices that trials upheld (section 14). The 2019 review pooled the 2014 trial with the others, so she updates once, on the review. Updating on the trial as well would count it twice. After the review, about 25 per cent (section 22): several trials point the same way, but at low certainty. Twenty-five per cent that it helps is 75 per cent that it does not, which is her sentence 3. Each step is dated, and neither end reaches 0 or 100.

Figures quoted in this illustration are from [33, 167, 230].

WHERE THIS PICTURE BREAKS. The lecturer and her 75 per cent are made up. Another careful reader could put the number at 70 or 80 on the same trials; what must not change is that she states one, with its grounds.

All the trials behind the case were in high-income countries, mostly short, in adults. The case says nothing about breakfast in Indian adults, in children, or about anything except weight. Her new position is about weight loss only, and she should say so if asked about breakfast in general.

Illustration 2. Researchers have written statements like these about their own published findings. Rohrer and colleagues collected 13 of them, from psychologists, in the Loss-of-Confidence Project (2021). Their paper is at https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/. Find the statement by Raphael Silberzahn and Eric Uhlmann, and match four of its sentences to the four.

Sentence 1, what we said and when:

> *In 2013, we published an article providing evidence that the meaning of a person's name might affect the person's career outcomes.*

The article was about names that sound noble, such as Kaiser or König.

Sentence 2, what new evidence appeared:

> *A postpublication reanalysis by Uri Simonsohn using an expanded version of our data set identified a curvilinear name-frequency confound in the data*

That is, another researcher, using more data, found a third factor (how common each name is) that the authors' analysis had not handled. Curvilinear: the link with name frequency bends rather than runs straight, and the authors had allowed only for a straight one.

Sentence 3, what we now think:

> *Germans with noble and nonnoble names are equally well represented in managerial positions.*

Sentence 4, what we did:

> *We subsequently coauthored a collaborative commentary (Silberzahn et al., 2014) reporting the new results.*

Now a sentence from another statement, by Christopher Chabris, about a 1992 study of chess masters. He no longer thinks his experiment supports its claims, and he says how far he has moved:

> *These claims may be true, but not because of our experiment.*

That is the second rule in one line. He has not swung to "the claims are false". He has withdrawn his evidence for them, which is as far as his re-reading of his own analysis carries him. It is the move from the section on absence of evidence, said about your own work.

Last, what a silent update costs. Rohrer and colleagues tell how a major German weekly newspaper, in January 2018, cited the 2013 name finding without question. The finding had been corrected in 2014. When the journalist was told, she published her own correction within a month. The authors had changed their position out loud, in print. It still took years to reach one reader who had relied on the old version.

Figures quoted in this illustration are from [292].

WHERE THIS PICTURE BREAKS. Rohrer's project took only statements in which the author owned an error in their own study. Most of the 13 are about mistakes, not about new evidence from elsewhere. The authors single out one, about pigeons choosing between rewards, as closest to the usual picture: new experiments led to a revised conclusion. The breakfast lecturer's case is of that kind. The four sentences fit both, but sentence 2 differs. For an error, it says what you found wrong; for new evidence, it says what appeared.

These are psychologists' statements about published papers. A change of position in a lecture or on a ward reaches a smaller audience, and the channel is different.

MUST KNOW POINTS FOR YOU.

- "Changing my mind in public will make people trust me less" is the fear to set aside. Change your position anyway, because the people you taught are acting on the old one.
- Use the four sentences, in order. Say them to the people who heard the old position.
- Changing your slides without saying so is a silent update. It protects the next batch and leaves every earlier one with the old advice.
- Move only as far as the new evidence carries you.
- Say what was reasonable then, and what was not even then, as two separate sentences. Do not abase yourself ("I taught you nonsense"), and do not rewrite your past ("as I always said").
- To a patient you advised, the four sentences fit in a minute: "Some years ago I told you to eat breakfast to help your weight. Newer trials, though few and short, suggest it does not help weight loss. So eat breakfast if it suits you, and skip it if it does not. The trials so far suggest your weight will not hang on that choice."
- If the error is in your own published data or analysis, a sentence to the class is not enough. The journal's record needs correcting too, and a later section of this book sets out how journals do that.
- Your new position is a dated belief like the old one, held at a stated confidence. Do not present it as final.

COMMON MISREADING. "Changing my position means admitting I was wrong." Not always. If the old position was reasonable on the evidence of its time, changing it is updating: the evidence moved, and you moved with it. Saying "I was wrong about that" is for a position that was not supported even when you held it. A later section of this book is about saying it.

The two are often mixed in one case, as in the breakfast lecture. The association was fair to teach; the causal claim was not. Say which part is which. Lumping them together either over-apologises for the fair part or hides the unfair one.

REPORTING IT. In [where] on [date], I said that [old position]. Since then, [new evidence, with its size and its certainty]. I now think that [new position]; I am about [number] per cent sure ("[word]" on the [named] scale), because [grounds]. From now on I will [action].

EXERCISE 1 (retrieval). From memory: give the four sentences of a change of position, in order. Then give the two rules that go with them.

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (critique). A senior resident posts this in the department's WhatsApp group. It is made up for this exercise.

> *Folks, ignore what I said about breakfast last year. Turns out skipping breakfast is better for weight loss: a BMJ meta-analysis proved it. My bad!*

Find what is wrong with it, and rewrite it.

EXERCISE 3 (teaching). A journalist rings you. "In 2012 you told our readers to eat breakfast to lose weight. Were you wrong?" You have three minutes. Give one quotable sentence, and say how sure the evidence lets you be.

EXERCISE 4 (design). Choose one thing you have taught, written or posted that the evidence has since moved on. Write your four sentences for it, add the sentence on what was reasonable then, and name who needs to hear it and where.

18 · How science corrects its record: correction, expression of concern, retraction, republication

Checked against the sources as they stood on 2026-10-10.

DEFINITION. A journal can change what it has already published with four instruments. The International Committee of Medical Journal Editors (ICMJE), a group of medical journal editors, sets them out in its Recommendations (updated January 2026), Sections III.A and III.B. The Committee on Publication Ethics (COPE), a body that advises editors and publishers, sets out the same tools in its Retraction Guidelines (version 3, August 2025).

ICMJE asks for a correction for "errors of fact that should have been recognized at the time of publication". The mistake was in the paper when it appeared. New evidence did not show it up later. It is used when the error does not change the direction or implications of the results, interpretations and conclusions. Citations should go to the most recent version.

An expression of concern is a notice an editor may publish while a question about a paper is open: while an investigation runs, or while the editor lacks the information to decide.

A retraction is published when a paper's findings and conclusions can no longer be relied on. The cause may be honest error, a naive mistake or misconduct.

Retraction with republication, which COPE calls retraction with replacement, is used when an honest error changes the results or conclusions substantially but the underlying study stands. The original is retracted and a corrected version is published beside it, with an explanation.

None of the four is for a disagreement, or for evidence that has moved on since publication. The decision is the editor's. [293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316]

IN PLAIN TERMS. A correction fixes an honest mistake that leaves the conclusion standing. A retraction withdraws a paper whose findings cannot be relied on.

An expression of concern is a warning flag. Something may be wrong, and nobody knows yet.

Retraction with republication withdraws a paper whose honest mistake changed its result, and publishes the fixed version beside it.

None of these is a punishment. A paper that newer evidence contradicts stays as it is. The newer studies are the correction.

So before you cite a paper, look at its page for any of these notices. Then cite the latest version.



A correction in action. Fang, Steen and Casadevall (2012) listed the 20 most-cited retracted articles with the reason for each: 12 for fraud and 8 for error as first published. Their 2013 correction changed one row, giving 13 and 7. The counts describe 20 papers chosen for their citations, not a rate.

Illustration 1. Start with a paper about retractions that was itself corrected. Fang, Steen and Casadevall (2012) studied all 2,047 English-language articles PubMed listed as retracted on 3 May 2012. PubMed is the index of medical journals. The paper appeared in the Proceedings of the National Academy of Sciences of the USA (PNAS). It is free to read at https://pmc.ncbi.nlm.nih.gov/articles/PMC3479492/.

Find Table 3. It lists the 20 most-cited retracted articles, with the reason each was retracted. Read the last row: Hwang, in Science, published 2005, retracted 2006, cited 234 times. The reason given is “Error”.

Now count the reasons as printed. You get 12 rows that say “Fraud” and 8 that say “Error”.

Suppose you stop there and put “8 of the 20 most-cited retracted papers were retracted for error” on a slide. You have quoted a number the authors themselves withdrew.

Here is how you would find out. Look up the paper’s record in Europe PMC, a free index of the medical literature (https://europepmc.org/article/MED/23027971). Its list of comments and corrections has an entry labelled “Erratum in”. It points to Proc Natl Acad Sci U S A. 2013 Jan 15;110(3):1137. That is the journal’s short name, the date of the issue, volume 110, issue 3, page 1137.

“Erratum” is an older word for a correction. Open it, at https://pmc.ncbi.nlm.nih.gov/articles/PMC3549137/. The whole notice is two sentences and a table:

> *The authors note that Table 3 appeared incorrectly. The corrected table appears below.*

In the corrected table, the last row is the same paper, with the same years and citations. Its reason now reads “Fraud”. Count again.

reason for retraction	as first published (2012)	as corrected (2013)

Fraud	12	13
Error	8	7

Now name the instrument. One entry in one table was wrong, the authors reported it, and the notice replaces that table. The ICMJE Recommendations call for a correction for an error of fact, as long as it does not change the direction of the conclusions. So a correction is the fitting tool, not a retraction.

Last, which table do you quote? The Recommendations say that "the citation should be to the most recent version". So if you use Table 3, take it from the 2013 correction, and say so.

Figures quoted in this illustration are from [317].

WHERE THIS PICTURE BREAKS. This correction was easy to judge: one cell in one table, reported by the authors, with the notice saying exactly what changed. Many are harder. A correction can change a number that the conclusion leans on. Then whether it "changes the direction or implications" of the conclusions is a judgement, and the editor makes it.

The counts of 12 and 13 describe 20 heavily cited papers chosen by citation count. They are not a rate of fraud among retractions. The paper's main counts, over all 2,047 articles, are the subject of the next section.

Illustration 2. Now a retracted paper from the obesity literature, and a habit that keeps it alive.

Go back to Table 3 in Fang and colleagues' paper and find Fukuhara. The paper appeared in Science in 2005 and was retracted in 2007, for error. By June 2012 it had been cited 686 times. Now search their Discussion for "visfatin":

> *Many scientists continue to cite the article by Fukuhara et al. for its initial identification of visfatin as an adipocytokine, even though the article was retracted because of concerns about the cytokine's reported insulin mimetic properties*

An adipocytokine is a signalling protein released by fat tissue. A cytokine is a signalling protein; "insulin mimetic" means acting like insulin. The sentence before says both the web page and the PDF of the article "are clearly marked as retracted". So the label was there, and people cited the paper anyway.

COPE's Retraction Guidelines give two reasons this happens. Search them for "did not verify":

> *the article was downloaded before it was retracted and the authors who cited it did not verify its status, or the authors did not notice that it was retracted*

Think of the PDFs in your thesis folder. Each one shows the paper as it was on the day you saved it. A retraction published after that day is not on your copy.

So check each reference where it lives today. Open the article's page on the journal's site and its record in PubMed or Europe PMC. Look for four things. "Retracted" in front of the title, or a "RE-

TRACTED" stamp across the pages; COPE calls both best practice. A linked notice of retraction. A linked expression of concern. A correction, which may be labelled "Erratum".

If you find a retraction and still need the paper, for its history or its method, COPE allows it. You cite it with the retraction: "Authors should not cite retracted papers without acknowledging the retraction."

Figures quoted in this illustration are from [317].

WHERE THIS PICTURE BREAKS. The four checks find only what an editor has acted on and a database has recorded. A paper can be wrong with no notice on it, and a notice can take years to appear. Fang and colleagues found that retractions for fraud took a mean of 46.8 months.

Fang's Discussion does not say whether those who cite Fukuhara for visfatin's discovery are right to. It says only that they do, and that the retraction concerned another claim in the same paper. COPE allows such a citation when the retraction is acknowledged.

MUST KNOW POINTS FOR YOU.

- "It was retracted, so the authors cheated" is the error. Read the notice's stated reason before you say anything about the people.
- Before you cite a paper in a thesis, a slide or a guideline, open its page and its database record on the day you write. Look for a correction, an expression of concern or a retraction.
- An expression of concern is not a verdict. Do not quote the paper as sound, and do not call it fraud.
- A clean status tells you that no editor has acted. It does not tell you the finding still holds. To know whether a finding stands, read a current review, not the article's page.
- The tool follows from one question: does the error change what the paper concludes? If not, a correction. If it does and the study is sound, retraction with republication. If the findings cannot be relied on at all, retraction. If nobody knows yet, an expression of concern.
- If you confirm an error in your own published paper, write to the journal's editor and publisher. The editor then decides.

COMMON MISREADING. "The paper has not been retracted, so it is sound." A retraction marks only what an editor has acted on. Fang and colleagues give an example from nutrition. Two journals, The Lancet and the British Medical Journal, "expressed serious reservations" about the Indo-Mediterranean Diet Heart Study. Its main author could not show the original records of ethics review and consent. Yet in 2012 the papers had not been retracted.

A paper can also be wrong because better evidence came later, and that never earns a notice. No notice means no editor has acted, nothing more.

REPORTING IT. The figures from Table 3 are taken from the authors' 2013 correction (Proc Natl Acad Sci USA 2013;110:1137), which replaced that table.

EXERCISE 1 (retrieval). From memory, name the four instruments a journal uses to change a published paper, and say in one line when each is used. Then name the kind of problem that gets none of them, and where it goes instead.

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (interpretation). Five situations, all made up. For each, name the instrument the ICMJE Recommendations and the COPE guidelines point to, or say that none applies.

(a) In a published survey from a district hospital, two column headings in Table 2 were swapped. The text, the abstract and the conclusions use the right figures.

(b) A reader writes that an observational paper on breakfast used causal words. A later trial found that advice about breakfast made no detectable difference to weight.

(c) A coding error swapped the two study groups. Corrected, the main difference points the other way. The data and the design are sound.

(d) A university is investigating claims that images in a paper were altered. The investigation has not finished.

(e) The investigation in (d) finds that the images were fabricated.

EXERCISE 3 (critique). A slide in a journal club on fat tissue reads (made up for this exercise):

> *Visfatin, a fat-tissue hormone, mimics insulin (Fukuhara et al., Science 2005).*

Say what is wrong with the citation, and rewrite the slide so it can stand.

EXERCISE 4 (teaching). For a first-year resident, ten minutes, a whiteboard. She has found that one number in a table of her published paper is wrong. She asks: "Will they retract it? Will everyone think I cheated?" Answer her.

19 · Being wrong in public is common, and owning it costs less than hiding it

Checked against the sources as they stood on 2026-10-10.

DEFINITION. Being publicly wrong is a measured, common event in medicine and science. Four sets of measurements show its shape.

Established practices are often overturned. Of 363 articles that tested an established practice, 146 (40.2 per cent) found it no better than a lesser or earlier standard.

Retraction is rare, and most of it is for misconduct. Lu and colleagues (2013) found 0.14 retractions for every 1,000 papers in biology and medicine journals in the Web of Science, a large index of journals. Fang, Steen and Casadevall (2012) classified all 2,047 retracted biomedical and life-science articles in PubMed in May 2012. Error accounted for 21.3 per cent of them. Misconduct accounted for 67.4 per cent: fraud or suspected fraud 43.4, duplicate publication 14.2 and plagiarism 9.8 per cent.

Owning an error is associated with a smaller cost than leaving it to others. Where the authors reported the error themselves, which was true of 21.9 per cent of retractions, their earlier papers showed no statistically significant loss of citations.

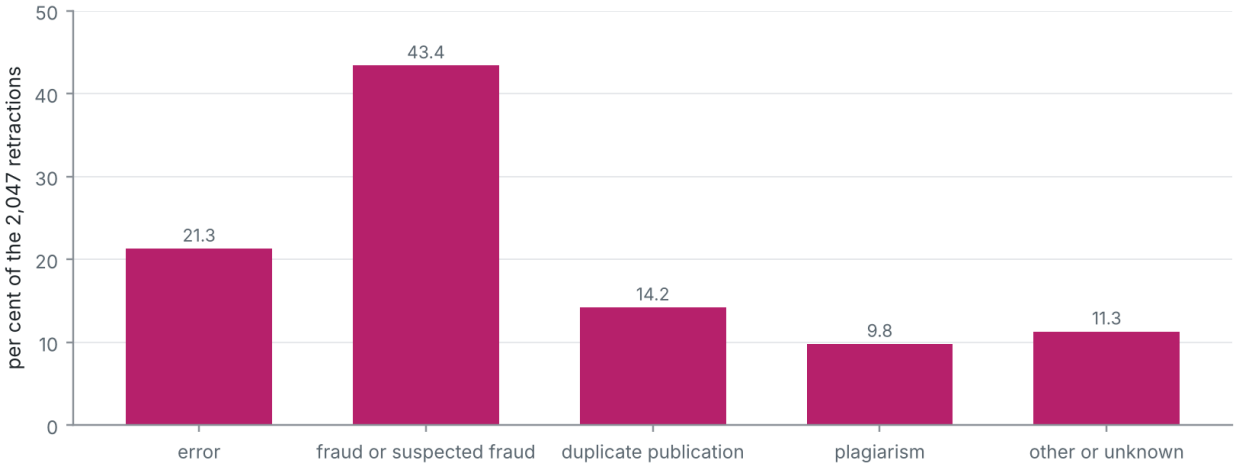
Private doubt is common and seldom made public. In an online survey, 44 per cent of 316 researchers said they had lost confidence in at least one of their own published findings (Rohrer and colleagues, 2021). Most were recruited through psychology groups and mailing lists.

Each figure comes from particular journals, databases or volunteers, mostly in high-income countries, and none was measured in India. [318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342]

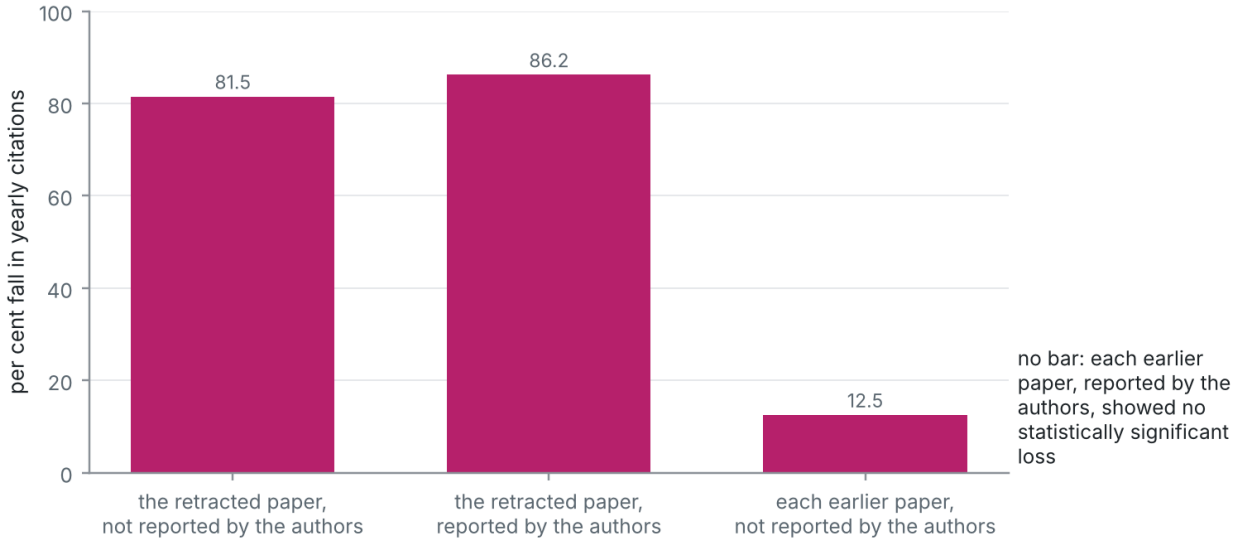
IN PLAIN TERMS. Being wrong in public is normal in medicine. Take one leading journal over ten years. About four in ten of its articles that tested an established practice found it no better than a lesser or earlier standard.

Retraction, the formal withdrawal of a paper, is different. It is rare: about 0.14 papers in every 1,000 in biology and medicine. And most retractions are for misconduct, not for honest mistakes.

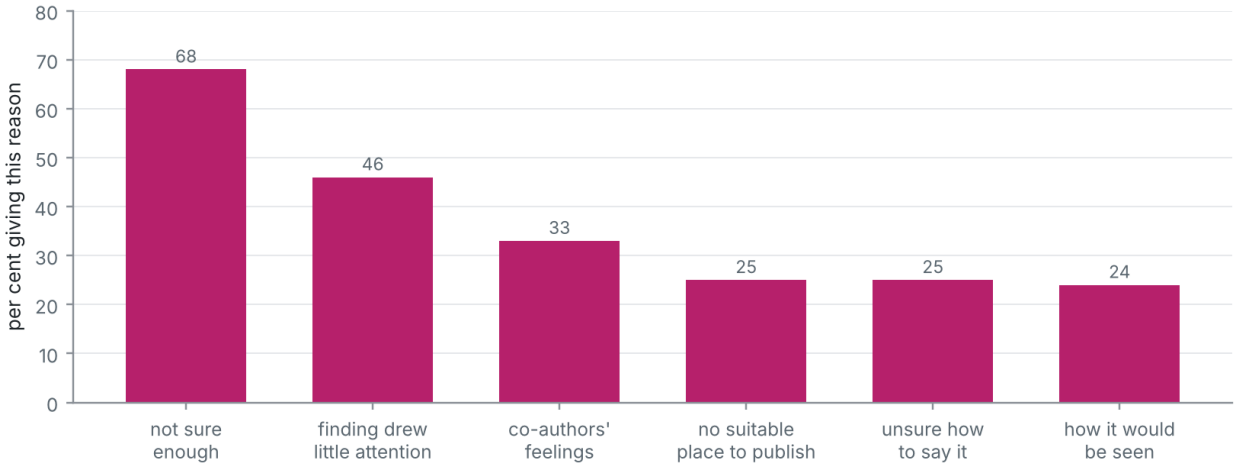
When authors report their own error, they seem to pay less for it.



Why 2,047 biomedical and life-science articles in PubMed had been retracted, as classified by Fang, Steen and Casadevall (2012): error 21.3 per cent; fraud or suspected fraud 43.4; duplicate publication 14.2; plagiarism 9.8; other or unknown 11.3 (worked out as the remainder). Shares of retractions, not of all research: retractions were about 0.14 per 1,000 papers in biology and medicine.



What a retraction cost in citations, five or more years on (Lu and colleagues, 2013). The retracted paper lost most of its yearly citations whether or not the authors had reported the error: 81.5 against 86.2 per cent. Each earlier paper by the same authors lost 12.5 per cent after a retraction they had not reported, and showed no statistically significant loss after one they had. An association: the errors people report may be milder ones.



Why researchers in Rohrer and colleagues' survey kept a loss of confidence in their own finding private. The commonest reason, given by 68 per cent, was not being sure enough. Each person could give several reasons, so the bars add to more than 100. The respondents were volunteers, not a representative sample.

Illustration 1. Start with a retraction notice that said "error", and was wrong.

Open Fang, Steen and Casadevall (2012) at https://pmc.ncbi.nlm.nih.gov/articles/PMC3479492/. In the Results, search for "FOXO1a". A retraction notice in the Journal of Cell Biology, written by the paper's own authors, read as an honest correction:

> *a retraction notice published by the authors of a manuscript in the Journal of Cell Biology stated that "In follow-up experiments . . . we have shown that the lack of FOXO1a expression reported in figure 1 is not correct" (11). A subsequent report from the Office of Research Integrity states that the first author committed "research misconduct*

The Office of Research Integrity is the US government office that oversees misconduct allegations in federally funded research. Fang and colleagues did not take notices at their word. They checked them against such reports, the news and other public records. Search for "118 of 742":

> *Use of these additional sources of information resulted in the reclassification of 118 of 742 (15.9%) retractions in an earlier study (4) from error to fraud.*

Check the share.

```
118 divided by 742 = 0.159, which is 15.9 per cent
```

So 118 of the earlier study's 742 retractions, about one in six of all of them, moved from error to fraud.

Now read their own count, for all 2,047 retracted articles. Search the abstract for "21.3%".

reason for retraction	per cent of the 2,047 retractions
error	21.3

fraud or suspected fraud	43.4
duplicate publication	14.2
plagiarism	9.8
other or unknown	11.3

The last row is not printed as a number. The paper says "miscellaneous reasons or unknown causes accounted for the remainder", so work it out.

```
43.4 plus 14.2 plus 9.8 = 67.4, the share for misconduct
21.3 plus 67.4 = 88.7
100 minus 88.7 = 11.3, the share for other or unknown reasons
```

The paper also says that where the reason is known, "three-quarters" were for misconduct. Check that too. Leave out the other-or-unknown share and divide.

```
67.4 divided by 88.7 = 0.760, about three-quarters
21.3 divided by 88.7 = 0.240, about one-quarter
```

Two lessons sit in this table. A notice's stated reason is where you start, not the verdict. And retraction for honest error is the smaller part of all retraction, which is why the word "retracted" sounds like an accusation.

Figures quoted in this illustration are from [317, 343].

WHERE THIS PICTURE BREAKS. These are shares of retractions, not of research. Lu and colleagues found about 0.14 retractions per 1,000 papers in biology and medicine. So 43.4 per cent of retractions is a small fraction of all papers. Fang and colleagues add that "only a fraction of fraudulent articles are retracted", so it is also an undercount of fraud.

The 2,047 were English-language articles in PubMed up to May 2012. Retraction practice has changed since; COPE's 2025 guidelines, for instance, describe batch retractions of many papers from paper mills, people or groups who repeatedly cheat the publishing process, for example to publish fabricated research. Do not carry these shares to a later year or another database without a newer count.

Illustration 2. Now the cost of owning an error. Lu, Jin, Uzzi and Jones (2013) asked what a retraction does to the citations of an author's earlier papers, the ones nobody retracted. Their paper is at https://pmc.ncbi.nlm.nih.gov/articles/PMC3818648/.

Read the method first. For each earlier paper they found papers in the same field and year that had been cited in the same way before the retraction. Then they compared what happened after it. They kept only authors with a single retraction: 667 retracted papers, 1,737 authors, 45,039 earlier papers.

Then they split the retractions in two. Across all 1,423 retractions in the database, 312 had been reported by the authors themselves. Check the share.

```
312 divided by 1,423 = 0.219, which is 21.9 per cent
```

The citation results below come from the 667 papers. Lu and colleagues do not say how many of those were self-reported.

Now read two results side by side, from the Results section.

what lost citations, five or more years after retraction	not reported by the authors	reported by the authors
the retracted paper itself	81.5 per cent	86.2 per cent
each earlier paper by the same authors	12.5 per cent	no statistically significant loss

The retracted paper loses most of its citations either way. Reporting your own error does not save the paper. The difference is in the earlier work. After a retraction the authors did not report, each earlier paper lost 12.5 per cent of its yearly citations, five years on. After one they reported, Lu and colleagues found no statistically significant loss.

Earlier papers are many. On average each author had 25.9 of them. Lu and colleagues also give the loss as a count. Five or more years after a retraction the authors did not report, each earlier paper lost about 0.208 citations a year. Multiply by 25.9 for a typical author's total.

```
0.208 times 25.9 = 5.39 citations a year
```

Now say what this does not show. The authors were not assigned at random to report or to stay silent. Lu and colleagues give two readings themselves. Self-reported errors "may reflect more innocuous or explainable errors". Or readers may reward people who correct their own mistakes. The study cannot tell these apart.

Figures quoted in this illustration are from [343].

WHERE THIS PICTURE BREAKS. This is an association between self-reporting and later citations. It is not a measured effect of choosing to self-report, because the errors people choose to report may be milder ones.

Citations are one cost. The study did not measure jobs, grants or how colleagues treat you.

It covers retractions in the Web of Science from 2000 onwards, across all fields. It does not cover corrections, which are smaller events, and it measured nothing in India.

Illustration 3. Last, the errors that never reach a journal. Rohrer and colleagues (2021) asked researchers whether they had ever lost confidence in one of their own published findings. Their paper is at https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/. Search for "316 scientists".

Of 316 who answered, 44 per cent said yes, at least once. Another 14 per cent were not sure. Each was then asked about the one case they felt most responsible for. In 17 per cent of those cases, the loss of confidence was a matter of public record.

Then read why the rest stayed silent. Each reason could be given alongside others.

reason given for not making it public	per cent
not sure enough about the loss of confidence	68

reason given for not making it public	per cent
the finding had not attracted much attention	46
worry about hurting co-authors' feelings	33
no suitable place to publish it	25
unsure how best to say it	25
worry about how it would be seen	24

Look at the first row. Two in three waited to be sure. But a loss of confidence is a change in how sure you are, and it can be said as one: "I am now much less sure of this than I was, and here is why." Waiting to be sure that you were wrong holds your doubt to a standard the original finding never met.

Figures quoted in this illustration are from [292].

WHERE THIS PICTURE BREAKS. The authors warn that "our respondents do not constitute a representative sample of researchers". They answered a link posted in online groups, 37 per cent from psychology departments, nearly all in Europe or the United States. People who care about self-correction may have been more likely to answer. Read 44 per cent as "common among these volunteers", not as a rate for medicine or for India.

MUST KNOW POINTS FOR YOU.

- Tell trainees that being overturned is ordinary.
- "Owning a mistake will finish my career" is the fear. One large study points the other way. When authors reported the error themselves, their earlier papers showed no statistically significant loss of citations. The study prints no interval in its text. When they had not, each earlier paper lost 12.5 per cent of its citations a year, five years on. The two groups were not made by chance: the errors people report may be milder.
- A retraction notice may give the wrong reason.
- You do not need to be sure you were wrong before you say your confidence has dropped. Say how sure you now are, and why, in the words the section on stating confidence taught.
- These numbers come from high-income-country journals (92 per cent of Herrera-Perez's reversals were tested in high-income populations), English-language databases and a volunteer survey. None was measured in India. Use them to show that being wrong is common, not to put a figure on how often Indian research is wrong.
- A reversal and a retraction are different events with different denominators. Never set the two side by side as if they measured the same thing.

COMMON MISREADING. "Two-thirds of retracted papers involve misconduct, so most research cannot be trusted." The 67.4 per cent is a share of retractions. And retractions are rare: about 0.14 in every 1,000 papers in biology and medicine. Fang and colleagues say that papers retracted for fraud are "a very small percentage of the scientific literature". They also say only some fraudulent papers are ever retracted, so the true share is larger than the count. Still, neither

point makes most research false. The common way to be wrong is the other one: honest work that later evidence overturns.

EXERCISE 1 (retrieval). From memory: what share of 363 tests of an established practice in one journal reversed it? Of 2,047 retracted articles, roughly what share was retracted for error, and what share for misconduct? What did Lu and colleagues find about the earlier papers of authors who reported their own error?

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (interpretation). A colleague reads Lu and colleagues' table of results and says: "So if I report my own error, my retracted paper will keep its citations." Use the two rows of the table in this section to say what the study shows and what it does not.

EXERCISE 3 (critique). A newspaper health column reads (made up for this exercise):

> *Two-thirds of medical papers that get withdrawn turn out to be fraud. And 40 per cent of what doctors do is later found to be useless. You cannot trust medical research.*

Find each step that does not follow from the studies behind these figures.

EXERCISE 4 (teaching). For a health secretary, one page, answer first. After a news story on retracted papers, she asks: "Should the department stop relying on published research?"

20 · Saying "I was wrong about that"

DEFINITION. Saying "I was wrong about that" is a procedure of five steps.

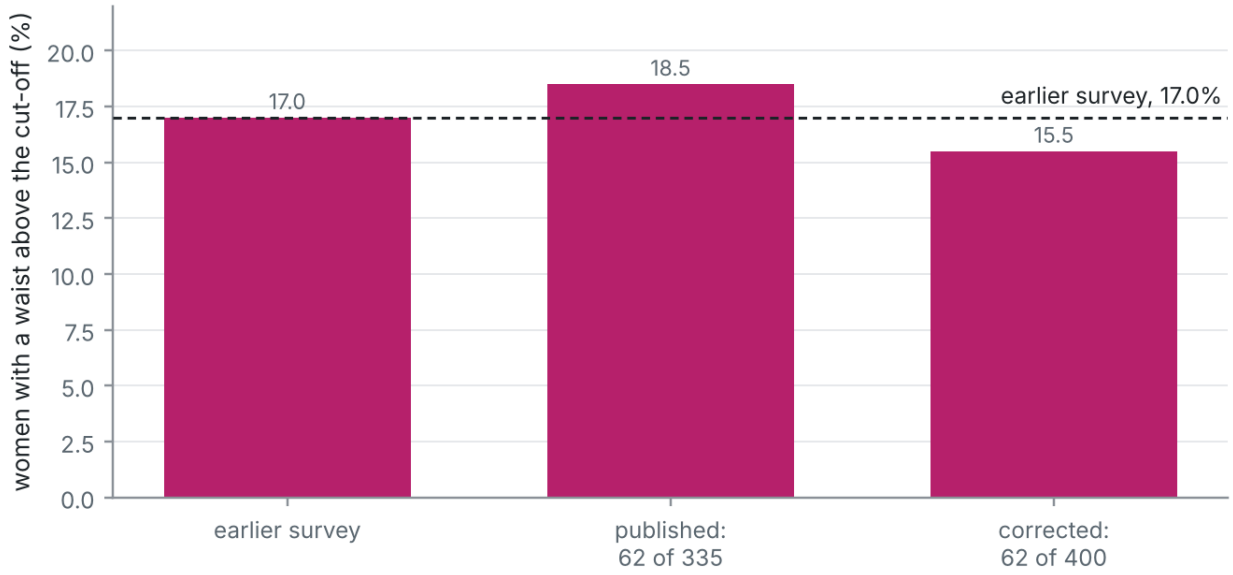
1. Find. List the confident claims you have made in public. For each, say what kind of statement it was and what it rested on. Then check it against a recent systematic review of the question, and that review's certainty of evidence.
2. Sort. Decide which of three things happened. The evidence moved: your claim fitted what was known when you made it, and later work changed the picture. You overstated: the evidence of the time did not carry the claim. Or you made an error of fact or analysis in your own work.
3. Choose the instrument. Correct the claim where its audience met it. A confirmed error in your own published work goes to the journal's editor, who decides between a correction, a retraction with republication and a retraction.
4. Say it, in the four sentences of the section on changing a stated position out loud. What I said, to whom, and when. What new evidence has appeared since. What I now think, how confident I am, and on what grounds. What I will do differently.
5. Date it and keep it, so that you, and anyone who asks, can see when you changed your mind and why.

In 2021 Rohrer and colleagues found no standard protocol for what to do when you lose confidence in your own finding; these steps are this book's. [2, 344, 345, 294, 295, 310, 346, 310, 347, 348, 349, 337, 350, 351, 3, 53, 352, 353, 354, 355, 356, 357, 330]

IN PLAIN TERMS. Sooner or later something you taught or published will turn out to be wrong.

When it does, put the correction where people met the claim. Tell the same class. Change the slides.

This is not an apology for honest work, and it is not a confession of cheating. It is how science stays right.



The made-up waist survey: the numbers are invented. The published 18.5 per cent (62 of 335) sits 1.5 percentage points above the earlier survey's 17.0 per cent: a rise. The corrected 15.5 per cent (62 of 400) sits 1.5 points below it: a fall. The same error leaves "a large share" standing and reverses "the share has gone up".

Illustration 1. Here is a slide from a postgraduate lecture on obesity, given in 2016. The slide is made up for this section. The studies named after it are real.

> *Breakfast skipping causes weight gain. Advise every patient to eat breakfast.*

Find. You search your old lecture files for "breakfast" and this slide comes up. Put the two questions from the section on belief beyond the evidence. What kind of statement is it? A cause, stated as fact. What did it rest on? Observational studies, which show only that people who skip breakfast more often have overweight or obesity.

Now check it against a systematic review. Sievert and colleagues (2019) pooled the randomised trials. Their paper is at https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/. They found a small difference in weight that favoured the people who skipped breakfast: a mean difference of 0.44 kg, with a 95 per cent confidence interval of 0.07 to 0.82 kg. They also wrote that the trials were short and "the quality of the included studies was mostly low". So the review does not support the slide.

Sort. Here is the trap. It is tempting to say "the evidence changed". Check the dates. In 2013, Brown and colleagues had already called the causal claim "only presumed true". In 2014 a trial of 309 adults had found that advice to eat or skip breakfast made no detectable difference to weight over 16 weeks. By 2016 the slide went beyond the evidence of its own time. The honest sort is "I overstated", with later evidence on top. The same slide shown in 2012, before both, would sort differently: the advice fitted the evidence of its time and the evidence moved, though the causal wording went too far even then (the section on changing a stated position).

Choose the instrument. The claim was taught, so correct it to the people who were taught it. Say it at the next session of the same course. Ask whoever runs the course to pass a note to the batch that heard it. Then change the slide.

Say it. One version, in its four sentences. Its confidence uses this book's scale for the breakfast claim, Intelligence Community Directive 203 (ICD 203):

> 1. *In my 2016 lecture on obesity, I said that skipping breakfast causes weight gain, and that every patient should be told to eat breakfast.*
> 2. *That rested on observational studies, and it went further than the evidence even then: a 2014 trial found that advice to eat or skip breakfast made no detectable difference to weight over 16 weeks. Since then, a 2019 review of trials found a small difference of 0.44 kg in favour of skipping, from short trials of mostly low quality.*
> 3. *I now think it "unlikely" (about 25 per cent, on the ICD 203 scale) that eating breakfast helps adults lose weight; my grounds are the 2019 review's trials, short and mostly of low quality.*
> 4. *I have changed the slide, and for advice about meals I will cite trials, not associations.*

Notice what is missing. There is no apology for teaching, and no claim that you always knew. Sentence 2 says plainly that the claim went too far. Sentence 3 gives a word, its scale and a number, so a listener knows how sure you now are. On ICD 203, "unlikely" runs from 20 to 45 per cent.

Date it. Write one line in the notebook where you keep such changes:

> *10 October 2026. Corrected the breakfast claim in my 2016 obesity lecture. Told the current course; note sent to the 2016 batch; slide changed. Now "unlikely" (ICD 203), about 25 per cent.*

Figures quoted in this illustration are from [167, 33, 230].

WHERE THIS PICTURE BREAKS. The confidence in sentence 3 is one reader's judgement from this evidence, not a number any study gives. Another careful reader might put it at 30 or 35. What the procedure requires is that a word, its scale and a number are stated, with the grounds beside them.

All the trials behind this case were in high-income countries; Sievert and colleagues' review was limited to them. Sentence 3 is about adults like those trial participants. It says nothing about children, or about settings where a missed meal means too little food.

Correcting a talk is easier than correcting a paper. Nobody keeps a public record of your old lecture, so the correction reaches only the people you can still reach.

Illustration 2. Now an error in your own published paper. The paper and its numbers are made up.

Your paper reported that 62 of 335 women screened at an urban health centre had a waist above the clinic's cut-off. Reusing the data for a new project, you find a filter in your analysis that wrongly left out 65 women who were measured. None of the 65 was above the cut-off. So the right count is 62 of 400.

```
335 plus 65 = 400
62 divided by 335 = 0.185, which is 18.5 per cent (published)
62 divided by 400 = 0.155, which is 15.5 per cent (correct)
```

Sort. This is an error in your own analysis. It is not evidence moving on.

Choose the instrument. The question that decides it: does the error change what the paper concludes? Take two versions of the paper.

In the first, the conclusion is that a large share of women at the centre had a waist above the cut-off, and that the centre should measure waists routinely. At 15.5 per cent instead of 18.5, that still stands. The tool is a correction.

In the second, the main claim is a rise. An earlier survey at the same centre had found 17.0 per cent, and the paper concluded that the share had gone up. Compare both figures with it.

```
18.5 minus 17.0 = 1.5 percentage points: a rise (published)
15.5 minus 17.0 = minus 1.5 percentage points: a fall (correct)
```

The direction of the main result has flipped, though the study itself is sound. That is the case the International Committee of Medical Journal Editors (ICMJE) describes for retraction with republication.

In both versions, the editor decides, not you. The Committee on Publication Ethics (COPE) asks you to raise a confirmed error with the editor and publisher, and to give "sufficient explanation and documentation" so that the editor can judge it.

Say it. Write to the editor. Give the citation of your article. Say which figure is wrong, how you found the error, and what the right figure is. Say whether the conclusion changes, and which instrument you think fits. Attach the corrected analysis.

Keep the notice you hope for as short as the real one you met in the section on correcting the record: Fang and colleagues' 2013 correction was two sentences and a table.

Date it. Note when you found the error and when you wrote to the journal.

WHERE THIS PICTURE BREAKS. Real errors rarely sort this cleanly. An error may change a secondary result and leave the main one standing, or shift a number the conclusion leans on without reversing it. Then whether it "changes the conclusion" is a judgement. Say in your letter what you think and why, and leave the decision to the editor.

Some errors are not honest. If you suspect that data were altered or made up, by anyone, this is no longer a matter for four sentences. COPE says misconduct "should also be raised with the institution".

MUST KNOW POINTS FOR YOU.

- Look for your own errors before someone else finds them. Once a year, search your slides, lecture notes, papers and posts for your confident claims.
- Do not write "the evidence has changed" when the evidence of the time already failed to carry your claim. Check the dates of the studies against the date of your claim.

- "Saying I was wrong is admitting I am careless" is the error. ICMJE calls honest errors "a part of science and publishing".
- Put the correction where the claim's audience met it: the same class, the same ward, the same journal.
- Ask the deciding question of a published error: does it change what the paper concludes? If not, ask the editor for a correction. If it flips or substantially changes the result of a sound study, ask for retraction with republication.
- This procedure covers honest error and changed evidence. It stops working the moment you suspect that data were altered or invented, by you or by a co-author. Then the matter goes to the journal and to the institution, and four sentences are the wrong tool.
- Before you write to an editor, read that journal's own policy on corrections and retractions, and follow it where it differs from these steps.

COMMON MISREADING. "Saying I was wrong means saying the claim is false." It means saying the claim was not supported as strongly as you said. Christopher Chabris, withdrawing confidence in his own 1992 experiment, put it exactly: "These claims may be true, but not because of our experiment." You met the line in the section on changing a stated position. Your new position can be "we do not know", stated with its confidence and its grounds. It does not have to be the opposite of the old one.

REPORTING IT. In my 2016 lecture I said that skipping breakfast causes weight gain. That went beyond the evidence even then. Trials in adults in high-income countries, one published before that lecture, did not find that eating breakfast helped weight loss. The trials were mostly of low quality. So I have withdrawn the slide.

EXERCISE 1 (retrieval). From memory: name the five steps for saying "I was wrong about that". Then name the three things that can have happened to a claim, and the four sentences of the statement itself.

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (design). Do it once, for real. Find one claim you have published or taught that current evidence no longer supports, or never did. Sort it. Choose the instrument. Write the four

sentences, with a confidence word from a named scale and its number. Write the dated line. Then deliver it: say it to the audience, or send the letter.

EXERCISE 3 (critique). Two attempts to correct the same lecture claim about breakfast. Both are made up. Say what is wrong with each, and say which of the four sentences each one leaves out.

> 1. *I am ashamed to say that everything I taught about breakfast was rubbish. I should never have been allowed to teach it. Please ignore my lectures.*

> 1. *As research in this area has evolved, guidance on breakfast and weight has been updated in line with emerging evidence.*

EXERCISE 4 (teaching). For a first-year resident, ten minutes. He realises a slide he presented at a departmental seminar last year cited a study that has since been retracted. He asks: "If I correct it now, won't the faculty think less of me?" Answer him.

21 · The dated belief register: what you believe, how sure you are, and when you said it

DEFINITION. A dated belief register is a written list of claims one person holds, one numbered line for each, kept so that each claim can be checked later against what happened. Each line records six things.

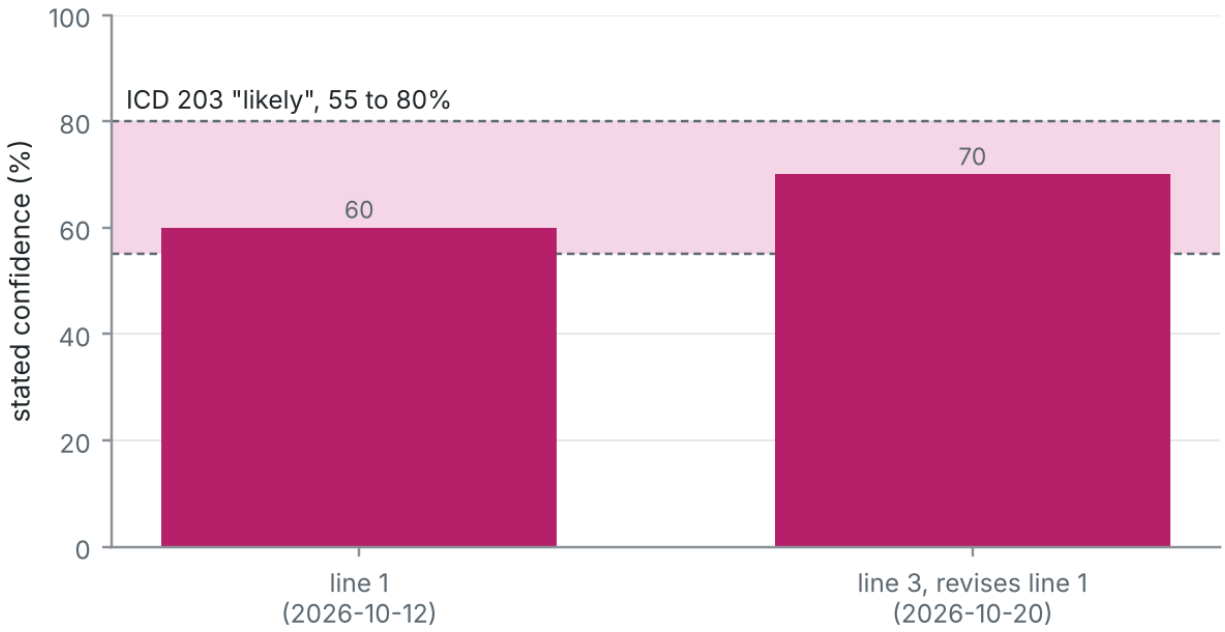
1. Its number, and the date it was written.
2. The claim, worded so that someone else could later judge it true or false: about what, in whom, measured how, and by when.
3. The person's confidence, as a word from one named scale with the number beside it.
4. The grounds: the kind of evidence the confidence rests on, and, where it has been rated, its certainty; where it has not, how directly it bears on the claim.
5. What would change the person's mind: a result or event named in advance.
6. A date to review the line.

A line is never edited once written. The register is only added to. A changed belief is a new dated line that names the line it revises, and the old line stays as it was. When a claim can be judged, its outcome is also added as a new dated line. [180, 231, 358, 359, 355]

IN PLAIN TERMS. A belief you never wrote down cannot be checked.

So keep a register. On each line write the date and the claim. Write how sure you are, as a word and a number. Write why, and what would make you change your mind. Write when you will look at it again.

Then never change a line. If your view changes, write a new line with a new date, and say which old line it replaces.



The made-up resident's confidence in her line 1 claim, then in line 3, which revises it. The number moved from 60 to 70 per cent; both sit inside ICD 203's "likely" band, 55 to 80 per cent, so the word stayed "likely". Only the number shows how far the new count moved her. Both numbers are her judgements.

Illustration 1. Start with the way most of us keep beliefs. Here is a line from a resident's journal-club notebook. It is made up for this section.

> *Reversals - probably rare. Most of what we teach is solid.*

Try to check it in ten years. You cannot. It has no date, so you cannot tell what she had read when she wrote it. "Rare" names no count and no time. "Probably" is a word with no scale, and the section on how one word is read as different numbers showed that readers turn such words into very different numbers. It gives no grounds, so you cannot tell whether she had looked at a count. Nothing in it could turn out false.

Now write it as a register line. Take the parts in order.

The claim. Make it something a stranger could judge. "Reversals are rare" cannot be judged. Pick something countable, with a date. Here is one: "At least one practice in my current postgraduate (PG) lecture slides will be reversed by a published randomised trial by 10 October 2036." Use Prasad and colleagues' meaning of a reversal, so the word is fixed: an established practice found to be no better than a lesser treatment.

The grounds. The section on base rates gave you a count. In one journal's decade, 2001-2010, 146 of 363 tests of an established practice reversed it.

```
146 divided by 363 is about 0.402
```

That is about 40 per cent of the tests. Write its limits beside it. It is one journal, and most of its practices were drugs and procedures. And it is a rate per test, not a rate per practice that you teach. So it is the nearest reference class, not a close one.

The word and the number. Pick one scale for the whole register and name it. This resident uses the seven bands of Intelligence Community Directive 203 (ICD 203), the US intelligence standard of the section on calibrated scales. Its bands meet but do not overlap; avoid a number on a band's edge. On that scale "likely" runs from 55 to 80 per cent. She judges 60 per cent, so she writes "likely". That is a judgement, not a calculation. Many practices in her slides could be reversed, which pushes her above 40. But each needs a trial published within ten years, which pulls her down. She settles on 60. She writes the claim on the side she leans to, so the number is 50 per cent or more, because that is how the section on calibration counts claims.

What would change her mind. Name it now, before anything happens. Here it is a finding that every practice in her slides has already been tested in a randomised trial and held up.

The review date. One year on, to look again, and the claim's own date, 10 October 2036, to close it.

Here is the line, with a second one she wrote the same week.

no.	date	claim	confidence (ICD 203)	grounds and their certainty	what would change my mind	review
1	2026-10-12	At least one practice in my current PG lecture slides will be reversed by a published randomised trial by 2036-10-10	"likely" (60 %)	One journal's decade: 146 of 363 tests of an established practice reversed it. A measured rate, but for a different class of practice	Finding that every practice in my slides has been tested and held up	2027-10-12; closes 2036-10-10
2	2026-10-14	If I write "unlikely" with no number in a patient leaflet, more than half of the next 20 patients I ask will read it as a larger chance than the 25% I mean	"likely" (65 %)	One UK study: mothers read "unlikely" as 27%, doctors as 16%. Indirect for my patients, none studied in India	Patients reading "unlikely" as doctors do	2027-01-31 (review and close)

Read across line 2. The claim names who, what and by when. Its grounds are one study in another country, so they are weak, and the line says so.

Figures quoted in this illustration are from [258, 230, 186].

WHERE THIS PICTURE BREAKS. Both lines are made up, and so are the resident's numbers, 60 and 65 per cent. They are judgements, not measurements. The counts in their grounds are real: Prasad and colleagues' for line 1, and Shaw and Dear's, as Hashim reports them, for line 2.

Line 1 borrows a rate from one American journal's trials of mostly drugs and procedures. It does not tell you how often a lecture on nutrition will be overturned.

Illustration 2. A week later the resident reads Herrera-Perez and colleagues. Across more than 3,000 randomised trials in three journals, they found 396 reversals. Their count is of all trials,

not only tests of an established practice, so it is no rate to set beside her 40 per cent. What it adds is breadth. Reversals turn up across three journals, so it is more likely that some practice in her slides is among them. She moves up a little.

The tempting move is to cross out "60%" and write "70%". Do not. Then the register would say she believed 70 per cent on 12 October, which is false. Instead she adds a line.

no.	date	claim	confidence (ICD 203)	grounds and their certainty	what would change my mind	review
3	2026-10-20	Revises line 1. Same claim	"likely" (70%)	As line 1, and: 396 reversals among more than 3,000 trials in three journals. Reversals are not one journal's quirk; still mostly drugs and procedures	As line 1	2027-10-12; closes 2036-10-10

Notice what the number did that the word could not. The word stayed "likely". The number moved from 60 to 70, and it shows by how much the new evidence moved her.

Now suppose it is 31 January 2027, and line 2 can be judged. She adds the outcome as line 4, with its date: true or false, and how many of the 20 patients read "unlikely" as more than 25 per cent. Line 2 stays as written. The pair of lines is what the tally of stated confidence against outcome counts.

Figures quoted in this illustration are from [259].

WHERE THIS PICTURE BREAKS. The move from 60 to 70 per cent is her judgement of how much a second count should move her. The register records the judgement. It does not show that the judgement was right.

A line about an event, like line 2, closes on its date. A line about an effect, such as "this diet lowers weight", may never close cleanly. Judge it at its review date against the best review then available, and say in the outcome line which review you used.

MUST KNOW POINTS FOR YOU.

- Write the date and the number on every line, before you know the outcome.
- Never edit a line. If your view changes, add a new dated line that names the line it revises.
- Word each claim so that a stranger could judge it: about what, in whom, measured how, by when. If no result could make the line false, it is not yet a claim.
- Use one named scale for the whole register, and write the number beside the word.
- Write what would change your mind before anything happens. Written later, it bends to fit what did happen.
- Before a committee or a departmental meeting discusses a question, write your own line first.
- Five lines cannot tell you whether you are well calibrated.

- A register shows what you believed and when. It does not show that you were right, or that your grounds were sound.
- When you ask residents to keep a register, review the closed lines together without blame.

COMMON MISREADING. "A register is a list of my opinions, so I can tidy it as my opinions change." That reading turns the register into a summary of what you think now. A summary is useful, but it cannot be checked.

The register's whole value is that its old lines are untouched. They are the only record of what you believed before you knew the outcome. Tidy a line, and you have destroyed the one thing the tally of stated confidence against outcome (the section on calibration) needs.

REPORTING IT. On [date] I recorded that [claim, with who, what and by when] was "[word]" ([number] per cent, on the [named] scale). My grounds were [kind of evidence and its certainty]. I said I would change my mind if [result or event], and I will review it on [date].

EXERCISE 1 (retrieval). From memory, list what goes on one line of a dated belief register, in order. Then state the one rule about changing a line, and why it exists.

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (critique). A colleague shows you the first line of her register. It is made up for this exercise.

> *2026-10-15. Intermittent fasting probably works. Very confident. Review: sometime next year.*

Name every fault, and rewrite the line in register form. Leave a marked gap for anything you would have to look up.

EXERCISE 3 (teaching). For a first-year resident, ten minutes with a whiteboard. She says: "Why write my beliefs down? I know what I think." Explain what a register does that her memory cannot.

EXERCISE 4 (build). This book's build: a written record of five current beliefs about your own field, with confidence levels, dated.

1. Choose five beliefs you would state to a resident, a district health officer or a reporter. Include at least two that will close on a date you name.
2. Word each as a claim someone else could judge: what, in whom, measured how, by when.
3. Choose one scale and name it at the top of the page. For each line, write the word and the number.
4. For each, write the grounds: the kind of evidence and its certainty, from a source you have opened. Say so where the evidence is from people unlike yours.
5. For each, write what would change your mind, and a review date.
6. Date the page. Then do not edit it. Add lines.

EXERCISE 5 (design). You are asked to start registers for the six residents in your department. Design how they will be kept and reviewed for two years, so that the lines can later be counted. Say where the lines live, who can change what, and when they are reviewed.

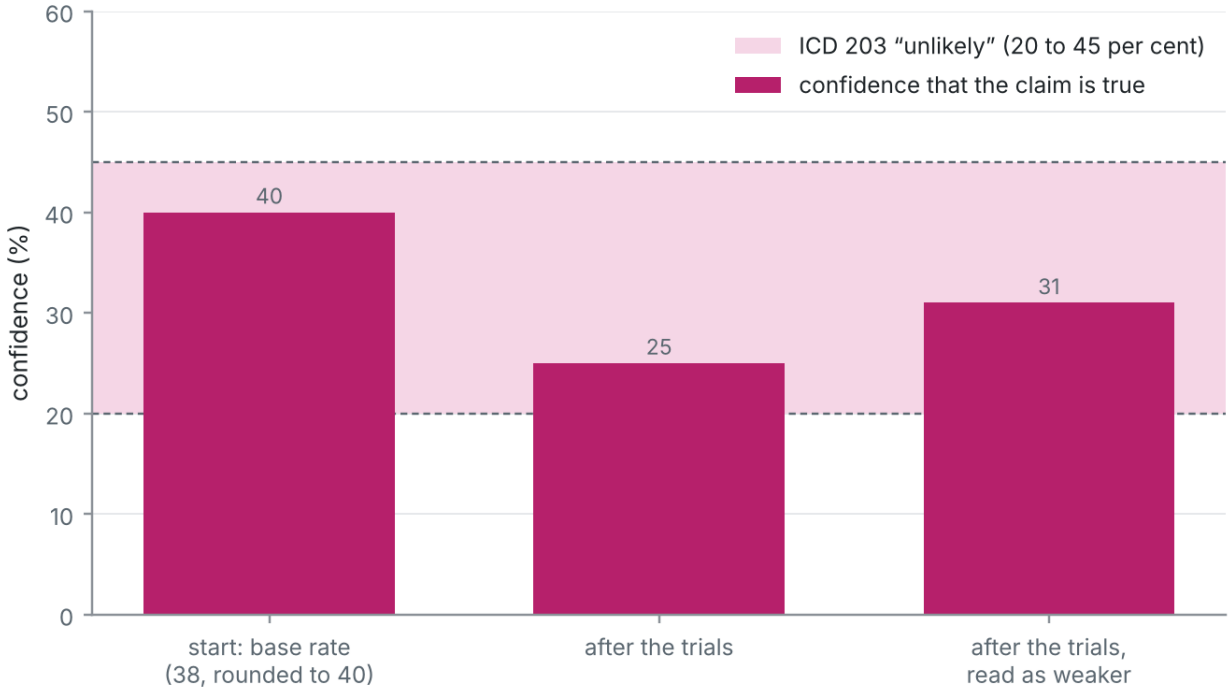
22 · Worked journey: One claim, from a belief to a dated line: “eating breakfast helps you lose weight”

DEFINITION. In this book, a claim’s journey is the order of work that takes one claim from the moment you meet it to a dated line in your register, and to saying so out loud.

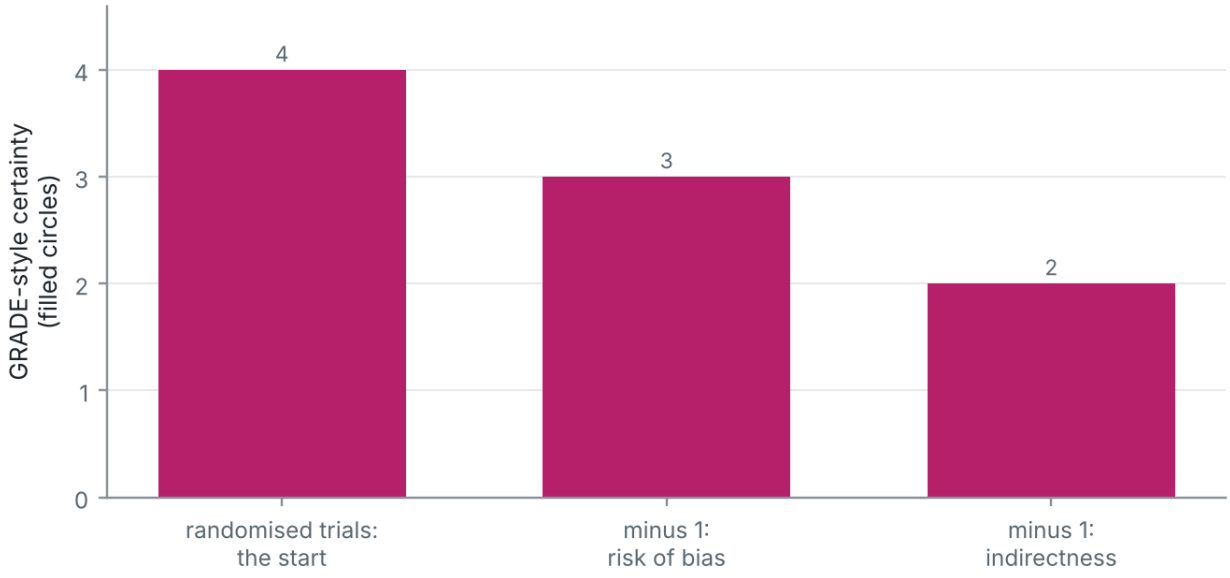
1. Sort it: hypothesis, finding or belief. Read its verb for a causal claim.
2. Find what it rests on, and list the explanations of that evidence besides the one claimed.
3. Ask the two questions of a causal claim: is there a measured difference between groups, and is there a shown mechanism? Say which leg is weak.
4. Before reading the trials, set a starting confidence from the base rate of claims like it, and say how well the reference class fits.
5. Read each trial for what it found and what it could not show.
6. Read the body of trials for its estimate and its certainty, and put the certainty into its words.
7. Update the starting confidence on that evidence, once, with a two-way table in natural frequencies (whole counts out of 1,000, as the section on updating taught).
8. Write a confidence statement in its four parts: the claim, a word from a named scale with its number, the grounds and their certainty, and what would change your mind.
9. Enter it as a dated line in your register.
10. If you taught or published the old view, change it out loud.
11. Say what the case does not show. [360, 3, 4, 77, 357, 287, 361, 53, 54, 362, 352, 363, 179, 364, 365, 366, 43, 367, 368, 369, 355]

IN PLAIN TERMS. Each earlier section taught one tool. Real work uses them one after another, on one claim.

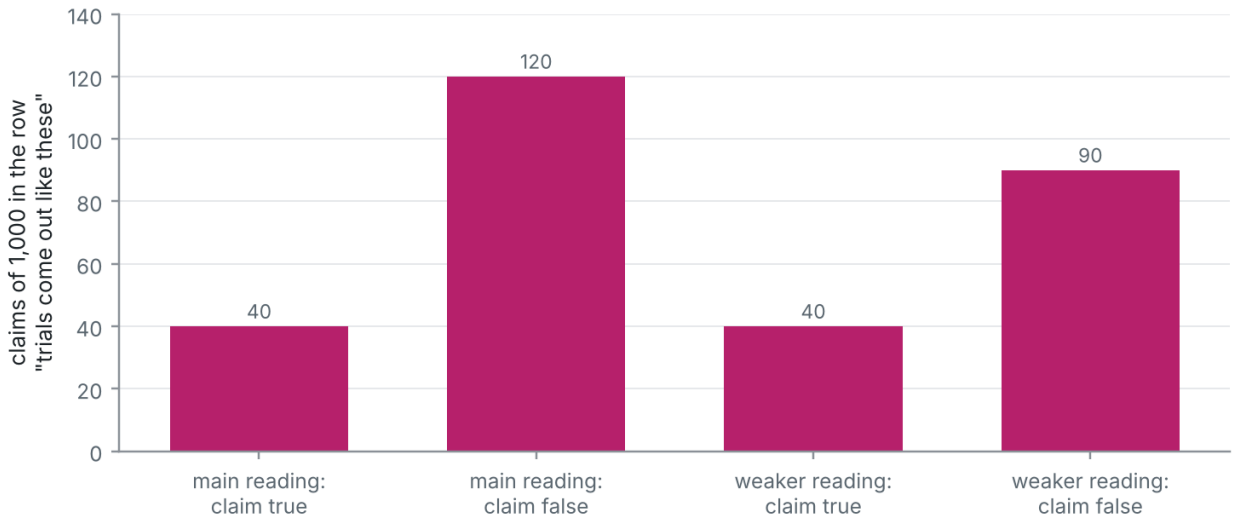
The claim is one you have heard on a ward round and in a newspaper column: eating breakfast helps you lose weight.



Confidence that eating breakfast helps adults lose weight: 40 per cent from the base rate, 25 per cent after the trials, or 31 per cent if the trials are read as weaker evidence. The shaded band is “unlikely” on the ICD 203 scale, 20 to 45 per cent. All three are judgements, not measurements.



This book’s GRADE-style reading of the review’s certainty, in filled circles. Randomised trials start at 4, high. One step down for risk of bias, one for indirectness (trials of a few weeks, a claim about losing weight): 2, low. The authors did not use GRADE.



The update in natural frequencies, out of 1,000 claims like this one. In the row of claims whose trials come out like these, 40 are true and 120 false: 40 of 160, 25 per cent. Read the trials as weaker, so that 90 false claims give such trials, and it is 40 of 130, about 31 per cent. Behind the counts are judged chances of such trials: 10 in 100 if the claim is true; 20 in 100 if it is false, or 15 in 100 on the weaker reading. They are judgements, not measured rates.

Illustration 1. Suppose a colleague's postgraduate (PG) lecture slide on obesity prevention says, in bold, "Never skip breakfast: eating breakfast helps you lose weight." The slide is made up. The claim is not; Brown and colleagues found it recommended by popular health figures, on health websites and even by the United States Surgeon General.

Sort it. It names no study and no people, so it is not a finding. Nobody is offering it to be tested, so it is not a hypothesis. It is a belief, stated in the voice of a finding. Now read the verb. "Helps you lose" says that eating breakfast changes weight: it is a causal claim.

What it rests on. Brown and colleagues pooled the observational studies, which watch people as they are. The association is firm: people who skip breakfast more often have overweight or obesity. In Brown's words, "this association supports the PEBO but does not show causation". Write the five explanations beside it, as the section on one finding with several explanations taught. Repetition makes chance a weak explanation. Bias, a third factor and the reverse direction are not, because every one of those studies shares the same design. Only the proposed cause is the claim.

The two legs. Ask the two questions of a causal claim. Was a difference in weight measured between groups formed by chance? Brown found very few trials and no consistent effect. Is there a shown mechanism? Brown again: "there is currently no clearly and consistently shown physiologic mechanism to explain the PEBO". Both legs are weak. In 2013 the claim was, in their phrase, "only presumed true".

Start from the base rate. Before you read the trials, ask how often claims of this kind hold up. Advice to eat breakfast was an established practice, recommended for weight. The nearest measured class is established practices tested in trials. Prasad and colleagues read one journal's decade of them.

```
146 divided by 363 is about 0.402
138 divided by 363 = 0.3801652893
138 divided by (146 plus 138) = 0.4859154930
```

Of 363 tests of an established practice, 146 reversed it and 138 upheld it; the rest gave no clear answer. A claim counts as true, for this purpose, if a trial would uphold it. A test that settled nothing did not uphold it, so count it against, as the section on base rates did: about 38 per cent. On clear answers only it would be 49; write which you used. For the table, round it: about 40 per cent (38, rounded), odds of 2 to 3.

Two cautions go beside that number. The class is a poor fit: one American journal, and mostly drugs and procedures, not food. And think before you move up for the observational studies. They show an association, not a cause, so they would move you little at most. Judge, too, that most practices in Prasad's count already had support of this kind when they were tested; moving up for it could count it twice. This journey leaves the start at 40 per cent and says so.

Figures quoted in this illustration are from [258].

WHERE THIS PICTURE BREAKS. The 40 per cent is a starting point taken from a different class of claims. A reader who moves up for the observational studies, to 50 per cent say, will end somewhere else after the same trials. Write down which start you used, and why, so that a reader can redo the work with another.

Illustration 2. The trial. Now read the first large trial. Dhurandhar and colleagues randomly assigned 309 adults with overweight or obesity who were trying to lose weight, in the United States and Denmark. One group was advised to eat breakfast, one to skip it, and one got general advice only. Over 16 weeks the advice changed what people said they ate at breakfast. It made no detectable difference to weight.

Read that without overreach, as the section on no evidence of an effect taught. The paper gives no confidence interval for the difference between the groups. So it does not show that breakfast has no effect on weight. It shows that this advice, in these adults, over 16 weeks, did not produce a difference the trial could detect.

The body of trials. Then read a 2019 review that pooled the trials then published. Sievert and colleagues (2019) found seven randomised trials of breakfast and weight, with 486 people, all in high-income countries. Pooled, those assigned to eat breakfast ended slightly heavier. The mean difference was 0.44 kg, favouring skipping, with a 95 per cent confidence interval of 0.07 to 0.82 kg. The interval excludes zero, and it holds only differences of under 1 kg. Follow-up averaged seven weeks.

Now rate the certainty, as GRADE does. Randomised trials start at high, four filled circles. The authors judged the quality of the body of evidence low. They gave two reasons: every trial was at high or unclear risk of bias in at least one respect, and every trial was short. Take the first as risk of bias. Take the second as indirectness: Cochrane's handbook calls trials indirect when they answer "a restricted version of the main review question". The claim is about losing weight that lasts; trials of a few weeks answer a restricted version.

```
4 minus 1 minus 1 = 2
```

Two filled circles is low. The authors did not use GRADE; these steps are this book's reading of the two reasons they gave. Put low certainty into its words, as the section on certainty words taught: low takes "may". Eating breakfast may not help adults lose weight.

Figures quoted in this illustration are from [33, 167].

WHERE THIS PICTURE BREAKS. The step count is a reading of the authors' two reasons, not their own GRADE table. The authors reported some inconsistency across the trials' results; a rater who counted it as serious would rate it very low. It is hard to see a case for rating it higher.

The 0.44 kg is an average over short trials. It is not a forecast of what one person will gain over a year by eating breakfast.

Illustration 3. Update, once. You started at about 40 per cent (38, rounded). Now move by as much as the trials warrant, with a two-way table in natural frequencies, as the section on updating taught: whole counts of 1,000 imagined claims like this one.

First, a trap. The 2014 trial is one of the seven trials Sievert pooled. Update on the review only. If you update on the trial and then again on the review, you count the same adults twice.

Fill the table from three numbers. Two of them are judgements, and you must say so.

- Of 1,000 claims like this, 400 are true. That is the starting 40 per cent. These are the two columns.
- If the claim were true, how often would a body of trials like this come out favouring skipping, with an interval that excludes zero? Seldom. Judge 10 in 100.
- If the claim were false, how often? More often, but far from always, because short trials with a high risk of bias can miss or blur a difference. Judge 20 in 100. These two rates give the rows.

```
400 times 0.10 = 40
600 times 0.20 = 120
40 plus 120 = 160
40 divided by 160 = 0.25
```

	the claim is true	the claim is false	total
trials come out like these	40	120	160
trials do not	360	480	840
total	400	600	1,000

Take the row of claims whose trials come out like these: 160. Of them, 40 are true. Your confidence moves from 40 to 25 per cent. A result like this turns up twice as often among false claims as among true ones, 20 in 100 against 10 in 100. That is why it counts against the claim. It is low-certainty evidence, so you judged it only twice as common, and it moves you some way, not all the way.

Now test how much hangs on your judgement. Suppose you read the trials as weaker still, so that a result like this would turn up only 15 times in 100 if the claim were false.

```
600 times 0.15 = 90
40 plus 90 = 130
40 divided by 130 = 0.3076923077
```

Then you move only from 40 to about 31 per cent. Weak evidence moves you a little. Both answers say the claim is less likely than not. Neither says it is settled.

WHERE THIS PICTURE BREAKS. The 10 in 100, 20 in 100 and 15 in 100 are judgements, not measured rates. No study has counted how often bodies of trials like this one come out this way for true and for false claims. The table makes the judgements visible so another reader can change them and redo the arithmetic.

A body of trials is not a laboratory test with a known catch rate. The table borrows the form of Book 0's two-way table, and its numbers are only as good as the judgements put into it.

Illustration 4. The confidence statement. Choose one scale and name it. The register of the section before this one uses ICD 203, whose bands meet but do not overlap. On it, "unlikely" runs from 20 to 45 per cent, and 25 per cent falls inside. Now write the four parts, as the section on stating confidence taught, with the grounds in a sentence of their own.

> *I think it "unlikely" (about 25 per cent, on the ICD 203 scale) that eating breakfast helps adults lose weight. Randomised trials have compared eating breakfast with skipping it in adults in high-income countries, over seven weeks on average. The difference in weight favoured skipping breakfast, by a mean of 0.44 kg (Sievert and colleagues, 2019). The authors judged the quality of this evidence low. A well-run trial of a year or more, ideally in Indian adults, in which those advised to eat breakfast lost clearly more weight would change my mind.*

Give the complement too: about a 3 in 4 chance that breakfast does not help, odds of 3 to 1 against.

The register. Enter it as a new line in your own register; here it is your first. The tally that will one day count these lines takes each claim on the side you lean to, so turn the claim round. A 25 per cent chance that breakfast helps is a 75 per cent chance that it does not. On ICD 203, "likely" runs from 55 to 80 per cent, so 75 per cent is "likely".

no.	date	claim	confidence (ICD 203)	grounds and their certainty	what would change my mind	review
1	2026-10-25	In adults in high-income countries, eating breakfast does not lead to more weight loss than skipping it, over weeks to months	"likely" (75%)	Base rate from tested established practices (poor fit), then 7 trials in high-income countries: 0.44 kg (0.07 to 0.82) favouring skipping, mean 7 weeks; quality judged low by the authors	A well-run trial of 12 months or more showing clearly more weight loss with breakfast	When a new review of breakfast trials appears, or 2028-10-25

Out loud. Your colleague taught the old view, in bold, to a room of residents, last year. First sort it, as the section on saying "I was wrong" taught: check the dates. The 2014 trial and the 2019 review were both out before her lecture, so the evidence did not move after she spoke. She overstated. The right instrument is a sentence to the same class, not a letter to a journal, because the claim was made in a lecture. She has read your working and adopts it. At the next lecture she says the four sentences of the section on changing a stated position:

> *In my lecture on obesity prevention last year, I told you that eating breakfast helps you lose weight, and I said it as a fact. That went beyond the evidence even then: trials published in 2014 and 2019 did not find that adults assigned to eat breakfast lost more weight; over a few weeks they ended slightly heavier, at low certainty. I now think it "unlikely", about 25 per cent on the ICD 203 scale, that breakfast helps weight loss; the evidence is low certainty, short and from high-income countries. I have changed the slide, and from now on I will not give breakfast as weight-loss advice.*

She adds two sentences on what was reasonable and what was not. The association is real, repeated in many studies, and public health authorities commonly recommended breakfast. Teaching it as a cause was never supported by those studies, and teaching it as fact after the trials was not reasonable. That is not an apology for honest work. It is "I was wrong about that", said with its reasons.

What the case does not show. Say these in the same breath as the conclusion.

- It is about adults. Sievert's review left out children and adolescents.
- It is about high-income countries. The weight trials were from the UK and the US, and the 2014 trial also had a site in Denmark; none was from India. The authors themselves call for studies in low- and middle-income countries, where undernutrition and obesity can sit side by side.
- It is about the short term: seven weeks on average, 16 weeks at the longest. It says nothing about a year.
- It is about weight. It says nothing about the other reasons people eat or skip breakfast.
- It does not show that breakfast causes weight gain. The 0.44 kg is small, and its certainty is low.
- The 25 per cent is a judgement, built from a borrowed base rate and two judged numbers. It is not a measurement.

Figures quoted in this illustration are from [230, 167].

WHERE THIS PICTURE BREAKS. The lecture, the slide and the register line are made up; the trials and their numbers are real. A lecturer who taught the old view in a paper, not a lecture, needs a different instrument, chosen as the section on saying "I was wrong" set out.

The confidence word fits ICD 203, and the weaker reading of 31 per cent gets the same word. The word hides the difference between the two readings; the number shows it. Write both, and name the scale every time.

MUST KNOW POINTS FOR YOU.

- Work in order.
- Set your starting confidence before you read the trials.
- Count each piece of evidence once.
- Never write "breakfast does not matter for weight" or "skipping breakfast makes you thinner".
- Mark every judgement in your working as a judgement.
- No trial in this review was done in India, and none lasted longer than 16 weeks.
- If you taught the old view, say so to the same audience, in the four sentences.
- The journey gives you a defensible number, not a true one.

COMMON MISREADING. "So the journey proved that breakfast does not help weight loss." It did not. It moved one person's confidence from 40 to 25 per cent, on low-certainty evidence from short trials in high-income countries. That is a statement about how sure to be, with its grounds. Three in four is not 100 per cent. A long, well-run trial could still move the number back up, and the line says so.

REPORTING IT. I think it "unlikely" (about 25 per cent, on the ICD 203 scale) that eating breakfast helps adults lose weight. Randomised trials have compared eating breakfast with skipping it in adults in high-income countries, over seven weeks on average. The difference in weight favoured skipping breakfast, by a mean of 0.44 kg (Sievert and colleagues, 2019). The authors judged the quality of this evidence low, so eating breakfast may increase weight slightly compared with skipping it. A good trial of a year or more could change my mind, if those told to eat breakfast lost clearly more weight.

EXERCISE 1 (retrieval). From memory, list the steps of a claim's journey in order, from first hearing the claim to what the case does not show. For each step, name the one question you ask.

Record your confidence as a percentage before turning to the answer.

EXERCISE 2 (critique). A newspaper health column reads: "New research proves skipping breakfast is the secret to weight loss." The column cites Sievert and colleagues' 2019 review. Name what the column gets wrong, using the steps of the journey, and write one sentence it could have printed.

EXERCISE 3 (teaching). For a journalist, three minutes. She asks: "Should people skip breakfast to lose weight?" Give her one quotable sentence, and tell her how sure you are without becoming useless to her.

EXERCISE 4 (design). Take a claim you have heard stated with confidence in your own department, such as on a ward round or in a lecture. Run it through the eleven steps on one sheet. End with a register line, and, if you have taught the claim, the four sentences you will say.

References

1. Clark, Mary Ann and Douglas, Matthew and Choi, Jung. Biology 2e. 2018. 1.1 The Science of Biology, The Process of Science, second paragraph. https://openstax.org/books/biology-2e (accessed 2026-09-23).
2. Clark, Mary Ann and Douglas, Matthew and Choi, Jung. Biology 2e. 2018. 1.1 The Science of Biology, The Scientific Method, Testing a Hypothesis. https://openstax.org/books/biology-2e (accessed 2026-09-23).
3. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Abstract, Results. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).
4. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Results, Research does not clearly confirm or refute the PEBO, Observational evidence. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).
5. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Methods, Cumulative meta-analysis to assess RLPV. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).
6. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Results, Table 2, the Mayo Clinic row. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).
7. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Abstract, Conclusions. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).
8. Kent, Sherman. Words of estimative probability. 1964. Opening section, the airfield passage (CIA web reprint, block 1). https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf (accessed 2026-10-09).
9. Kent, Sherman. Words of estimative probability. 1964. Opening section, the airfield passage, statement 2. https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf (accessed 2026-10-09).
10. Kent, Sherman. Words of estimative probability. 1964. Opening section, the airfield passage, statement 3. https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf (accessed 2026-10-09).
11. Kent, Sherman. Words of estimative probability. 1964. Opening section, the airfield passage. https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf (accessed 2026-10-09).
12. Kent, Sherman. Words of estimative probability. 1964. Estimative Uncertainty, first paragraph. https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf (accessed 2026-10-09).
13. Kent, Sherman. Words of estimative probability. 1964. Estimative Uncertainty, second paragraph (the third kind of statement). https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf (accessed 2026-10-09).
14. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. figures quoted in the text. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).
15. Kent, Sherman. Words of estimative probability. 1964. figures quoted in the text. https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf (accessed 2026-10-09).
16. Clark, Mary Ann and Douglas, Matthew and Choi, Jung. Biology 2e. 2018. 1.1 The Science of Biology, The Scientific Method, Testing a Hypothesis, first sentences. https://openstax.org/books/biology-2e (accessed 2026-09-23).

17. Clark, Mary Ann and Douglas, Matthew and Choi, Jung. Biology 2e. 2018. 1.1 The Science of Biology, The Scientific Method, Proposing a Hypothesis, second paragraph. https://openstax.org/books/biology-2e (accessed 2026-09-23).

18. Clark, Mary Ann and Douglas, Matthew and Choi, Jung. Biology 2e. 2018. 1.1 The Science of Biology, The Process of Science, third paragraph. https://openstax.org/books/biology-2e (accessed 2026-09-23).

19. Barseghyan, Hakob and Overgaard, Nicholas and Rupik, Gregory. Introduction to History and Philosophy of Science. 2018. 1.3 Scientific Method, Acceptance, Use, and Pursuit, first paragraph. https://human.libretexts.org/Bookshelves/Philosophy/Philosophy_of_Science/Introduction_to_History_and_Philosophy_of_Science_(Barseghyan_Overgaard_and_Rupik) (accessed 2026-10-09).

20. Barseghyan, Hakob and Overgaard, Nicholas and Rupik, Gregory. Introduction to History and Philosophy of Science. 2018. 1.3 Scientific Method, Acceptance, Use, and Pursuit, fourth paragraph. https://human.libretexts.org/Bookshelves/Philosophy/Philosophy_of_Science/Introduction_to_History_and_Philosophy_of_Science_(Barseghyan_Overgaard_and_Rupik) (accessed 2026-10-09).

21. Barseghyan, Hakob and Overgaard, Nicholas and Rupik, Gregory. Introduction to History and Philosophy of Science. 2018. 1.3 Scientific Method, Acceptance, Use, and Pursuit, fourth paragraph (the bridge example). https://human.libretexts.org/Bookshelves/Philosophy/Philosophy_of_Science/Introduction_to_History_and_Philosophy_of_Science_(Barseghyan_Overgaard_and_Rupik) (accessed 2026-10-09).

22. Barseghyan, Hakob and Overgaard, Nicholas and Rupik, Gregory. Introduction to History and Philosophy of Science. 2018. 1.3 Scientific Method, Acceptance, Use, and Pursuit, fifth paragraph. https://human.libretexts.org/Bookshelves/Philosophy/Philosophy_of_Science/Introduction_to_History_and_Philosophy_of_Science_(Barseghyan_Overgaard_and_Rupik) (accessed 2026-10-09).

23. Barseghyan, Hakob and Overgaard, Nicholas and Rupik, Gregory. Introduction to History and Philosophy of Science. 2018. 1.3 Scientific Method, Acceptance, Use, and Pursuit, closing paragraphs. https://human.libretexts.org/Bookshelves/Philosophy/Philosophy_of_Science/Introduction_to_History_and_Philosophy_of_Science_(Barseghyan_Overgaard_and_Rupik) (accessed 2026-10-09).

24. Barseghyan, Hakob and Overgaard, Nicholas and Rupik, Gregory. Introduction to History and Philosophy of Science. 2018. 1.3 Scientific Method, Acceptance, Use, and Pursuit, third paragraph. https://human.libretexts.org/Bookshelves/Philosophy/Philosophy_of_Science/Introduction_to_History_and_Philosophy_of_Science_(Barseghyan_Overgaard_and_Rupik) (accessed 2026-10-09).

25. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Introduction, last paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

26. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Introduction, second-last paragraph (the earlier trial, Schlundt et al.). https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

27. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Abstract, Results. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

28. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Subjects and Methods, Randomization. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

29. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Results, interaction of assignment with prerandomization status. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

30. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Results, Effect of control condition. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

31. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Discussion, second paragraph (control-group skippers who began eating breakfast). https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

32. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Discussion, the paragraph on limitations. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

33. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. figures quoted in the text. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

34. Clark, Mary Ann and Douglas, Matthew and Choi, Jung. Biology 2e. 2018. 1.1 The Science of Biology, The Scientific Method, Proposing a Hypothesis, first paragraph. https://openstax.org/books/biology-2e (accessed 2026-09-23).

35. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Results, Table 2 (statements of the PEBO as true), the Daily Mail row. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

36. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Discussion, third paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

37. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Introduction, second paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

38. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Methods, Intervention, last sentence of the paragraph on follow-up and measurement. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

39. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Abstract, Results (the four examples of biased reporting). https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

40. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Introduction, first paragraph (the claim defined). https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

41. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Results, The PEBO is widely stated as true, first paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

42. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Results, RLPV, second paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

43. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Results, Research does not clearly confirm or refute the PEBO, RCTs. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

44. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Results, Improper use of causal language to describe one's own results. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

45. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Methods, Assessing BRR, item 2. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

46. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Methods, Assessing BRR, item 2 (qualified causal language). https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

47. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Discussion, fifth paragraph (one trial registered). https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

48. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Results, Improper use of causal language in citing others' work. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

49. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Methods, Assessing BRR, item 4 (Wyatt et al. quoted). https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

50. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Methods, Assessing BRR, the paragraph after Table 1. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

51. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Methods, Assessing BRR, item 2 (causal language not appropriate). https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

52. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on

weight loss: a randomized controlled trial. 2014. Abstract, Design and Results. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

53. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Abstract, Conclusions. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

54. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Subjects and Methods, Study design. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

55. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Discussion, third paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

56. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Subjects and Methods, Intervention (the three arms). https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

57. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Introduction (Schlundt's interaction). https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

58. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Discussion, fourth paragraph (the second reading). https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

59. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Methods, Cumulative meta-analysis to assess RLPV (independent estimates). https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

60. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Methods, Assessing BRR (how the ratings were made). https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

61. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Results, Biased interpretation of one's own results. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

62. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Results, Research does not clearly confirm or refute the PEBO, Intermediate endpoint studies. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

63. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Discussion (the

mere exposure effect, offered as a hypothesis). https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

64. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Discussion (untested). https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

65. Sievert, Katherine and Hussain, Sultana Monira and Page, Matthew J and Wang, Yuanyuan and Hughes, Harrison J and Malek, Mary and Cicuttini, Flavia M. Effect of breakfast on weight and energy intake: systematic review and meta-analysis of randomised controlled trials. PMID 30700403, PMC6352874, PROSPERO CRD42017057687. Open access under CC BY-NC 4.0. Quoted from the PMC open-access text. Discussion, Conclusions and future implications. https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/ (accessed 2026-10-09).

66. Howick, Jeremy and Glasziou, Paul and Aronson, Jeffrey K. The evolution of evidence hierarchies: what can Bradford Hill's `guidelines for causation' contribute?. 2009. Revising Bradford Hill's guidelines, the three categories. https://pmc.ncbi.nlm.nih.gov/articles/PMC2677430/ (accessed 2026-10-09).

67. Howick, Jeremy and Glasziou, Paul and Aronson, Jeffrey K. The evolution of evidence hierarchies: what can Bradford Hill's `guidelines for causation' contribute?. 2009. Mechanistic evidence, third paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC2677430/ (accessed 2026-10-09).

68. Howick, Jeremy and Glasziou, Paul and Aronson, Jeffrey K. The evolution of evidence hierarchies: what can Bradford Hill's `guidelines for causation' contribute?. 2009. Mechanistic evidence, third paragraph, last sentence. https://pmc.ncbi.nlm.nih.gov/articles/PMC2677430/ (accessed 2026-10-09).

69. Howick, Jeremy and Glasziou, Paul and Aronson, Jeffrey K. The evolution of evidence hierarchies: what can Bradford Hill's `guidelines for causation' contribute?. 2009. Mechanistic evidence, first paragraph (the black box). https://pmc.ncbi.nlm.nih.gov/articles/PMC2677430/ (accessed 2026-10-09).

70. Howick, Jeremy and Glasziou, Paul and Aronson, Jeffrey K. The evolution of evidence hierarchies: what can Bradford Hill's `guidelines for causation' contribute?. 2009. Plausible mechanism, first paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC2677430/ (accessed 2026-10-09).

71. Howick, Jeremy and Glasziou, Paul and Aronson, Jeffrey K. The evolution of evidence hierarchies: what can Bradford Hill's `guidelines for causation' contribute?. 2009. Size of effect not attributable to plausible confounding, first paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC2677430/ (accessed 2026-10-09).

72. Howick, Jeremy and Glasziou, Paul and Aronson, Jeffrey K. The evolution of evidence hierarchies: what can Bradford Hill's `guidelines for causation' contribute?. 2009. Introduction, third paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC2677430/ (accessed 2026-10-09).

73. Howick, Jeremy and Glasziou, Paul and Aronson, Jeffrey K. The evolution of evidence hierarchies: what can Bradford Hill's `guidelines for causation' contribute?. 2009. Size of effect not attributable to plausible confounding, third paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC2677430/ (accessed 2026-10-09).

74. Howick, Jeremy and Glasziou, Paul and Aronson, Jeffrey K. The evolution of evidence hierarchies: what can Bradford Hill's `guidelines for causation' contribute?. 2009. Dose-responsiveness. https://pmc.ncbi.nlm.nih.gov/articles/PMC2677430/ (accessed 2026-10-09).

75. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.1, paragraph on study design. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

76. Hill, Austin Bradford. The environment and disease: association or causation?. 1965. p. 299, the paragraph after (9) Analogy. https://pmc.ncbi.nlm.nih.gov/articles/PMC1898525/ (accessed 2026-10-09).

77. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Results, Intermediate endpoint studies. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

78. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Table 4, rows Kral 2011 and Levitsky 2013 (2). https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

79. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Table 4, row Levitsky 2013 (2). https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

80. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Results, Effect of treatment on weight loss. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

81. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Results, Effect of treatment on weight loss, paragraph after Table 2. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

82. Howick, Jeremy and Glasziou, Paul and Aronson, Jeffrey K. The evolution of evidence hierarchies: what can Bradford Hill's `guidelines for causation' contribute?. 2009. figures quoted in the text. https://pmc.ncbi.nlm.nih.gov/articles/PMC2677430/ (accessed 2026-10-09).

83. Clark, Mary Ann and Douglas, Matthew and Choi, Jung. Biology 2e. 2018. 1.1 The Science of Biology, Scientific Reasoning, third paragraph. https://openstax.org/books/biology-2e (accessed 2026-09-23).

84. Clark, Mary Ann and Douglas, Matthew and Choi, Jung. Biology 2e. 2018. 1.1 The Science of Biology, Scientific Reasoning, fourth paragraph. https://openstax.org/books/biology-2e (accessed 2026-09-23).

85. Clark, Mary Ann and Douglas, Matthew and Choi, Jung. Biology 2e. 2018. 1.1 The Science of Biology, The Scientific Method, Testing a Hypothesis, first paragraph. https://openstax.org/books/biology-2e (accessed 2026-09-23).

86. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.1 Factors determining the quality of evidence, paragraph after 'For authors of systematic reviews'. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

87. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.1, note on case series. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

88. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.2.1 Study limitations (Risk of Bias), opening paragraph. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

89. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.2.2 Inconsistency of results, first sentence. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

90. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.2.3 Indirectness of evidence, first paragraph. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

91. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.2.3, source 3, Differences in outcome measures (surrogate outcomes), second paragraph. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

92. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.2.4 Imprecision, opening paragraph. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

93. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.2.5 Publication bias, opening paragraph. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

94. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.3.1 Large magnitude of an effect, first paragraph. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

95. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.3.2 Dose-response gradient, first paragraph. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

96. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.1, paragraph on expert opinion. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

97. Balshem, Howard and Helfand, Mark and Sch"unemann, Holger J and Oxman, Andrew D and Kunz, Regina and Brozek, Jan and Vist, Gunn E and Falck-Ytter, Yngve and Meerpohl, Joerg and Norris, Susan and Guyatt, Gordon H. GRADE guidelines: 3. Rating the quality of evidence. 2011. Section 3, Opinion is not evidence, end of the endocrinologist example. https://doi.org/10.1016/j.jclinepi.2010.07.015 (accessed 2026-10-09).

98. Balshem, Howard and Helfand, Mark and Sch"unemann, Holger J and Oxman, Andrew D and Kunz, Regina and Brozek, Jan and Vist, Gunn E and Falck-Ytter, Yngve and Meerpohl, Joerg and Norris, Susan and Guyatt, Gordon H. GRADE guidelines: 3. Rating the quality of evidence. 2011. Section 6, Quality in GRADE means more than risk of bias, second paragraph. https://doi.org/10.1016/j.jclinepi.2010.07.015 (accessed 2026-10-09).

99. Balshem, Howard and Helfand, Mark and Sch"unemann, Holger J and Oxman, Andrew D and Kunz, Regina and Brozek, Jan and Vist, Gunn E and Falck-Ytter, Yngve and Meerpohl, Joerg and Norris, Susan and Guyatt, Gordon H. GRADE guidelines: 3. Rating the quality of evidence. 2011. Section 3, Opinion is not evidence, the paragraph on other explanations of the endocrinologist's observations. https://doi.org/10.1016/j.jclinepi.2010.07.015 (accessed 2026-10-09).

100. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.1, Example 1, Quality of evidence may differ from one outcome to another within a single study. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

101. Guyatt, Gordon H and Oxman, Andrew D and Vist, Gunn E and Kunz, Regina and Falck-Ytter, Yngve and Alonso-Coello, Pablo and Schünemann, Holger J. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. 2008. Series note at the end of the article. https://pmc.ncbi.nlm.nih.gov/articles/PMC2335261/ (accessed 2026-10-09).

102. Guyatt, Gordon H and Oxman, Andrew D and Vist, Gunn E and Kunz, Regina and Falck-Ytter, Yngve and Alonso-Coello, Pablo and Schünemann, Holger J. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. 2008. What is quality of evidence and why is it important?, second paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC2335261/ (accessed 2026-10-09).

103. Guyatt, Gordon H and Oxman, Andrew D and Vist, Gunn E and Kunz, Regina and Falck-Ytter, Yngve and Alonso-Coello, Pablo and Schünemann, Holger J. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. 2008. What is quality of evidence and why is it important?, third paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC2335261/ (accessed 2026-10-09).

104. Guyatt, Gordon H and Oxman, Andrew D and Vist, Gunn E and Kunz, Regina and Falck-Ytter, Yngve and Alonso-Coello, Pablo and Schünemann, Holger J. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. 2008. What is quality of evidence and why is it important?, third paragraph, last sentence. https://pmc.ncbi.nlm.nih.gov/articles/PMC2335261/ (accessed 2026-10-09).

105. Guyatt, Gordon H and Oxman, Andrew D and Vist, Gunn E and Kunz, Regina and Falck-Ytter, Yngve and Alonso-Coello, Pablo and Schünemann, Holger J. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. 2008. What is quality of evidence and why is it important?, fourth paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC2335261/ (accessed 2026-10-09).

106. Guyatt, Gordon H and Oxman, Andrew D and Vist, Gunn E and Kunz, Regina and Falck-Ytter, Yngve and Alonso-Coello, Pablo and Schünemann, Holger J. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. 2008. What is quality of evidence and why is it important?, fourth paragraph, last sentence. https://pmc.ncbi.nlm.nih.gov/articles/PMC2335261/ (accessed 2026-10-09).

107. Guyatt, Gordon H and Oxman, Andrew D and Vist, Gunn E and Kunz, Regina and Falck-Ytter, Yngve and Alonso-Coello, Pablo and Schünemann, Holger J. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. 2008. Competing interests. https://pmc.ncbi.nlm.nih.gov/articles/PMC2335261/ (accessed 2026-10-09).

108. Balshem, Howard and Helfand, Mark and Sch"unemann, Holger J and Oxman, Andrew D and Kunz, Regina and Brozek, Jan and Vist, Gunn E and Falck-Ytter, Yngve and Meerpohl, Joerg and Norris, Susan and Guyatt, Gordon

H. GRADE guidelines: 3. Rating the quality of evidence. 2011. Section 9, Rationale for using GRADE's definition of quality, first paragraph. https://doi.org/10.1016/j.jclinepi.2010.07.015 (accessed 2026-10-09).

109. Balshem, Howard and Helfand, Mark and Sch"unemann, Holger J and Oxman, Andrew D and Kunz, Regina and Brozek, Jan and Vist, Gunn E and Falck-Ytter, Yngve and Meerpohl, Joerg and Norris, Susan and Guyatt, Gordon H. GRADE guidelines: 3. Rating the quality of evidence. 2011. Section 9, second paragraph. https://doi.org/10.1016/j.jclinepi.2010.07.015 (accessed 2026-10-09).

110. Balshem, Howard and Helfand, Mark and Sch"unemann, Holger J and Oxman, Andrew D and Kunz, Regina and Brozek, Jan and Vist, Gunn E and Falck-Ytter, Yngve and Meerpohl, Joerg and Norris, Susan and Guyatt, Gordon H. GRADE guidelines: 3. Rating the quality of evidence. 2011. Section 9, third paragraph. https://doi.org/10.1016/j.jclinepi.2010.07.015 (accessed 2026-10-09).

111. Balshem, Howard and Helfand, Mark and Sch"unemann, Holger J and Oxman, Andrew D and Kunz, Regina and Brozek, Jan and Vist, Gunn E and Falck-Ytter, Yngve and Meerpohl, Joerg and Norris, Susan and Guyatt, Gordon H. GRADE guidelines: 3. Rating the quality of evidence. 2011. Section 9, third paragraph, last sentence. https://doi.org/10.1016/j.jclinepi.2010.07.015 (accessed 2026-10-09).

112. Balshem, Howard and Helfand, Mark and Sch"unemann, Holger J and Oxman, Andrew D and Kunz, Regina and Brozek, Jan and Vist, Gunn E and Falck-Ytter, Yngve and Meerpohl, Joerg and Norris, Susan and Guyatt, Gordon H. GRADE guidelines: 3. Rating the quality of evidence. 2011. figures quoted in the text. https://doi.org/10.1016/j.jclinepi.2010.07.015 (accessed 2026-10-09).

113. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. About the Handbook, second paragraph. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

114. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. About the Handbook, important disclaimer. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

115. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. 14.2.1 The GRADE approach, first paragraph. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

116. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. Key Points, fourth point. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

117. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.1, definition for authors of systematic reviews. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

118. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.1, Table 5.1 Quality of Evidence Grades, row High. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

119. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.1, Table 5.1, row Moderate. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

120. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.1, Table 5.1, row Low. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

121. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.1, Table 5.1, row Very Low. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

122. Balshem, Howard and Helfand, Mark and Sch"unemann, Holger J and Oxman, Andrew D and Kunz, Regina and Brozek, Jan and Vist, Gunn E and Falck-Ytter, Yngve and Meerpohl, Joerg and Norris, Susan and Guyatt, Gordon H. GRADE guidelines: 3. Rating the quality of evidence. 2011. Abstract, fourth sentence. https://doi.org/10.1016/j.jclinepi.2010.07.015 (accessed 2026-10-09).

123. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of

the evidence. 2024. 14.2.1, paragraph on ROBINS-I. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

124. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.1, paragraph before Table 5.2. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

125. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.1, Table 5.2 Factors that can reduce the quality of the evidence, first row. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

126. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. 14.2.2, opening paragraph. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

127. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. 14.2.1, paragraph after the ROBINS-I paragraph. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

128. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. 14.2.1, same paragraph, last sentence. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

129. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.1, Table 5.3 Factors that can increase the quality of the evidence. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

130. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.1, Table 5.3, row Dose-response gradient. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

131. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. 14.2.3, item 1, Large effects. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

132. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.3, opening paragraph. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

133. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.3, note after the opening paragraph. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

134. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.1, paragraph after Table 5.3. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

135. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.1, paragraph on borderline decisions. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

136. Balshem, Howard and Helfand, Mark and Sch"unemann, Holger J and Oxman, Andrew D and Kunz, Regina and Brozek, Jan and Vist, Gunn E and Falck-Ytter, Yngve and Meerpohl, Joerg and Norris, Susan and Guyatt, Gordon H. GRADE guidelines: 3. Rating the quality of evidence. 2011. Section 10, Conclusion, second paragraph. https://doi.org/10.1016/j.jclinepi.2010.07.015 (accessed 2026-10-09).

137. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. Key Points, second point. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

138. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.3.3, Example 1, When confounding is expected to reduce a demonstrated effect. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

139. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. Figure 14.1.b, footnote b. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

140. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. Figure 14.1.b, footnote c. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

141. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. Figure 14.1.b, certainty cell of the row for children over 5 years. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

142. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. Figure 14.1.b, certainty and comments cells of the row for children under 5 years. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

143. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. 14.2.1, second paragraph. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

144. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.1, note on study designs. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

145. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. 14.2.3, item 2, note after the rofecoxib example. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

146. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. 14.2.2, item 5, High probability of publication bias. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

147. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.3.1, Table 5.9, note. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

148. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. figures quoted in the text. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

149. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. figures quoted in the text. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

150. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 5.1 Factors determining the quality of evidence, the definition for authors of systematic reviews. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

151. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 6.1 Recommendations and their strength, first paragraph. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

152. Guyatt, Gordon H and Oxman, Andrew D and Vist, Gunn E and Kunz, Regina and Falck-Ytter, Yngve and Alonso-Coello, Pablo and Schünemann, Holger J. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. 2008. How does the GRADE system consider strength of recommendation?, first paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC2335261/ (accessed 2026-10-09).

153. Guyatt, Gordon H and Oxman, Andrew D and Vist, Gunn E and Kunz, Regina and Falck-Ytter, Yngve and Alonso-Coello, Pablo and Schünemann, Holger J. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. 2008. How does the GRADE system consider strength of recommendation?, second paragraph and Table 1. https://pmc.ncbi.nlm.nih.gov/articles/PMC2335261/ (accessed 2026-10-09).

154. Guyatt, Gordon H and Oxman, Andrew D and Vist, Gunn E and Kunz, Regina and Falck-Ytter, Yngve and Alonso-Coello, Pablo and Schünemann, Holger J. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. 2008. What makes a good grading system?, first paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC2335261/ (accessed 2026-10-09).

155. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 1.2 Separation of confidence in effect estimates from strength of recommendations, paragraph before Example 1. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

156. Sch"unemann, Holger J and Vist, Gunn E and Higgins, Julian PT and Santesso, Nancy and Deeks, Jonathan J and Glasziou, Paul and Akl, Elie A and Guyatt, Gordon H. Chapter 15: Interpreting results and drawing conclusions. 2024. 15.6.2 Implications for practice, end of the first paragraph. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-15 (accessed 2026-10-09).

157. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 1 Overview, For authors of systematic reviews. https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

158. Sch"unemann, Holger J and Vist, Gunn E and Higgins, Julian PT and Santesso, Nancy and Deeks, Jonathan J and Glasziou, Paul and Akl, Elie A and Guyatt, Gordon H. Chapter 15: Interpreting results and drawing conclusions. 2024. 15.6.4 Reaching conclusions, the sentence introducing Table 15.6.b. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-15 (accessed 2026-10-09).

159. Sch"unemann, Holger J and Vist, Gunn E and Higgins, Julian PT and Santesso, Nancy and Deeks, Jonathan J and Glasziou, Paul and Akl, Elie A and Guyatt, Gordon H. Chapter 15: Interpreting results and drawing conclusions. 2024. 15.6.4, Table 15.6.b, moderate certainty band, moderate effect row. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-15 (accessed 2026-10-09).

160. Sch"unemann, Holger J and Vist, Gunn E and Higgins, Julian PT and Santesso, Nancy and Deeks, Jonathan J and Glasziou, Paul and Akl, Elie A and Guyatt, Gordon H. Chapter 15: Interpreting results and drawing conclusions. 2024. 15.6.4, Table 15.6.b, very low certainty band, any effect row. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-15 (accessed 2026-10-09).

161. Guyatt, Gordon H and Oxman, Andrew D and Vist, Gunn E and Kunz, Regina and Falck-Ytter, Yngve and Alonso-Coello, Pablo and Schünemann, Holger J. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. 2008. What is “strength of recommendation” and why is it important?, first paragraph (illustration 1, case 1). https://pmc.ncbi.nlm.nih.gov/articles/PMC2335261/ (accessed 2026-10-09).

162. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 1.2, Example 2: Weak recommendation based on high quality evidence (illustration 1, case 2). https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

163. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. 1.2, Example 3: Strong recommendation based on low or very low quality evidence (illustration 1, case 3). https://gdt.gradepro.org/app/handbook/handbook.html (accessed 2026-10-09).

164. Guyatt, Gordon H and Oxman, Andrew D and Vist, Gunn E and Kunz, Regina and Falck-Ytter, Yngve and Alonso-Coello, Pablo and Schünemann, Holger J. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. 2008. What makes a good grading system?, aspirin and paracetamol example (illustration 1, case 4). https://pmc.ncbi.nlm.nih.gov/articles/PMC2335261/ (accessed 2026-10-09).

165. Sievert, Katherine and Hussain, Sultana Monira and Page, Matthew J and Wang, Yuanyuan and Hughes, Harrison J and Malek, Mary and Cicuttini, Flavia M. Effect of breakfast on weight and energy intake: systematic review and meta-analysis of randomised controlled trials. PMID 30700403, PMC6352874, PROSPERO CRD42017057687. Open access under CC BY-NC 4.0. Quoted from the PMC open-access text. Abstract, Eligibility criteria (illustration 2). https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/ (accessed 2026-10-09).

166. Sievert, Katherine and Hussain, Sultana Monira and Page, Matthew J and Wang, Yuanyuan and Hughes, Harrison J and Malek, Mary and Cicuttini, Flavia M. Effect of breakfast on weight and energy intake: systematic review and meta-analysis of randomised controlled trials. PMID 30700403, PMC6352874, PROSPERO CRD42017057687. Open access under CC BY-NC 4.0. Quoted from the PMC open-access text. Discussion, Quality of evidence, first sentence (illustration 2, reporting sentence). https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/ (accessed 2026-10-09).

167. Sievert, Katherine and Hussain, Sultana Monira and Page, Matthew J and Wang, Yuanyuan and Hughes, Harrison J and Malek, Mary and Cicuttini, Flavia M. Effect of breakfast on weight and energy intake: systematic review and meta-analysis of randomised controlled trials. PMID 30700403, PMC6352874, PROSPERO CRD42017057687. Open access under CC BY-NC 4.0. Quoted from the PMC open-access text. figures quoted in the text. https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/ (accessed 2026-10-09).

168. Altman, Douglas G and Bland, J Martin. Absence of evidence is not evidence of absence. 1995. Statistics Notes, first paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC2550545/ (accessed 2026-10-09).

169. Altman, Douglas G and Bland, J Martin. Absence of evidence is not evidence of absence. 1995. Statistics Notes, second paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC2550545/ (accessed 2026-10-09).

170. Altman, Douglas G and Bland, J Martin. Absence of evidence is not evidence of absence. 1995. Statistics Notes, fourth paragraph (the fibrinolytic overview). https://pmc.ncbi.nlm.nih.gov/articles/PMC2550545/ (accessed 2026-10-09).

171. Altman, Douglas G and Bland, J Martin. Absence of evidence is not evidence of absence. 1995. Statistics Notes, last paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC2550545/ (accessed 2026-10-09).

172. Altman, Douglas G and Bland, J Martin. Absence of evidence is not evidence of absence. 1995. Statistics Notes, fifth paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC2550545/ (accessed 2026-10-09).

173. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Discussion, paragraph comparing the Schlundt et al study. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

174. Sch"unemann, Holger J and Vist, Gunn E and Higgins, Julian PT and Santesso, Nancy and Deeks, Jonathan J and Glasziou, Paul and Akl, Elie A and Guyatt, Gordon H. Chapter 15: Interpreting results and drawing conclusions. 2024. 15.6.4 Reaching conclusions, first sentence. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-15 (accessed 2026-10-09).

175. Sch"unemann, Holger J and Vist, Gunn E and Higgins, Julian PT and Santesso, Nancy and Deeks, Jonathan J and Glasziou, Paul and Akl, Elie A and Guyatt, Gordon H. Chapter 15: Interpreting results and drawing conclusions. 2024. 15.6.4 Reaching conclusions, first paragraph. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-15 (accessed 2026-10-09).

176. Sch"unemann, Holger J and Vist, Gunn E and Higgins, Julian PT and Santesso, Nancy and Deeks, Jonathan J and Glasziou, Paul and Akl, Elie A and Guyatt, Gordon H. Chapter 15: Interpreting results and drawing conclusions. 2024. Key Points, fourth point. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-15 (accessed 2026-10-09).

177. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Results, Effect of treatment on weight loss, first sentence. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

178. Sievert, Katherine and Hussain, Sultana Monira and Page, Matthew J and Wang, Yuanyuan and Hughes, Harrison J and Malek, Mary and Cicuttini, Flavia M. Effect of breakfast on weight and energy intake: systematic review and meta-analysis of randomised controlled trials. PMID 30700403, PMC6352874, PROSPERO CRD42017057687. Open access under CC BY-NC 4.0. Quoted from the PMC open-access text. Results, Breakfast eating and weight

change, subgroup analysis by baseline body mass index. https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/ (accessed 2026-10-09).

179. Sievert, Katherine and Hussain, Sultana Monira and Page, Matthew J and Wang, Yuanyuan and Hughes, Harrison J and Malek, Mary and Cicuttini, Flavia M. Effect of breakfast on weight and energy intake: systematic review and meta-analysis of randomised controlled trials. PMID 30700403, PMC6352874, PROSPERO CRD42017057687. Open access under CC BY-NC 4.0. Quoted from the PMC open-access text. Results, caption of Fig 3 (random effects meta-analysis of weight). https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/ (accessed 2026-10-09).

180. H'ajek, Alan. Interpretations of Probability. 2023. §3.3 The Subjective Interpretation, §3.3.1 Probability as degree of belief, first paragraph. https://plato.stanford.edu/archives/win2023/entries/probability-interpret/ (accessed 2026-10-09).

181. H'ajek, Alan. Interpretations of Probability. 2023. §3.3.1 Probability as degree of belief, first paragraph, opening sentence. https://plato.stanford.edu/archives/win2023/entries/probability-interpret/ (accessed 2026-10-09).

182. H'ajek, Alan. Interpretations of Probability. 2023. §3.3.1 Probability as degree of belief, second paragraph. https://plato.stanford.edu/archives/win2023/entries/probability-interpret/ (accessed 2026-10-09).

183. Mastrandrea, Michael D and Field, Christopher B and Stocker, Thomas F and Edenhofer, Ottmar and Ebi, Kristie L and Frame, David J and Held, Hermann and Kriegler, Elmar and Mach, Katharine J and Matschoss, Patrick R and Plattner, Gian-Kasper and Yohe, Gary W and Zwiers, Francis W. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. 2010. Paragraph 10, first and second sentences. https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf (accessed 2026-10-09).

184. Mastrandrea, Michael D and Field, Christopher B and Stocker, Thomas F and Edenhofer, Ottmar and Ebi, Kristie L and Frame, David J and Held, Hermann and Kriegler, Elmar and Mach, Katharine J and Matschoss, Patrick R and Plattner, Gian-Kasper and Yohe, Gary W and Zwiers, Francis W. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. 2010. Introductory box, second bullet under 'two metrics'. https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf (accessed 2026-10-09).

185. Mastrandrea, Michael D and Field, Christopher B and Stocker, Thomas F and Edenhofer, Ottmar and Ebi, Kristie L and Frame, David J and Held, Hermann and Kriegler, Elmar and Mach, Katharine J and Matschoss, Patrick R and Plattner, Gian-Kasper and Yohe, Gary W and Zwiers, Francis W. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. 2010. Paragraph 10, near the end. https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf (accessed 2026-10-09).

186. Hashim, M Jawad. Verbal probability terms for communicating clinical risk - a systematic review. 2024. figures quoted in the text. https://pmc.ncbi.nlm.nih.gov/articles/PMC11067312/ (accessed 2026-10-09).

187. Kent, Sherman. Words of estimative probability. 1964. 'Early Brush with Ambiguity'. https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf (accessed 2026-10-09).

188. Kent, Sherman. Words of estimative probability. 1964. 'Early Brush with Ambiguity', the paragraph on National Intelligence Estimate 29-51. https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf (accessed 2026-10-09).

189. Wintle, Bonnie C and Fraser, Hannah and Wills, Ben C and Nicholson, Ann E and Fidler, Fiona. Verbal probabilities: very likely to be somewhat more confusing than numbers. 2019. Abstract. https://pmc.ncbi.nlm.nih.gov/articles/PMC6469752/ (accessed 2026-10-09).

190. Wintle, Bonnie C and Fraser, Hannah and Wills, Ben C and Nicholson, Ann E and Fidler, Fiona. Verbal probabilities: very likely to be somewhat more confusing than numbers. 2019. Results, the paragraph pooling percentage overlap across the four phrases. https://pmc.ncbi.nlm.nih.gov/articles/PMC6469752/ (accessed 2026-10-09).

191. Wintle, Bonnie C and Fraser, Hannah and Wills, Ben C and Nicholson, Ann E and Fidler, Fiona. Verbal probabilities: very likely to be somewhat more confusing than numbers. 2019. Results, paragraph on best estimates within the guidelines (Fig 5). https://pmc.ncbi.nlm.nih.gov/articles/PMC6469752/ (accessed 2026-10-09).

192. Wintle, Bonnie C and Fraser, Hannah and Wills, Ben C and Nicholson, Ann E and Fidler, Fiona. Verbal probabilities: very likely to be somewhat more confusing than numbers. 2019. Introduction, first paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC6469752/ (accessed 2026-10-09).

193. Hashim, M Jawad. Verbal probability terms for communicating clinical risk - a systematic review. 2024. Results, first paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC11067312/ (accessed 2026-10-09).

194. Hashim, M Jawad. Verbal probability terms for communicating clinical risk - a systematic review. 2024. Discussion, second paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC11067312/ (accessed 2026-10-09).

195. Hashim, M Jawad. Verbal probability terms for communicating clinical risk - a systematic review. 2024. Table 1, row Shaw & Dear. https://pmc.ncbi.nlm.nih.gov/articles/PMC11067312/ (accessed 2026-10-09).

196. Wintle, Bonnie C and Fraser, Hannah and Wills, Ben C and Nicholson, Ann E and Fidler, Fiona. Verbal probabilities: very likely to be somewhat more confusing than numbers. 2019. figures quoted in the text. https://pmc.ncbi.nlm.nih.gov/articles/PMC6469752/ (accessed 2026-10-09).

197. Kent, Sherman. Words of estimative probability. 1964. Block 5, the five-band chart as printed on p. 8 of the reprint, row ‘Almost certain’. https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf (accessed 2026-10-09).

198. Kent, Sherman. Words of estimative probability. 1964. Block 5, row ‘Probable’. https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf (accessed 2026-10-09).

199. Kent, Sherman. Words of estimative probability. 1964. Block 5, row ‘Chances about even’. https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf (accessed 2026-10-09).

200. Kent, Sherman. Words of estimative probability. 1964. Block 5, row ‘Probably not’. https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf (accessed 2026-10-09).

201. Kent, Sherman. Words of estimative probability. 1964. Block 5, row ‘Almost certainly not’. https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf (accessed 2026-10-09).

202. Kent, Sherman. Words of estimative probability. 1964. Block 3, ‘The Growth of Variants’, table of synonyms, the ‘Probable’ entry. https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf (accessed 2026-10-09).

203. Kent, Sherman. Words of estimative probability. 1964. Block 3, table of synonyms, the ‘Probably not’ entry. https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf (accessed 2026-10-09).

204. Kent, Sherman. Words of estimative probability. 1964. Block 3, table of synonyms, the ‘Almost certain’ entry. https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf (accessed 2026-10-09).

205. Kent, Sherman. Words of estimative probability. 1964. Block 1, ‘Early Brush with Ambiguit’, the paragraph after NIE 29-51. https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf (accessed 2026-10-09).

206. Hashim, M Jawad. Verbal probability terms for communicating clinical risk - a systematic review. 2024. Table 3, Standardized terms for expressing probabilities, row ‘Likely / Probable’. https://pmc.ncbi.nlm.nih.gov/articles/PMC11067312/ (accessed 2026-10-09).

207. Hashim, M Jawad. Verbal probability terms for communicating clinical risk - a systematic review. 2024. Table 3, row ‘Very likely’. https://pmc.ncbi.nlm.nih.gov/articles/PMC11067312/ (accessed 2026-10-09).

208. Wintle, Bonnie C and Fraser, Hannah and Wills, Ben C and Nicholson, Ann E and Fidler, Fiona. Verbal probabilities: very likely to be somewhat more confusing than numbers. 2019. Discussion, paragraph comparing with Budescu et al. https://pmc.ncbi.nlm.nih.gov/articles/PMC6469752/ (accessed 2026-10-09).

209. Wintle, Bonnie C and Fraser, Hannah and Wills, Ben C and Nicholson, Ann E and Fidler, Fiona. Verbal probabilities: very likely to be somewhat more confusing than numbers. 2019. Materials and methods, Participants and procedure. https://pmc.ncbi.nlm.nih.gov/articles/PMC6469752/ (accessed 2026-10-09).

210. Mastrandrea, Michael D and Field, Christopher B and Stocker, Thomas F and Edenhofer, Ottmar and Ebi, Kristie L and Frame, David J and Held, Hermann and Kriegler, Elmar and Mach, Katharine J and Matschoss, Patrick R

and Plattner, Gian-Kasper and Yohe, Gary W and Zwiers, Francis W. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. 2010. Table 1, Likelihood Scale, row 'Likely' (page 3). https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf (accessed 2026-10-09).

211. Mastrandrea, Michael D and Field, Christopher B and Stocker, Thomas F and Edenhofer, Ottmar and Ebi, Kristie L and Frame, David J and Held, Hermann and Kriegler, Elmar and Mach, Katharine J and Matschoss, Patrick R and Plattner, Gian-Kasper and Yohe, Gary W and Zwiers, Francis W. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. 2010. Table 1, row 'Very likely'. https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf (accessed 2026-10-09).

212. Mastrandrea, Michael D and Field, Christopher B and Stocker, Thomas F and Edenhofer, Ottmar and Ebi, Kristie L and Frame, David J and Held, Hermann and Kriegler, Elmar and Mach, Katharine J and Matschoss, Patrick R and Plattner, Gian-Kasper and Yohe, Gary W and Zwiers, Francis W. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. 2010. Table 1, row 'About as likely as not'. https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf (accessed 2026-10-09).

213. Mastrandrea, Michael D and Field, Christopher B and Stocker, Thomas F and Edenhofer, Ottmar and Ebi, Kristie L and Frame, David J and Held, Hermann and Kriegler, Elmar and Mach, Katharine J and Matschoss, Patrick R and Plattner, Gian-Kasper and Yohe, Gary W and Zwiers, Francis W. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. 2010. Paragraph 10, sentences on fuzzy boundaries and the meaning of 'likely'. https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf (accessed 2026-10-09).

214. Mastrandrea, Michael D and Field, Christopher B and Stocker, Thomas F and Edenhofer, Ottmar and Ebi, Kristie L and Frame, David J and Held, Hermann and Kriegler, Elmar and Mach, Katharine J and Matschoss, Patrick R and Plattner, Gian-Kasper and Yohe, Gary W and Zwiers, Francis W. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. 2010. Paragraph 9, first sentence. https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf (accessed 2026-10-09).

215. Mastrandrea, Michael D and Field, Christopher B and Stocker, Thomas F and Edenhofer, Ottmar and Ebi, Kristie L and Frame, David J and Held, Hermann and Kriegler, Elmar and Mach, Katharine J and Matschoss, Patrick R and Plattner, Gian-Kasper and Yohe, Gary W and Zwiers, Francis W. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. 2010. Paragraph 8, first sentence. https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf (accessed 2026-10-09).

216. Mastrandrea, Michael D and Field, Christopher B and Stocker, Thomas F and Edenhofer, Ottmar and Ebi, Kristie L and Frame, David J and Held, Hermann and Kriegler, Elmar and Mach, Katharine J and Matschoss, Patrick R and Plattner, Gian-Kasper and Yohe, Gary W and Zwiers, Francis W. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. 2010. Paragraph 9, near the end. https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf (accessed 2026-10-09).

217. Office of the Director of National Intelligence. Intelligence Community Directive 203: Analytic Standards. 2015. D.6.e.(2)(a), the likelihood table, band row (page 5). https://www.dni.gov/files/documents/ICD/ICD-203.pdf (accessed 2026-10-09).

218. Office of the Director of National Intelligence. Intelligence Community Directive 203: Analytic Standards. 2015. D.6.e.(2)(a), the likelihood table, second row of terms (page 5). https://www.dni.gov/files/documents/ICD/ICD-203.pdf (accessed 2026-10-09).

219. Office of the Director of National Intelligence. Intelligence Community Directive 203: Analytic Standards. 2015. D.6.e.(2)(a), the sentence after the band row. https://www.dni.gov/files/documents/ICD/ICD-203.pdf (accessed 2026-10-09).

220. Office of the Director of National Intelligence. Intelligence Community Directive 203: Analytic Standards. 2015. D.6.e.(2)(b). https://www.dni.gov/files/documents/ICD/ICD-203.pdf (accessed 2026-10-09).

221. Office of the Director of National Intelligence. Intelligence Community Directive 203: Analytic Standards. 2015. D.6.e.(2)(b), opening words. https://www.dni.gov/files/documents/ICD/ICD-203.pdf (accessed 2026-10-09).

222. Office of the Director of National Intelligence. Intelligence Community Directive 203: Analytic Standards. 2015. D.6.e.(2), the paragraph before (a). https://www.dni.gov/files/documents/ICD/ICD-203.pdf (accessed 2026-10-09).

223. Mastrandrea, Michael D and Field, Christopher B and Stocker, Thomas F and Edenhofer, Ottmar and Ebi, Kristie L and Frame, David J and Held, Hermann and Kriegler, Elmar and Mach, Katharine J and Matschoss, Patrick R and Plattner, Gian-Kasper and Yohe, Gary W and Zwiers, Francis W. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. 2010. Table 1, footnote (page 3). https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf (accessed 2026-10-09).

224. Office of the Director of National Intelligence. Intelligence Community Directive 203: Analytic Standards. 2015. F. Effective date, signature line. https://www.dni.gov/files/documents/ICD/ICD-203.pdf (accessed 2026-10-09).

225. Office of the Director of National Intelligence. Intelligence Community Directive 203: Analytic Standards. 2015. Technical amendment page, signature and date. https://www.dni.gov/files/documents/ICD/ICD-203.pdf (accessed 2026-10-09).

226. Wintle, Bonnie C and Fraser, Hannah and Wills, Ben C and Nicholson, Ann E and Fidler, Fiona. Verbal probabilities: very likely to be somewhat more confusing than numbers. 2019. Materials and methods, Analysis, Equation 3. https://pmc.ncbi.nlm.nih.gov/articles/PMC6469752/ (accessed 2026-10-09).

227. Wintle, Bonnie C and Fraser, Hannah and Wills, Ben C and Nicholson, Ann E and Fidler, Fiona. Verbal probabilities: very likely to be somewhat more confusing than numbers. 2019. Results, the paragraph on 'active' participants. https://pmc.ncbi.nlm.nih.gov/articles/PMC6469752/ (accessed 2026-10-09).

228. Wintle, Bonnie C and Fraser, Hannah and Wills, Ben C and Nicholson, Ann E and Fidler, Fiona. Verbal probabilities: very likely to be somewhat more confusing than numbers. 2019. Introduction, paragraph on Budescu et al. https://pmc.ncbi.nlm.nih.gov/articles/PMC6469752/ (accessed 2026-10-09).

229. Mastrandrea, Michael D and Field, Christopher B and Stocker, Thomas F and Edenhofer, Ottmar and Ebi, Kristie L and Frame, David J and Held, Hermann and Kriegler, Elmar and Mach, Katharine J and Matschoss, Patrick R and Plattner, Gian-Kasper and Yohe, Gary W and Zwiers, Francis W. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. 2010. figures quoted in the text. https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf (accessed 2026-10-09).

230. Office of the Director of National Intelligence. Intelligence Community Directive 203: Analytic Standards. 2015. figures quoted in the text. https://www.dni.gov/files/documents/ICD/ICD-203.pdf (accessed 2026-10-09).

231. Mastrandrea, Michael D and Field, Christopher B and Stocker, Thomas F and Edenhofer, Ottmar and Ebi, Kristie L and Frame, David J and Held, Hermann and Kriegler, Elmar and Mach, Katharine J and Matschoss, Patrick R and Plattner, Gian-Kasper and Yohe, Gary W and Zwiers, Francis W. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. 2010. Paragraph 2 (the traceable account). https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf (accessed 2026-10-09).

232. Mastrandrea, Michael D and Field, Christopher B and Stocker, Thomas F and Edenhofer, Ottmar and Ebi, Kristie L and Frame, David J and Held, Hermann and Kriegler, Elmar and Mach, Katharine J and Matschoss, Patrick R and Plattner, Gian-Kasper and Yohe, Gary W and Zwiers, Francis W. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. 2010. Paragraph 3. https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf (accessed 2026-10-09).

233. Office of the Director of National Intelligence. Intelligence Community Directive 203: Analytic Standards. 2015. D.6.e.(2), first sentence. https://www.dni.gov/files/documents/ICD/ICD-203.pdf (accessed 2026-10-09).

234. Office of the Director of National Intelligence. Intelligence Community Directive 203: Analytic Standards. 2015. D.6.e.(2), later in the paragraph. https://www.dni.gov/files/documents/ICD/ICD-203.pdf (accessed 2026-10-09).

235. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. Figure 14.1.b, row for children under 5 years, numbers and absolute effects. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

236. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. Figure 14.1.b, row for children under 5 years, 95 per cent CI of the risk ratio. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

237. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. Figure 14.1.b, first column, follow-up of the incidence outcome. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

238. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. Figure 14.1.b, row for children over 5 years, numbers and absolute effects. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

239. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. Figure 14.1.b, row Adverse events, certainty cell. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

240. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. Abstract, Methods (Europe PMC record; block 1 of the held file): the four types of article. https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

241. Herrera-Perez, Diana and Haslam, Alyson and Crain, Tyler and Gill, Jennifer and Livingston, Catherine and Kaestner, Victoria and Hayes, Michael and Morgan, Dan and Cifu, Adam S and Prasad, Vinay. A comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals. PMID 31182188, PMC6559784. Open access under CC BY 4.0. Quoted from the PMC open-access text. Introduction, second paragraph, the definition. https://pmc.ncbi.nlm.nih.gov/articles/PMC6559784/ (accessed 2026-10-09).

242. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. Abstract, Methods and Results (block 1): the journal, the years and the 2044 and 1344. https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

243. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. Abstract, Results (block 1): the 363 tests of an established practice. https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

244. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. Results, paragraph on articles testing an existing practice (block 2): the 79 inconclusive. https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

245. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. Abstract, Results (block 1): the new practices. https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

246. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. Abstract, Results (block 1): replacement and back to the drawing board. https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

247. Herrera-Perez, Diana and Haslam, Alyson and Crain, Tyler and Gill, Jennifer and Livingston, Catherine and Kaestner, Victoria and Hayes, Michael and Morgan, Dan and Cifu, Adam S and Prasad, Vinay. A comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals. PMID 31182188, PMC6559784. Open access under CC BY 4.0. Quoted from the PMC open-access text. Results, first paragraph: journals, years, 3017 trials and 396 reversals. https://pmc.ncbi.nlm.nih.gov/articles/PMC6559784/ (accessed 2026-10-09).

248. Herrera-Perez, Diana and Haslam, Alyson and Crain, Tyler and Gill, Jennifer and Livingston, Catherine and Kaestner, Victoria and Hayes, Michael and Morgan, Dan and Cifu, Adam S and Prasad, Vinay. A comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals. PMID 31182188, PMC6559784. Open access under CC BY 4.0. Quoted from the PMC open-access text. Results, first paragraph, last sentence. https://pmc.ncbi.nlm.nih.gov/articles/PMC6559784/ (accessed 2026-10-09).

249. Herrera-Perez, Diana and Haslam, Alyson and Crain, Tyler and Gill, Jennifer and Livingston, Catherine and Kaestner, Victoria and Hayes, Michael and Morgan, Dan and Cifu, Adam S and Prasad, Vinay. A comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals. PMID 31182188, PMC6559784. Open access under CC BY 4.0. Quoted from the PMC open-access text. Results, third paragraph: where the reversal trials were done. https://pmc.ncbi.nlm.nih.gov/articles/PMC6559784/ (accessed 2026-10-09).

250. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. Discussion, limitations paragraph: how rarely standard care is tested. https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

251. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. Table 1, by year (block 4; cropped layout extraction checked against printed p. 792). https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

252. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. Table 1, by year (block 4): the 2006 row. https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

253. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. Results, last paragraph: no trend over the decade. https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

254. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. Table 1, by year (block 4): the 2009 row. https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

255. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. Introduction, first paragraph: an earlier review of highly cited studies, as Prasad and colleagues report it. https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

256. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. Introduction, first paragraph: the 7 contradicted. https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

257. Herrera-Perez, Diana and Haslam, Alyson and Crain, Tyler and Gill, Jennifer and Livingston, Catherine and Kaestner, Victoria and Hayes, Michael and Morgan, Dan and Cifu, Adam S and Prasad, Vinay. A comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals. PMID 31182188, PM-

C6559784. Open access under CC BY 4.0. Quoted from the PMC open-access text. Results, third paragraph: type of intervention. https://pmc.ncbi.nlm.nih.gov/articles/PMC6559784/ (accessed 2026-10-09).

258. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. figures quoted in the text. https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

259. Herrera-Perez, Diana and Haslam, Alyson and Crain, Tyler and Gill, Jennifer and Livingston, Catherine and Kaestner, Victoria and Hayes, Michael and Morgan, Dan and Cifu, Adam S and Prasad, Vinay. A comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals. PMID 31182188, PMC6559784. Open access under CC BY 4.0. Quoted from the PMC open-access text. figures quoted in the text. https://pmc.ncbi.nlm.nih.gov/articles/PMC6559784/ (accessed 2026-10-09).

260. Illowsky, Barbara and Dean, Susan and OpenStax. Introductory Statistics 2e. 2023. §3.4 Contingency Tables, opening paragraph. https://openstax.org/details/books/introductory-statistics-2e (accessed 2026-09-23).

261. Illowsky, Barbara and Dean, Susan and OpenStax. Introductory Statistics 2e. 2023. §3.4, worked Problem, solution (e): dividing by the row or column that is given. https://openstax.org/details/books/introductory-statistics-2e (accessed 2026-09-23).

262. H'ajek, Alan. Interpretations of Probability. 2023. §3.3.4, the second item: the rule for updating, named. https://plato.stanford.edu/archives/win2023/entries/probability-interpret/ (accessed 2026-10-09).

263. Morgan, Daniel J and Pineles, Lisa and Owczarzak, Jill and Magder, Larry and Scherer, Laura and Brown, Jessica P and Pfeiffer, Chris and Terndrup, Chris and Leykum, Luci and Feldstein, David and Foy, Andrew and Stevens, Deborah and Koch, Christina and Masnick, Max and Weisenberg, Scott and Korenstein, Deborah. Accuracy of practitioner estimates of probability of diagnosis before and after testing. 2021. Results, Estimates of Disease Probability, last paragraph: the hypothetical test. https://pmc.ncbi.nlm.nih.gov/articles/PMC8022260/ (accessed 2026-10-09).

264. Morgan, Daniel J and Pineles, Lisa and Owczarzak, Jill and Magder, Larry and Scherer, Laura and Brown, Jessica P and Pfeiffer, Chris and Terndrup, Chris and Leykum, Luci and Feldstein, David and Foy, Andrew and Stevens, Deborah and Koch, Christina and Masnick, Max and Weisenberg, Scott and Korenstein, Deborah. Accuracy of practitioner estimates of probability of diagnosis before and after testing. 2021. Results, Estimates of Disease Probability, last paragraph: the answers. https://pmc.ncbi.nlm.nih.gov/articles/PMC8022260/ (accessed 2026-10-09).

265. Morgan, Daniel J and Pineles, Lisa and Owczarzak, Jill and Magder, Larry and Scherer, Laura and Brown, Jessica P and Pfeiffer, Chris and Terndrup, Chris and Leykum, Luci and Feldstein, David and Foy, Andrew and Stevens, Deborah and Koch, Christina and Masnick, Max and Weisenberg, Scott and Korenstein, Deborah. Accuracy of practitioner estimates of probability of diagnosis before and after testing. 2021. Results, Estimates of Disease Probability, last paragraph: the answers for a negative result. https://pmc.ncbi.nlm.nih.gov/articles/PMC8022260/ (accessed 2026-10-09).

266. Morgan, Daniel J and Pineles, Lisa and Owczarzak, Jill and Magder, Larry and Scherer, Laura and Brown, Jessica P and Pfeiffer, Chris and Terndrup, Chris and Leykum, Luci and Feldstein, David and Foy, Andrew and Stevens, Deborah and Koch, Christina and Masnick, Max and Weisenberg, Scott and Korenstein, Deborah. Accuracy of practitioner estimates of probability of diagnosis before and after testing. 2021. Results, Estimates of Disease Probability: cardiac ischemia. https://pmc.ncbi.nlm.nih.gov/articles/PMC8022260/ (accessed 2026-10-09).

267. Morgan, Daniel J and Pineles, Lisa and Owczarzak, Jill and Magder, Larry and Scherer, Laura and Brown, Jessica P and Pfeiffer, Chris and Terndrup, Chris and Leykum, Luci and Feldstein, David and Foy, Andrew and Stevens, Deborah and Koch, Christina and Masnick, Max and Weisenberg, Scott and Korenstein, Deborah. Accuracy of practitioner estimates of probability of diagnosis before and after testing. 2021. Key Points, Findings: who was surveyed. https://pmc.ncbi.nlm.nih.gov/articles/PMC8022260/ (accessed 2026-10-09).

268. Morgan, Daniel J and Pineles, Lisa and Owczarzak, Jill and Magder, Larry and Scherer, Laura and Brown, Jessica P and Pfeiffer, Chris and Terndrup, Chris and Leykum, Luci and Feldstein, David and Foy, Andrew and Stevens, Deborah and Koch, Christina and Masnick, Max and Weisenberg, Scott and Korenstein, Deborah. Accuracy of

practitioner estimates of probability of diagnosis before and after testing. 2021. Abstract, Results. https://pmc.ncbi.nlm.nih.gov/articles/PMC8022260/ (accessed 2026-10-09).

269. Morgan, Daniel J and Pineles, Lisa and Owczarzak, Jill and Magder, Larry and Scherer, Laura and Brown, Jessica P and Pfeiffer, Chris and Terndrup, Chris and Leykum, Luci and Feldstein, David and Foy, Andrew and Stevens, Deborah and Koch, Christina and Masnick, Max and Weisenberg, Scott and Korenstein, Deborah. Accuracy of practitioner estimates of probability of diagnosis before and after testing. 2021. Discussion, first paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC8022260/ (accessed 2026-10-09).

270. Morgan, Daniel J and Pineles, Lisa and Owczarzak, Jill and Magder, Larry and Scherer, Laura and Brown, Jessica P and Pfeiffer, Chris and Terndrup, Chris and Leykum, Luci and Feldstein, David and Foy, Andrew and Stevens, Deborah and Koch, Christina and Masnick, Max and Weisenberg, Scott and Korenstein, Deborah. Accuracy of practitioner estimates of probability of diagnosis before and after testing. 2021. Discussion, fifth paragraph: where the move after a test was too large. https://pmc.ncbi.nlm.nih.gov/articles/PMC8022260/ (accessed 2026-10-09).

271. Morgan, Daniel J and Pineles, Lisa and Owczarzak, Jill and Magder, Larry and Scherer, Laura and Brown, Jessica P and Pfeiffer, Chris and Terndrup, Chris and Leykum, Luci and Feldstein, David and Foy, Andrew and Stevens, Deborah and Koch, Christina and Masnick, Max and Weisenberg, Scott and Korenstein, Deborah. Accuracy of practitioner estimates of probability of diagnosis before and after testing. 2021. Discussion, sixth paragraph, last sentence. https://pmc.ncbi.nlm.nih.gov/articles/PMC8022260/ (accessed 2026-10-09).

272. Morgan, Daniel J and Pineles, Lisa and Owczarzak, Jill and Magder, Larry and Scherer, Laura and Brown, Jessica P and Pfeiffer, Chris and Terndrup, Chris and Leykum, Luci and Feldstein, David and Foy, Andrew and Stevens, Deborah and Koch, Christina and Masnick, Max and Weisenberg, Scott and Korenstein, Deborah. Accuracy of practitioner estimates of probability of diagnosis before and after testing. 2021. figures quoted in the text. https://pmc.ncbi.nlm.nih.gov/articles/PMC8022260/ (accessed 2026-10-09).

273. Morgan, Daniel J and Pineles, Lisa and Owczarzak, Jill and Magder, Larry and Scherer, Laura and Brown, Jessica P and Pfeiffer, Chris and Terndrup, Chris and Leykum, Luci and Feldstein, David and Foy, Andrew and Stevens, Deborah and Koch, Christina and Masnick, Max and Weisenberg, Scott and Korenstein, Deborah. Accuracy of practitioner estimates of probability of diagnosis before and after testing. 2021. Key Points, Findings (block 2). https://pmc.ncbi.nlm.nih.gov/articles/PMC8022260/ (accessed 2026-10-09).

274. Morgan, Daniel J and Pineles, Lisa and Owczarzak, Jill and Magder, Larry and Scherer, Laura and Brown, Jessica P and Pfeiffer, Chris and Terndrup, Chris and Leykum, Luci and Feldstein, David and Foy, Andrew and Stevens, Deborah and Koch, Christina and Masnick, Max and Weisenberg, Scott and Korenstein, Deborah. Accuracy of practitioner estimates of probability of diagnosis before and after testing. 2021. Abstract, Design, Setting, and Participants (block 2). https://pmc.ncbi.nlm.nih.gov/articles/PMC8022260/ (accessed 2026-10-09).

275. Morgan, Daniel J and Pineles, Lisa and Owczarzak, Jill and Magder, Larry and Scherer, Laura and Brown, Jessica P and Pfeiffer, Chris and Terndrup, Chris and Leykum, Luci and Feldstein, David and Foy, Andrew and Stevens, Deborah and Koch, Christina and Masnick, Max and Weisenberg, Scott and Korenstein, Deborah. Accuracy of practitioner estimates of probability of diagnosis before and after testing. 2021. Results, Estimates of Disease Probability, first paragraph (block 3). https://pmc.ncbi.nlm.nih.gov/articles/PMC8022260/ (accessed 2026-10-09).

276. Morgan, Daniel J and Pineles, Lisa and Owczarzak, Jill and Magder, Larry and Scherer, Laura and Brown, Jessica P and Pfeiffer, Chris and Terndrup, Chris and Leykum, Luci and Feldstein, David and Foy, Andrew and Stevens, Deborah and Koch, Christina and Masnick, Max and Weisenberg, Scott and Korenstein, Deborah. Accuracy of practitioner estimates of probability of diagnosis before and after testing. 2021. Discussion, first paragraph (block 3). https://pmc.ncbi.nlm.nih.gov/articles/PMC8022260/ (accessed 2026-10-09).

277. Wintle, Bonnie C and Fraser, Hannah and Wills, Ben C and Nicholson, Ann E and Fidler, Fiona. Verbal probabilities: very likely to be somewhat more confusing than numbers. 2019. Introduction, second paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC6469752/ (accessed 2026-10-09).

278. Wintle, Bonnie C and Fraser, Hannah and Wills, Ben C and Nicholson, Ann E and Fidler, Fiona. Verbal probabilities: very likely to be somewhat more confusing than numbers. 2019. Discussion, the paragraph on negatively worded expressions. https://pmc.ncbi.nlm.nih.gov/articles/PMC6469752/ (accessed 2026-10-09).

279. Mastrandrea, Michael D and Field, Christopher B and Stocker, Thomas F and Edenhofer, Ottmar and Ebi, Kristie L and Frame, David J and Held, Hermann and Kriegler, Elmar and Mach, Katharine J and Matschoss, Patrick R and Plattner, Gian-Kasper and Yohe, Gary W and Zwiers, Francis W. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. 2010. Paragraph 10, first sentence (block 7). https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf (accessed 2026-10-09).

280. Mastrandrea, Michael D and Field, Christopher B and Stocker, Thomas F and Edenhofer, Ottmar and Ebi, Kristie L and Frame, David J and Held, Hermann and Kriegler, Elmar and Mach, Katharine J and Matschoss, Patrick R and Plattner, Gian-Kasper and Yohe, Gary W and Zwiers, Francis W. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. 2010. Paragraph 3, third sentence (block 4). https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf (accessed 2026-10-09).

281. Morgan, Daniel J and Pineles, Lisa and Owczarzak, Jill and Magder, Larry and Scherer, Laura and Brown, Jessica P and Pfeiffer, Chris and Terndrup, Chris and Leykum, Luci and Feldstein, David and Foy, Andrew and Stevens, Deborah and Koch, Christina and Masnick, Max and Weisenberg, Scott and Korenstein, Deborah. Accuracy of practitioner estimates of probability of diagnosis before and after testing. 2021. Discussion, Limitations, last sentence. https://pmc.ncbi.nlm.nih.gov/articles/PMC8022260/ (accessed 2026-10-09).

282. Barseghyan, Hakob and Overgaard, Nicholas and Rupik, Gregory. Introduction to History and Philosophy of Science. 2018. 1.3 Scientific Method, Acceptance, Use, and Pursuit, third paragraph after the definitions. https://human.libretexts.org/Bookshelves/Philosophy/Philosophy_of_Science/Introduction_to_History_and_Philosophy_of_Science_(Barseghyan_Overgaard_and_Rupik) (accessed 2026-10-09).

283. Barseghyan, Hakob and Overgaard, Nicholas and Rupik, Gregory. Introduction to History and Philosophy of Science. 2018. 1.3 Scientific Method, Intro, first paragraph. https://human.libretexts.org/Bookshelves/Philosophy/Philosophy_of_Science/Introduction_to_History_and_Philosophy_of_Science_(Barseghyan_Overgaard_and_Rupik) (accessed 2026-10-09).

284. Rohrer, Julia M and Tierney, Warren and Uhlmann, Eric L and DeBruine, Lisa M and Heyman, Tom and Jones, Benedict and Schmukle, Stefan C and Silberzahn, Raphael and Will'en, Rebecca M and Carlsson, Rickard and Lucas, Richard E and Strand, Julia and Vazire, Simine and Witt, Jessica K and Zentall, Thomas R and Chabris, Christopher F and Yarkoni, Tal. Putting the self in self-correction: findings from the Loss-of-Confidence Project. PMC8564260. Creative Commons Attribution 4.0. Abstract. https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/ (accessed 2026-10-09).

285. Rohrer, Julia M and Tierney, Warren and Uhlmann, Eric L and DeBruine, Lisa M and Heyman, Tom and Jones, Benedict and Schmukle, Stefan C and Silberzahn, Raphael and Will'en, Rebecca M and Carlsson, Rickard and Lucas, Richard E and Strand, Julia and Vazire, Simine and Witt, Jessica K and Zentall, Thomas R and Chabris, Christopher F and Yarkoni, Tal. Putting the self in self-correction: findings from the Loss-of-Confidence Project. PMC8564260. Creative Commons Attribution 4.0. Discussion of the Loss-of-Confidence Statements, Invalid inference (on Zentall's statement). https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/ (accessed 2026-10-09).

286. Rohrer, Julia M and Tierney, Warren and Uhlmann, Eric L and DeBruine, Lisa M and Heyman, Tom and Jones, Benedict and Schmukle, Stefan C and Silberzahn, Raphael and Will'en, Rebecca M and Carlsson, Rickard and Lucas, Richard E and Strand, Julia and Vazire, Simine and Witt, Jessica K and Zentall, Thomas R and Chabris, Christopher F and Yarkoni, Tal. Putting the self in self-correction: findings from the Loss-of-Confidence Project. PMC8564260. Creative Commons Attribution 4.0. General Discussion, Fostering a culture of self-correction. https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/ (accessed 2026-10-09).

287. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Abstract, Background. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

288. Sievert, Katherine and Hussain, Sultana Monira and Page, Matthew J and Wang, Yuanyuan and Hughes, Harrison J and Malek, Mary and Cicuttini, Flavia M. Effect of breakfast on weight and energy intake: systematic review and meta-analysis of randomised controlled trials. PMID 30700403, PMC6352874, PROSPERO CRD42017057687.

Open access under CC BY-NC 4.0. Quoted from the PMC open-access text. Abstract, Conclusion. https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/ (accessed 2026-10-09).

289. Sievert, Katherine and Hussain, Sultana Monira and Page, Matthew J and Wang, Yuanyuan and Hughes, Harrison J and Malek, Mary and Cicuttini, Flavia M. Effect of breakfast on weight and energy intake: systematic review and meta-analysis of randomised controlled trials. PMID 30700403, PMC6352874, PROSPERO CRD42017057687. Open access under CC BY-NC 4.0. Quoted from the PMC open-access text. Results, Fig 3 legend. https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/ (accessed 2026-10-09).

290. Rohrer, Julia M and Tierney, Warren and Uhlmann, Eric L and DeBruine, Lisa M and Heyman, Tom and Jones, Benedict and Schmukle, Stefan C and Silberzahn, Raphael and Will'en, Rebecca M and Carlsson, Rickard and Lucas, Richard E and Strand, Julia and Vazire, Simine and Witt, Jessica K and Zentall, Thomas R and Chabris, Christopher F and Yarkoni, Tal. Putting the self in self-correction: findings from the Loss-of-Confidence Project. PMC8564260. Creative Commons Attribution 4.0. Statement on Chabris and Hamilton (1992), last sentence. https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/ (accessed 2026-10-09).

291. Rohrer, Julia M and Tierney, Warren and Uhlmann, Eric L and DeBruine, Lisa M and Heyman, Tom and Jones, Benedict and Schmukle, Stefan C and Silberzahn, Raphael and Will'en, Rebecca M and Carlsson, Rickard and Lucas, Richard E and Strand, Julia and Vazire, Simine and Witt, Jessica K and Zentall, Thomas R and Chabris, Christopher F and Yarkoni, Tal. Putting the self in self-correction: findings from the Loss-of-Confidence Project. PMC8564260. Creative Commons Attribution 4.0. Statement on Silberzahn and Uhlmann (2013). https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/ (accessed 2026-10-09).

292. Rohrer, Julia M and Tierney, Warren and Uhlmann, Eric L and DeBruine, Lisa M and Heyman, Tom and Jones, Benedict and Schmukle, Stefan C and Silberzahn, Raphael and Will'en, Rebecca M and Carlsson, Rickard and Lucas, Richard E and Strand, Julia and Vazire, Simine and Witt, Jessica K and Zentall, Thomas R and Chabris, Christopher F and Yarkoni, Tal. Putting the self in self-correction: findings from the Loss-of-Confidence Project. PMC8564260. Creative Commons Attribution 4.0. figures quoted in the text. https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/ (accessed 2026-10-09).

293. International Committee of Medical Journal Editors. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals: III.A Corrections and Version Control; III.B Scientific Misconduct, Expressions of Concern, and Retraction. 2026. Recommendations landing page, version line. ICMJE. https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/corrections-and-version-control.html (accessed 2026-10-09).

294. International Committee of Medical Journal Editors. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals: III.A Corrections and Version Control; III.B Scientific Misconduct, Expressions of Concern, and Retraction. 2026. Section III.A Corrections and Version Control, first paragraph. ICMJE. https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/corrections-and-version-control.html (accessed 2026-10-09).

295. International Committee of Medical Journal Editors. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals: III.A Corrections and Version Control; III.B Scientific Misconduct, Expressions of Concern, and Retraction. 2026. Section III.A, first paragraph, third and fourth sentences. ICMJE. https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/corrections-and-version-control.html (accessed 2026-10-09).

296. International Committee of Medical Journal Editors. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals: III.A Corrections and Version Control; III.B Scientific Misconduct, Expressions of Concern, and Retraction. 2026. Section III.A, minimum standards for a correction, first three bullets and the last. ICMJE. https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/corrections-and-version-control.html (accessed 2026-10-09).

297. International Committee of Medical Journal Editors. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals: III.A Corrections and Version Control; III.B Scientific Misconduct, Expressions of Concern, and Retraction. 2026. Section III.A, minimum standards for a correction, last bullet. ICMJE. https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/corrections-and-version-control.html (accessed 2026-10-09).

298. International Committee of Medical Journal Editors. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals: III.A Corrections and Version Control; III.B Scientific Misconduct, Expressions of Concern, and Retraction. 2026. Section III.A, third paragraph (pervasive errors). ICMJE. https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/corrections-and-version-control.html (accessed 2026-10-09).

299. International Committee of Medical Journal Editors. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals: III.A Corrections and Version Control; III.B Scientific Misconduct, Expressions of Concern, and Retraction. 2026. Section III.A, fourth paragraph (retraction, and retraction with republication). ICMJE. https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/corrections-and-version-control.html (accessed 2026-10-09).

300. International Committee of Medical Journal Editors. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals: III.A Corrections and Version Control; III.B Scientific Misconduct, Expressions of Concern, and Retraction. 2026. Section III.B Scientific Misconduct, first paragraph. ICMJE. https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/corrections-and-version-control.html (accessed 2026-10-09).

301. International Committee of Medical Journal Editors. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals: III.A Corrections and Version Control; III.B Scientific Misconduct, Expressions of Concern, and Retraction. 2026. Section III.B, second paragraph. ICMJE. https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/corrections-and-version-control.html (accessed 2026-10-09).

302. COPE Council. COPE Guidelines: Retraction Guidelines. 2025. Web page, Guidance preview, The purpose of retraction. Committee on Publication Ethics. https://publicationethics.org/guidance/guideline/retraction-guidelines (accessed 2026-10-09).

303. COPE Council. COPE Guidelines: Retraction Guidelines. 2025. Web page, Guidance preview, When retraction is not appropriate. Committee on Publication Ethics. https://publicationethics.org/guidance/guideline/retraction-guidelines (accessed 2026-10-09).

304. COPE Council. COPE Guidelines: Retraction Guidelines. 2025. PDF p. 6, Other forms of retraction, Retraction with replacement. Committee on Publication Ethics. https://publicationethics.org/guidance/guideline/retraction-guidelines (accessed 2026-10-09).

305. COPE Council. COPE Guidelines: Retraction Guidelines. 2025. PDF p. 7, Which publications should be retracted?, second paragraph. Committee on Publication Ethics. https://publicationethics.org/guidance/guideline/retraction-guidelines (accessed 2026-10-09).

306. COPE Council. COPE Guidelines: Retraction Guidelines. 2025. Web page, Last reviewed line and citation. Committee on Publication Ethics. https://publicationethics.org/guidance/guideline/retraction-guidelines (accessed 2026-10-09).

307. COPE Council. COPE Guidelines: Retraction Guidelines. 2025. PDF p. 8, Content of retraction notices, Unmistakably identified. Committee on Publication Ethics. https://publicationethics.org/guidance/guideline/retraction-guidelines (accessed 2026-10-09).

308. COPE Council. COPE Guidelines: Retraction Guidelines. 2025. PDF p. 13, Citing retracted articles. Committee on Publication Ethics. https://publicationethics.org/guidance/guideline/retraction-guidelines (accessed 2026-10-09).

309. COPE Council. COPE Guidelines: Retraction Guidelines. 2025. PDF p. 13, Citing retracted articles, last two sentences. Committee on Publication Ethics. https://publicationethics.org/guidance/guideline/retraction-guidelines (accessed 2026-10-09).

310. COPE Council. COPE Guidelines: Retraction Guidelines. 2025. PDF p. 12, Authorship and retracted articles, third paragraph. Committee on Publication Ethics. https://publicationethics.org/guidance/guideline/retraction-guidelines (accessed 2026-10-09).

311. Fang, Ferric C and Steen, R Grant and Casadevall, Arturo. Misconduct accounts for the majority of retracted scientific publications. 2012. Correction, PNAS 2013;110(3):1137, PMC page, first two sentences. https://pmc.ncbi.nlm.nih.gov/articles/PMC3479492/ (accessed 2026-10-09).

312. Fang, Ferric C and Steen, R Grant and Casadevall, Arturo. Misconduct accounts for the majority of retracted scientific publications. 2012. Europe PMC record of Fang et al. 2012, comment and correction list. https://pmc.ncbi.nlm.nih.gov/articles/PMC3479492/ (accessed 2026-10-09).

313. Fang, Ferric C and Steen, R Grant and Casadevall, Arturo. Misconduct accounts for the majority of retracted scientific publications. 2012. Discussion, fourth paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC3479492/ (accessed 2026-10-09).

314. COPE Council. COPE Guidelines: Retraction Guidelines. 2025. PDF p. 8, Content of retraction notices, Objective and factual. Committee on Publication Ethics. https://publicationethics.org/guidance/guideline/retraction-guidelines (accessed 2026-10-09).

315. Fang, Ferric C and Steen, R Grant and Casadevall, Arturo. Misconduct accounts for the majority of retracted scientific publications. 2012. Discussion, second paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC3479492/ (accessed 2026-10-09).

316. Fang, Ferric C and Steen, R Grant and Casadevall, Arturo. Misconduct accounts for the majority of retracted scientific publications. 2012. Methods, first sentence. https://pmc.ncbi.nlm.nih.gov/articles/PMC3479492/ (accessed 2026-10-09).

317. Fang, Ferric C and Steen, R Grant and Casadevall, Arturo. Misconduct accounts for the majority of retracted scientific publications. 2012. figures quoted in the text. https://pmc.ncbi.nlm.nih.gov/articles/PMC3479492/ (accessed 2026-10-09).

318. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. Abstract (Europe PMC record), Results. https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

319. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. Abstract (Europe PMC record), Methods. https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

320. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. Methods, first paragraph (supplied PDF, p. 791). https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

321. Rohrer, Julia M and Tierney, Warren and Uhlmann, Eric L and DeBruine, Lisa M and Heyman, Tom and Jones, Benedict and Schmukle, Stefan C and Silberzahn, Raphael and Will'en, Rebecca M and Carlsson, Rickard and Lucas, Richard E and Strand, Julia and Vazire, Simine and Witt, Jessica K and Zentall, Thomas R and Chabris, Christopher F and Yarkoni, Tal. Putting the self in self-correction: findings from the Loss-of-Confidence Project. PMC8564260. Creative Commons Attribution 4.0. Fostering a culture of self-correction, Different media for self-correction, second paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/ (accessed 2026-10-09).

322. COPE Council. COPE Guidelines: Retraction Guidelines. 2025. Web page, Key points, last point. Committee on Publication Ethics. https://publicationethics.org/guidance/guideline/retraction-guidelines (accessed 2026-10-09).

323. Herrera-Perez, Diana and Haslam, Alyson and Crain, Tyler and Gill, Jennifer and Livingston, Catherine and Kaestner, Victoria and Hayes, Michael and Morgan, Dan and Cifu, Adam S and Prasad, Vinay. A comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals. PMID 31182188, PMC6559784. Open access under CC BY 4.0. Quoted from the PMC open-access text. Abstract. https://pmc.ncbi.nlm.nih.gov/articles/PMC6559784/ (accessed 2026-10-09).

324. Lu, Susan Feng and Jin, Ginger Zhe and Uzzi, Brian and Jones, Benjamin. The retraction penalty: evidence from the Web of Science. PMID 24192909, PMC3818648. Creative Commons Attribution 3.0 Unported. Figure 1 legend, (b). https://pmc.ncbi.nlm.nih.gov/articles/PMC3818648/ (accessed 2026-10-09).

325. Fang, Ferric C and Steen, R Grant and Casadevall, Arturo. Misconduct accounts for the majority of retracted scientific publications. 2012. Abstract. https://pmc.ncbi.nlm.nih.gov/articles/PMC3479492/ (accessed 2026-10-09).

326. Fang, Ferric C and Steen, R Grant and Casadevall, Arturo. Misconduct accounts for the majority of retracted scientific publications. 2012. Results, Causes of Retraction, last sentences. https://pmc.ncbi.nlm.nih.gov/articles/PMC3479492/ (accessed 2026-10-09).

327. Fang, Ferric C and Steen, R Grant and Casadevall, Arturo. Misconduct accounts for the majority of retracted scientific publications. 2012. Results, Causes of Retraction, the sentence on reclassification. https://pmc.ncbi.nlm.nih.gov/articles/PMC3479492/ (accessed 2026-10-09).

328. Fang, Ferric C and Steen, R Grant and Casadevall, Arturo. Misconduct accounts for the majority of retracted scientific publications. 2012. Results, Causes of Retraction, the FOXO1a example. https://pmc.ncbi.nlm.nih.gov/articles/PMC3479492/ (accessed 2026-10-09).

329. Fang, Ferric C and Steen, R Grant and Casadevall, Arturo. Misconduct accounts for the majority of retracted scientific publications. 2012. Discussion, last paragraph but one. https://pmc.ncbi.nlm.nih.gov/articles/PMC3479492/ (accessed 2026-10-09).

330. Lu, Susan Feng and Jin, Ginger Zhe and Uzzi, Brian and Jones, Benjamin. The retraction penalty: evidence from the Web of Science. PMID 24192909, PMC3818648. Creative Commons Attribution 3.0 Unported. Abstract. https://pmc.ncbi.nlm.nih.gov/articles/PMC3818648/ (accessed 2026-10-09).

331. Lu, Susan Feng and Jin, Ginger Zhe and Uzzi, Brian and Jones, Benjamin. The retraction penalty: evidence from the Web of Science. PMID 24192909, PMC3818648. Creative Commons Attribution 3.0 Unported. Results, the paragraph after the description of Figure 3. https://pmc.ncbi.nlm.nih.gov/articles/PMC3818648/ (accessed 2026-10-09).

332. Lu, Susan Feng and Jin, Ginger Zhe and Uzzi, Brian and Jones, Benjamin. The retraction penalty: evidence from the Web of Science. PMID 24192909, PMC3818648. Creative Commons Attribution 3.0 Unported. Results, the same paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC3818648/ (accessed 2026-10-09).

333. Lu, Susan Feng and Jin, Ginger Zhe and Uzzi, Brian and Jones, Benjamin. The retraction penalty: evidence from the Web of Science. PMID 24192909, PMC3818648. Creative Commons Attribution 3.0 Unported. Results, first paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC3818648/ (accessed 2026-10-09).

334. Lu, Susan Feng and Jin, Ginger Zhe and Uzzi, Brian and Jones, Benjamin. The retraction penalty: evidence from the Web of Science. PMID 24192909, PMC3818648. Creative Commons Attribution 3.0 Unported. Discussion, first paragraph, last sentence. https://pmc.ncbi.nlm.nih.gov/articles/PMC3818648/ (accessed 2026-10-09).

335. Rohrer, Julia M and Tierney, Warren and Uhlmann, Eric L and DeBruine, Lisa M and Heyman, Tom and Jones, Benedict and Schmukle, Stefan C and Silberzahn, Raphael and Will'en, Rebecca M and Carlsson, Rickard and Lucas, Richard E and Strand, Julia and Vazire, Simine and Witt, Jessica K and Zentall, Thomas R and Chabris, Christopher F and Yarkoni, Tal. Putting the self in self-correction: findings from the Loss-of-Confidence Project. PMC8564260. Creative Commons Attribution 4.0. Results and Discussion of the Anonymous Online Survey, first two paragraphs. https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/ (accessed 2026-10-09).

336. Rohrer, Julia M and Tierney, Warren and Uhlmann, Eric L and DeBruine, Lisa M and Heyman, Tom and Jones, Benedict and Schmukle, Stefan C and Silberzahn, Raphael and Will'en, Rebecca M and Carlsson, Rickard and Lucas, Richard E and Strand, Julia and Vazire, Simine and Witt, Jessica K and Zentall, Thomas R and Chabris, Christopher F and Yarkoni, Tal. Putting the self in self-correction: findings from the Loss-of-Confidence Project. PMC8564260. Creative Commons Attribution 4.0. Results and Discussion of the Anonymous Online Survey, first paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/ (accessed 2026-10-09).

337. Rohrer, Julia M and Tierney, Warren and Uhlmann, Eric L and DeBruine, Lisa M and Heyman, Tom and Jones, Benedict and Schmukle, Stefan C and Silberzahn, Raphael and Will'en, Rebecca M and Carlsson, Rickard and Lucas, Richard E and Strand, Julia and Vazire, Simine and Witt, Jessica K and Zentall, Thomas R and Chabris, Christopher F and Yarkoni, Tal. Putting the self in self-correction: findings from the Loss-of-Confidence Project. PM-

C8564260. Creative Commons Attribution 4.0. Results and Discussion of the Anonymous Online Survey, third paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/ (accessed 2026-10-09).

338. Rohrer, Julia M and Tierney, Warren and Uhlmann, Eric L and DeBruine, Lisa M and Heyman, Tom and Jones, Benedict and Schmukle, Stefan C and Silberzahn, Raphael and Will'en, Rebecca M and Carlsson, Rickard and Lucas, Richard E and Strand, Julia and Vazire, Simine and Witt, Jessica K and Zentall, Thomas R and Chabris, Christopher F and Yarkoni, Tal. Putting the self in self-correction: findings from the Loss-of-Confidence Project. PMC8564260. Creative Commons Attribution 4.0. Results and Discussion of the Anonymous Online Survey, fifth paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/ (accessed 2026-10-09).

339. COPE Council. COPE Guidelines: Retraction Guidelines. 2025. Definitions box, paper mill (supplied PDF, p. 11). Committee on Publication Ethics. https://publicationethics.org/guidance/guideline/retraction-guidelines (accessed 2026-10-09).

340. Fang, Ferric C and Steen, R Grant and Casadevall, Arturo. Misconduct accounts for the majority of retracted scientific publications. 2012. Discussion, first paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC3479492/ (accessed 2026-10-09).

341. Rohrer, Julia M and Tierney, Warren and Uhlmann, Eric L and DeBruine, Lisa M and Heyman, Tom and Jones, Benedict and Schmukle, Stefan C and Silberzahn, Raphael and Will'en, Rebecca M and Carlsson, Rickard and Lucas, Richard E and Strand, Julia and Vazire, Simine and Witt, Jessica K and Zentall, Thomas R and Chabris, Christopher F and Yarkoni, Tal. Putting the self in self-correction: findings from the Loss-of-Confidence Project. PMC8564260. Creative Commons Attribution 4.0. Methods of the Anonymous Online Survey. https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/ (accessed 2026-10-09).

342. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. Results (supplied PDF). https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

343. Lu, Susan Feng and Jin, Ginger Zhe and Uzzi, Brian and Jones, Benjamin. The retraction penalty: evidence from the Web of Science. PMID 24192909, PMC3818648. Creative Commons Attribution 3.0 Unported. figures quoted in the text. https://pmc.ncbi.nlm.nih.gov/articles/PMC3818648/ (accessed 2026-10-09).

344. Barseghyan, Hakob and Overgaard, Nicholas and Rupik, Gregory. Introduction to History and Philosophy of Science. 2018. Section 1.3 Acceptance, use and pursuit, paragraph on acceptance. https://human.libretexts.org/Bookshelves/Philosophy/Philosophy_of_Science/Introduction_to_History_and_Philosophy_of_Science_(Barseghyan_Overgaard_and_Rupik) (accessed 2026-10-09).

345. Barseghyan, Hakob and Overgaard, Nicholas and Rupik, Gregory. Introduction to History and Philosophy of Science. 2018. Section 1.3, paragraph on acceptance, last sentence. https://human.libretexts.org/Bookshelves/Philosophy/Philosophy_of_Science/Introduction_to_History_and_Philosophy_of_Science_(Barseghyan_Overgaard_and_Rupik) (accessed 2026-10-09).

346. COPE Council. COPE Guidelines: Retraction Guidelines. 2025. PDF p. 12, Authorship and retracted articles, last sentences. Committee on Publication Ethics. https://publicationethics.org/guidance/guideline/retraction-guidelines (accessed 2026-10-09).

347. Rohrer, Julia M and Tierney, Warren and Uhlmann, Eric L and DeBruine, Lisa M and Heyman, Tom and Jones, Benedict and Schmukle, Stefan C and Silberzahn, Raphael and Will'en, Rebecca M and Carlsson, Rickard and Lucas, Richard E and Strand, Julia and Vazire, Simine and Witt, Jessica K and Zentall, Thomas R and Chabris, Christopher F and Yarkoni, Tal. Putting the self in self-correction: findings from the Loss-of-Confidence Project. PMC8564260. Creative Commons Attribution 4.0. Different media for self-correction, first paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/ (accessed 2026-10-09).

348. Rohrer, Julia M and Tierney, Warren and Uhlmann, Eric L and DeBruine, Lisa M and Heyman, Tom and Jones, Benedict and Schmukle, Stefan C and Silberzahn, Raphael and Will'en, Rebecca M and Carlsson, Rickard and Lucas, Richard E and Strand, Julia and Vazire, Simine and Witt, Jessica K and Zentall, Thomas R and Chabris, Christopher F and Yarkoni, Tal. Putting the self in self-correction: findings from the Loss-of-Confidence Project. PM-

C8564260. Creative Commons Attribution 4.0. Different media for self-correction, fifth paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/ (accessed 2026-10-09).

349. Rohrer, Julia M and Tierney, Warren and Uhlmann, Eric L and DeBruine, Lisa M and Heyman, Tom and Jones, Benedict and Schmukle, Stefan C and Silberzahn, Raphael and Will'en, Rebecca M and Carlsson, Rickard and Lucas, Richard E and Strand, Julia and Vazire, Simine and Witt, Jessica K and Zentall, Thomas R and Chabris, Christopher F and Yarkoni, Tal. Putting the self in self-correction: findings from the Loss-of-Confidence Project. PMC8564260. Creative Commons Attribution 4.0. Different media for self-correction, last paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/ (accessed 2026-10-09).

350. Rohrer, Julia M and Tierney, Warren and Uhlmann, Eric L and DeBruine, Lisa M and Heyman, Tom and Jones, Benedict and Schmukle, Stefan C and Silberzahn, Raphael and Will'en, Rebecca M and Carlsson, Rickard and Lucas, Richard E and Strand, Julia and Vazire, Simine and Witt, Jessica K and Zentall, Thomas R and Chabris, Christopher F and Yarkoni, Tal. Putting the self in self-correction: findings from the Loss-of-Confidence Project. PMC8564260. Creative Commons Attribution 4.0. Fostering a culture of self-correction, second paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/ (accessed 2026-10-09).

351. Rohrer, Julia M and Tierney, Warren and Uhlmann, Eric L and DeBruine, Lisa M and Heyman, Tom and Jones, Benedict and Schmukle, Stefan C and Silberzahn, Raphael and Will'en, Rebecca M and Carlsson, Rickard and Lucas, Richard E and Strand, Julia and Vazire, Simine and Witt, Jessica K and Zentall, Thomas R and Chabris, Christopher F and Yarkoni, Tal. Putting the self in self-correction: findings from the Loss-of-Confidence Project. PMC8564260. Creative Commons Attribution 4.0. Statement on Chabris and Hamilton (1992) by Christopher F. Chabris, end of first paragraph. https://pmc.ncbi.nlm.nih.gov/articles/PMC8564260/ (accessed 2026-10-09).

352. Sievert, Katherine and Hussain, Sultana Monira and Page, Matthew J and Wang, Yuanyuan and Hughes, Harrison J and Malek, Mary and Cicuttini, Flavia M. Effect of breakfast on weight and energy intake: systematic review and meta-analysis of randomised controlled trials. PMID 30700403, PMC6352874, PROSPERO CRD42017057687. Open access under CC BY-NC 4.0. Quoted from the PMC open-access text. Abstract, Results. https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/ (accessed 2026-10-09).

353. Sievert, Katherine and Hussain, Sultana Monira and Page, Matthew J and Wang, Yuanyuan and Hughes, Harrison J and Malek, Mary and Cicuttini, Flavia M. Effect of breakfast on weight and energy intake: systematic review and meta-analysis of randomised controlled trials. PMID 30700403, PMC6352874, PROSPERO CRD42017057687. Open access under CC BY-NC 4.0. Quoted from the PMC open-access text. Abstract, Results, last sentences. https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/ (accessed 2026-10-09).

354. Sievert, Katherine and Hussain, Sultana Monira and Page, Matthew J and Wang, Yuanyuan and Hughes, Harrison J and Malek, Mary and Cicuttini, Flavia M. Effect of breakfast on weight and energy intake: systematic review and meta-analysis of randomised controlled trials. PMID 30700403, PMC6352874, PROSPERO CRD42017057687. Open access under CC BY-NC 4.0. Quoted from the PMC open-access text. Objective. https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/ (accessed 2026-10-09).

355. Office of the Director of National Intelligence. Intelligence Community Directive 203: Analytic Standards. 2015. Section D.6.e.(2)(a), page 5, the band row of the likelihood table. https://www.dni.gov/files/documents/ICD/ICD-203.pdf (accessed 2026-10-09).

356. International Committee of Medical Journal Editors. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals: III.A Corrections and Version Control; III.B Scientific Misconduct, Expressions of Concern, and Retraction. 2026. Section III.A, retraction with republication paragraph. ICMJE. https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/corrections-and-version-control.html (accessed 2026-10-09).

357. Prasad, Vinay and Vandross, Andrae and Toomey, Caitlin and Cheung, Michael and Rho, Jason and Quinn, Steven and Chacko, Satish Jacob and Borkar, Durga and Gall, Victor and Selvaraj, Senthil and Ho, Nancy and Cifu, Adam. A decade of reversal: an analysis of 146 contradicted medical practices. 2013. Abstract, Results (Europe PMC record), last sentence. https://pubmed.ncbi.nlm.nih.gov/23871230/ (accessed 2026-10-09).

358. Mastrandrea, Michael D and Field, Christopher B and Stocker, Thomas F and Edenhofer, Ottmar and Ebi, Kristie L and Frame, David J and Held, Hermann and Kriegler, Elmar and Mach, Katharine J and Matschoss, Patrick R

and Plattner, Gian-Kasper and Yohe, Gary W and Zwiers, Francis W. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. 2010. Paragraph 3 (group overconfidence and anchoring). https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf (accessed 2026-10-09).

359. Office of the Director of National Intelligence. Intelligence Community Directive 203: Analytic Standards. 2015. Section D.6.e.(2), page 5, the paragraph before the likelihood table. https://www.dni.gov/files/documents/ICD/ICD-203.pdf (accessed 2026-10-09).

360. Illowsky, Barbara and Dean, Susan and OpenStax. Introductory Statistics 2e. 2023. §3.4 Contingency Tables, first sentence. https://openstax.org/details/books/introductory-statistics-2e (accessed 2026-09-23).

361. Dhurandhar, Emily J and Dawson, John and Alcorn, Amy and Larsen, Lesli H and Thomas, Elizabeth A and Cardel, Michelle and Bourland, Ashley C and Astrup, Arne and St-Onge, Marie-Pierre and Hill, James O and Apovian, Caroline M and Shikany, James M and Allison, David B. The effectiveness of breakfast recommendations on weight loss: a randomized controlled trial. 2014. Abstract, Design. https://pmc.ncbi.nlm.nih.gov/articles/PMC4095657/ (accessed 2026-10-09).

362. Sievert, Katherine and Hussain, Sultana Monira and Page, Matthew J and Wang, Yuanyuan and Hughes, Harrison J and Malek, Mary and Cicuttini, Flavia M. Effect of breakfast on weight and energy intake: systematic review and meta-analysis of randomised controlled trials. PMID 30700403, PMC6352874, PROSPERO CRD42017057687. Open access under CC BY-NC 4.0. Quoted from the PMC open-access text. Abstract, Eligibility criteria for selecting studies. https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/ (accessed 2026-10-09).

363. Sievert, Katherine and Hussain, Sultana Monira and Page, Matthew J and Wang, Yuanyuan and Hughes, Harrison J and Malek, Mary and Cicuttini, Flavia M. Effect of breakfast on weight and energy intake: systematic review and meta-analysis of randomised controlled trials. PMID 30700403, PMC6352874, PROSPERO CRD42017057687. Open access under CC BY-NC 4.0. Quoted from the PMC open-access text. Discussion, Quality of evidence, first two sentences. https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/ (accessed 2026-10-09).

364. Sievert, Katherine and Hussain, Sultana Monira and Page, Matthew J and Wang, Yuanyuan and Hughes, Harrison J and Malek, Mary and Cicuttini, Flavia M. Effect of breakfast on weight and energy intake: systematic review and meta-analysis of randomised controlled trials. PMID 30700403, PMC6352874, PROSPERO CRD42017057687. Open access under CC BY-NC 4.0. Quoted from the PMC open-access text. Results, Table 2 (general characteristics of included studies), Betts row. https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/ (accessed 2026-10-09).

365. Sievert, Katherine and Hussain, Sultana Monira and Page, Matthew J and Wang, Yuanyuan and Hughes, Harrison J and Malek, Mary and Cicuttini, Flavia M. Effect of breakfast on weight and energy intake: systematic review and meta-analysis of randomised controlled trials. PMID 30700403, PMC6352874, PROSPERO CRD42017057687. Open access under CC BY-NC 4.0. Quoted from the PMC open-access text. Results, Table 2 (general characteristics of included studies), Dhurandhar row. https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/ (accessed 2026-10-09).

366. Brown, Andrew W and Bohan Brown, Michelle M and Allison, David B. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. 2013. Results, The PEBO is widely stated as true. https://pmc.ncbi.nlm.nih.gov/articles/PMC3798081/ (accessed 2026-10-09).

367. Sch"unemann, Holger J and Higgins, Julian PT and Vist, Gunn E and Glasziou, Paul and Akl, Elie A and Skoetz, Nicole and Guyatt, Gordon H. Chapter 14: Completing `Summary of findings' tables and grading the certainty of the evidence. 2024. 14.2.2 (3) Indirectness of evidence, second type. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14 (accessed 2026-10-09).

368. Sievert, Katherine and Hussain, Sultana Monira and Page, Matthew J and Wang, Yuanyuan and Hughes, Harrison J and Malek, Mary and Cicuttini, Flavia M. Effect of breakfast on weight and energy intake: systematic review and meta-analysis of randomised controlled trials. PMID 30700403, PMC6352874, PROSPERO CRD42017057687. Open access under CC BY-NC 4.0. Quoted from the PMC open-access text. Discussion, the paragraph on high-income countries. https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/ (accessed 2026-10-09).

369. Sievert, Katherine and Hussain, Sultana Monira and Page, Matthew J and Wang, Yuanyuan and Hughes, Harrison J and Malek, Mary and Cicuttini, Flavia M. Effect of breakfast on weight and energy intake: systematic review and meta-analysis of randomised controlled trials. PMID 30700403, PMC6352874, PROSPERO CRD42017057687.

Open access under CC BY-NC 4.0. Quoted from the PMC open-access text. Methods, Eligibility criteria. https://pmc.ncbi.nlm.nih.gov/articles/PMC6352874/ (accessed 2026-10-09).

Appendix · Worked answers

1 · Three kinds of statement: a hypothesis, a finding and a belief

EXERCISE 1. A hypothesis is a proposed explanation, stated so that a study can be designed to support or refute it. A finding is what a particular study observed, in its own people and setting, with its uncertainty. A belief is what a person holds to be true, held with some strength, on whatever grounds they have.

Kent's three kinds. First, a statement of something known, as close to indisputable fact as one can come. Second, a judgement on evidence, which the evidence makes likely but cannot make certain. Third, a judgement made almost without evidence.

EXERCISE 2.

1. A belief, dressed as a finding. "Studies prove" names no study, and no study can prove. Ask: which studies, in which children, and did they show that skipping caused the weight, or only that the two go together?
2. A hypothesis. It is offered to be tested, and it names what to compare. Ask: what result would count against it? For example, no difference in the share with overweight between the two groups of schools.
3. A finding: a count from one survey, in its own students. In Kent's terms it is close to his first kind. Ask: who were the 412, and how were they chosen?
4. A belief. "Surely" carries the strength; nothing in the sentence carries grounds. Ask: what is it resting on? Then ask what else has changed in the district at the same time.

EXERCISE 3. Draw three columns: hypothesis, finding, belief. Ask her which column her sentence belongs in.

It is written as a finding, before there are any data. That makes it a belief: it states the answer and gives no grounds. "Prove" makes it worse, because no study proves anything. A study supports or counts against.

Rewrite it with her as a hypothesis, something to test:

> *We will test whether students who skip breakfast on most days have a higher share with overweight than students who eat it.*

Then ask her for the result that would count against it. Write that under the hypothesis.

Last, tell her what her sentence may become if the study goes well: a finding, about her students, with its uncertainty. Whether it can ever show cause depends on the design. The book comes to that, and she does not need it yet.

EXERCISE 4. Because the evidence is circumstantial. The aircraft are military and the buildings look like other bases, which makes the case, in Kent's words, "say, 90 percent of the way". It "cannot justify a flat assertion". The qualifier tells the listener the statement is a judgement, not a count.

For your own sentences, look for any claim that reached beyond what was counted. A typical one says what a treatment does, built from studies that only watched people. Each of those was a judgement on evidence. Spoken flat, it sounded like a fact.

2 · A hypothesis is testable: the prediction, and the result that would count against it

EXERCISE 1. A hypothesis is testable when a prediction follows from it: "if it is true, then this result". Some possible result, named in advance, must count against it.

A matching result does not prove it because the argument "if the hypothesis, then the result; the result; so the hypothesis" does not carry. Other explanations can predict the same result.

Accepted: taken as the best available description. Used: taken as an adequate tool for a practical job. Pursued: thought worth developing and testing further.

EXERCISE 2. The second sentence is the escape. Whatever a study finds, the claim survives: if skippers gain weight, breakfast was the cause; if they do not, snacks hid the cause. No result is allowed to count against it, so it predicts nothing.

Here is a testable version. The snacks become part of the test rather than an excuse.

> *Adults told to skip breakfast gain more weight over 16 weeks than adults told to eat it, whatever they eat later in the day.*

Prediction: if this is true, then at 16 weeks the group told to skip has gained more weight, on average, than the group told to eat. Result that would count against it: the groups change by about the same, or the group told to eat breakfast gains more.

"Switches on your metabolism" is a proposed mechanism, a different claim. It needs its own test.

EXERCISE 3. Check your four lines this way.

- The hypothesis names who, what, compared with what, and over how long.
- The prediction is something you could see in your data, not a feeling.
- The result against it is specific enough that you would recognise it in your tables. "No difference" needs a definition: no difference in what, measured how.
- Your stance is almost certainly "pursued". If you wrote "accepted", ask what you will do with a result against it. If the honest answer is "explain it away", rewrite the stance.

EXERCISE 4. No. The trial tested a clear prediction and the prediction failed. Adults trying to lose weight who were told to switch their breakfast habit did not lose clearly more weight than those who were not. That counts against the idea that breakfast advice helps weight loss, in adults like these, over 16 weeks.

It did not prove that breakfast never matters. The trial was in the United States and Denmark, in adults, for 16 weeks. The authors themselves say a very small effect could need longer to show.

Quotable sentence:

> *In a 16-week trial, advice to eat or skip breakfast made no clear difference to weight loss in adults trying to lose weight.*

3 · One finding, several explanations

EXERCISE 1. Chance, bias, a third factor, reverse direction and the proposed cause.

Repeating the same design makes chance weaker, because a pattern made by luck would come and go. It leaves standing every rival the design shares. For observational studies these are usually a third factor, reverse direction and any bias they have in common.

EXERCISE 2.

- Chance: how many students were surveyed? A small survey can show a pattern by luck. Has anyone else found the same?
- Bias: how were students chosen, and was weight measured or self-reported? A convenience sample, or weights that some groups under-report, can push the result one way.
- A third factor: what goes with both late sleep and weight? Candidates to check include late-night eating, screen time, hostel life and exam stress. Each is a question, not a fact.
- Reverse direction: the survey measured both at one visit. Could weight, or something linked to it, affect sleep timing?
- The proposed cause: late sleep putting weight on. It stays possible.

The conclusion should read "late sleep was associated with overweight in these students". A claim about cause needs a design in which the order of events is known. It also needs the third factors made even between the groups.

EXERCISE 3. A third factor: largely removed. Chance decides who is told to eat breakfast, so routines, sleep and everything else end up spread between the groups, apart from chance differences.

Reverse direction: removed. The advice comes first and the weight change after it.

Chance: not removed. Any one trial can mislead by luck, so a single trial's result is still a finding, and trials are repeated.

Bias: not removed by randomisation on its own. If weight is measured badly, or many people drop out of one group, the trial can still be pushed one way.

The proposed cause: that is what the trial is testing.

EXERCISE 4. Answer: no, not as worded. The studies show that people who skip breakfast more often have overweight. They do not show that changing breakfast changes weight, which is what the campaign would claim.

Why: every one of those studies watched people as they were. Something else that goes with eating breakfast, such as a regular routine, could explain the pattern. Or weight could change breakfast habits rather than the reverse. More studies of the same kind do not settle either.

What would settle it: trials in which people are assigned, by chance, to advice to eat breakfast or not, and their weight is followed.

What to ask the note's authors: which trials, in which people, and what they found.

4 · Belief beyond the evidence

EXERCISE 1. About their own results:

1. A biased reading of their own results.
2. Causal language to describe their own observational finding.

About other people's work:

1. Citing another study's result in a misleading way.
2. Causal language when citing another study's association.

EXERCISE 2. First citation. Schlundt and colleagues ran a 12-week trial in women on a weight-loss plan. It found no main effect of breakfast on weight loss. Women who changed their usual habit, either way, lost more, and that hint was weak ($P < 0.06$). The slide cites a trial for something it did not find. Brown counted this exact error: 29 of 46 papers citing the trial about weight misread it, 28 of them towards breakfast.

Second citation. In the registry, 78 per cent reported eating breakfast every day. With no comparison group, that shows only that breakfast and kept-off weight occur together in these people. It is not even an association, because nobody without weight loss was measured. "Causes" and "helps people keep weight off" read a cause into a description.

A fair slide:

> *Many observational studies find that people who skip breakfast more often have overweight or obesity. This is an association. In one trial (Schlundt 1992), breakfast had no main effect on weight loss.*

EXERCISE 3. "Co-occurring" claims only that both are present in the same people. That is all a case series with no comparison group can show, so it is the right wording.

"Associated" claims that the two go together more than they would by chance. That needs a comparison: people without the kept-off weight loss, to see whether they eat breakfast less often. The registry has none.

A causal statement claims that breakfast helped keep the weight off. That needs a comparison in which something other than the people's own habits decided who ate breakfast.

So only 29 per cent of these 72 papers matched the study's design. Brown counts the "associated" and causal wordings together, 48 per cent of the citations, as claiming more than was warranted.

EXERCISE 4. A quotable sentence: "People who skip breakfast tend to weigh more. But when a trial told people to eat or skip it, the advice made no clear difference to how much weight they lost."

How sure, in words: fairly sure that the link seen in surveys is real; not convinced that breakfast itself changes weight.

Why: the link comes from many observational studies. They watch what people already do. They cannot tell whether breakfast, or something else about people who eat it, is behind the link. The largest single trial, in adults in the United States and Denmark trying to lose weight, found no detectable effect of the advice over 16 weeks. A 2019 review of the trials, which a later section reads, points the same way.

What would change your mind: longer trials, in other groups, showing a difference.

Leave out "myth". The evidence says the cause is unproven, not that it is disproved.

5 · Two kinds of support for a causal claim: a difference between groups, and a mechanism

EXERCISE 1. The first is a difference between groups. Compare groups that differ in the exposure, and see whether the outcome differs. It is best when chance decided the groups.

The second is a mechanism. Find evidence for each step of how the exposure leads to the outcome.

A claim cannot do without the first. With a mechanism alone, however good, it stays a hypothesis.

The third kind is parallel evidence: other studies that find the same thing.

EXERCISE 2. The slide offers a mechanism. Its chain runs from eating late, to a change in how the body handles food at night, to more fat stored, to weight. Each link is a claim that needs its own evidence, and the slide gives none for any of them.

It leaves out any difference between groups. Do people who eat dinner late gain more weight than people who eat early? If they do, ask how that was found. In an observational comparison, late eaters may differ in sleep, shift work or income. In a trial, chance decides the timing.

An honest slide: "Eating late may change how the body handles food. Whether it changes weight has to be shown by comparing groups. That is a hypothesis to test, not yet a finding."

EXERCISE 3. One sound design. Adults who agree to take part are split by a chance process, such as sealed numbered envelopes, into two groups. One group is asked to finish dinner by an early set time, the other to eat it late, with no other change to advice. Weight is measured at the start and after several months, by staff who do not know the group.

Chance deciding the group is the point: it spreads sleep, work hours and habits between the groups by chance, not by choice.

Ways it could still mislead (any one):

- People may not follow the timing. Record whether they did.
- People may eat more or less overall when told to change the timing. Then a difference in weight could come from the amount, not the timing.
- A short trial may miss a slow effect. The breakfast trial's own authors named its 16 weeks as a limit.
- A small trial can be unbalanced by bad luck. Check the groups' starting values.

EXERCISE 4. Draw two columns: "difference between groups" and "mechanism".

Under mechanism, write his chain: breakfast, less hunger, less at lunch, less over the day, lower weight. Then mark the links. Less at lunch: some short studies support it. Less over the whole day: none of the studies Brown and colleagues collected showed it, and two found more. The chain breaks at that link.

Under difference between groups, write two lines. Observational studies: breakfast skippers more often have overweight or obesity, but skippers may differ in other ways. The 2014 trial, where chance decided the advice: no detectable difference in weight over 16 weeks.

Then ask him which leg is weak. Both. His sentence is a hypothesis, and a fair one to test. It is not something to tell patients as fact.

6 · Why evidence comes in strengths

EXERCISE 1. Design: were the groups made by chance, or watched as they came?

Risk of bias: do flaws in how the studies were run tilt the answer?

Inconsistency: do the studies disagree, with no reason found?

Indirectness: do they answer a nearby question? Other people, another form of the treatment, another comparison, or a stand-in outcome all count.

Imprecision: were there few people or few events, so the range is wide?

Publication bias: are studies missing because nobody published them?

The two that can raise trust: a large effect, and an effect that grows with the dose.

No amount makes a conclusion certain. Evidence goes from some cases to a general claim, and new facts can always change what we know.

EXERCISE 2. "Published in leading journals" says nothing about strength. Strength belongs to the body of evidence for one outcome.

"20 studies agree" answers only the inconsistency question. Twenty studies with the same design can share the same weakness. If they are observational, the people chose whether to skip breakfast. Whatever led them to skip it could also affect their weight, so agreement does not remove the third factor.

"Causes" is a claim about cause, and observational studies alone cannot carry it far.

"Obesity" is the outcome, but measured how, in whom, and over what time? That is the indirectness question.

The slide is silent on the rest: risk of bias, imprecision, and whether studies that found nothing went unpublished.

Before you could judge it, the slide would need five things. The design of the studies. The outcome and how it was measured. The people studied, and where. The size of the effect, with its range. The main flaws. Then a word for how strong the evidence is, with the reason.

EXERCISE 3. No study proves a thing like this. A study shows what happened in some people; the claim is about people in general, and that step is never certain. The useful question is how strong the evidence is.

Then ask her what the study was. Did the researchers decide who ate late, by chance, and compare the groups? Or did they watch people who already ate late? If they watched, the people who eat late may differ in other ways, such as their work hours or their sleep. Those could explain the weight.

How many people, and how big was the difference? Is it one study, or do others agree? Who were they, and where? A result from one country's night-shift workers may not hold for students in Raipur.

The sentence to quote: "One study is a clue, not proof. How far to trust it depends on how it was done, and on whether other studies find the same."

Until you have read the study, that is all you can honestly say.

7 · GRADE certainty: four levels, where you start, and moving down or up

EXERCISE 1. High: the true effect lies close to the estimate. Moderate: it is likely to be close, but may be far off. Low: it may be far off. Very low: it is likely to be far off.

Randomised trials start at high; observational studies start at low.

Down: risk of bias, inconsistency, indirectness, imprecision, publication bias. Each moves the rating down one level if serious and two if very serious. Usually the total is three at most, and nothing goes below very low.

Up, mostly for observational evidence, there are three reasons. A large effect gives one level, and a very large effect two. A dose-response gradient gives one. All plausible bias working against the result gives one.

EXERCISE 2. Randomised trials start at high. Two reasons, each costing one level:

```
4 minus 1 minus 1 = 2
```

That is low, as the abstract says.

A sentence for the officer: "The trials suggest the programme lowers weight over 8 weeks, but we are not very sure of it. The true effect may be quite different. The trials had flaws, and they were all in adults in rich countries, followed for only 8 weeks."

The rating would rise if trials with fewer flaws were done. It would also rise if trials followed people for longer, in people like those the officer serves. Each would remove a step down. The rating cannot rise for these trials by rating up: GRADE has not found a convincing case for rating up randomised trials.

EXERCISE 3. Draw a ladder of four rungs: high, moderate, low, very low. Write the circles beside them: four filled for high, down to one for very low. Tell her the circles count rungs, nothing more.

Draw two starting points: randomised trials at the top rung, observational studies at the second from the bottom. Ask her why. Chance made the groups in a trial, so a third factor is less likely to explain the result.

Draw five arrows down: risk of bias, inconsistency, indirectness, imprecision, publication bias. One rung for a serious problem, two for a very serious one. Nothing below the bottom rung.

Draw three short arrows up, beside the observational start only: a large effect, a dose-response gradient, and bias that all pushes against the result.

Then her table: "⊕⊕◯◯ low" from trials means two rungs were lost. Send her to the footnotes to find which two. Finish with what low means: the true effect may be substantially different from the estimate, larger or smaller.

1. A serious reason costs one level; a very serious reason costs two.

```
4 minus 1 minus 2 = 1
```

Level 1 is very low.

2.

```
part (a): 2 plus 1 = 3
part (b): 2 minus 1 = 1
part (c): 2 minus 2 minus 1 = -1
```

Part (a) ends at moderate and part (b) at very low. In part (c) the count falls below 1, but nothing goes below very low, so the rating stops at very low.

3.

```
part (a): 4 minus 2 = 2
part (b): 4 minus 2 = 2
part (c): 4 minus 1 = 3
```

In part (a), two levels down: two serious reasons, or one very serious one.

In part (b), two levels up: a very large effect alone, or two one-level reasons together.

In part (c), three levels down. One reason can cost at most two levels, so one reason is not enough. Two reasons can do it: one very serious (two levels) and one serious (one level).

4. The studies are observational, so the rating starts at low.

Over-reporting would make a link appear where there is none. Every plausible bias here would have produced a link, and still none was found. That is GRADE's third reason to rate up: all plausible bias works against the result.

```
2 plus 1 = 3
```

Moderate. The Cochrane chapter calls this "moderate, or even high certainty" evidence against a causal link at typical doses. It does not show that metformin is safe at every dose, or for every other harm.

5. Compare each number with the thresholds: below 0.5 for large, below 0.2 for very large.

value	below 0.5?	below 0.2?
estimate 0.31	yes	no
lower end 0.26	yes	no
upper end 0.37	yes	no

The whole interval lies below 0.5, so the effect is large even at the end nearest no effect. No part of it reaches below 0.2, so it is not very large. Rate up one level.

```
2 plus 1 = 3
```

Moderate, for the outcome head injury. The Cochrane chapter reaches the same rating. One caution: the GRADE Handbook writes these thresholds for risk ratios and says they cannot always be applied to an odds ratio. It suggests converting an odds ratio to a risk ratio first. Cochrane applies them here directly.

6.

```
part (a): 2 plus 1 plus 1 = 4
part (b): 4 minus 1 minus 2 = 1
part (c): 2 minus 1 minus 1 = 0
```

In part (a), adding back the two one-level steps gives 4: the rating started at high, the start for randomised trials.

In part (b), very low. One rater's "serious" against another's "very serious" is the difference of a level, which is why GRADE asks for the reason in a footnote.

In part (c), the count falls below 1, so the row would stop at very low, not low.

7. The start broke. Observational studies start at low, not high.

```
2 minus 1 = 1
```

The rating is very low.

Why the wrong start looks reasonable: with the newer ROBINS-I tool, non-randomised studies may start at high. But they are then generally rated down two levels for having no randomisation, and the colleague left that out.

```
4 minus 2 minus 1 = 1
```

Either way, the evidence ends at very low.

8. The step up for a large effect is the one that breaks. It needs the absence of plausible confounders. The Cochrane chapter discusses this very case and says that residual confounding, a third factor the studies could not fully remove, is likely here. A second reason: the estimate, 2.19, is above 2, but the interval runs down to 1.64, appreciably

below 2. There, the chapter says, some would hesitate to rate up. And GRADE says to rate up only rarely when a serious limitation is present.

The step for the gradient holds: the risk ratio rises from 1.33 to 2.19 as the dose rises.

```
2 plus 1 = 3
```

A rater who takes the gradient alone ends at moderate. A rater who also takes the large effect must defend it against the interval.

Why the wrong answer looks reasonable: 2.19 really is above 2, and the rule is often quoted without its caution about the interval.

9. Find the average size of a study.

```
1432 divided by 14 is about 102.3
```

About 100 patients a study: many small studies, most with heavy sponsor involvement. The Cochrane chapter gives this as the typical picture that may raise a suspicion of publication bias. Small studies that found nothing may never have been published.

So a rater may lower the rating one level for suspected publication bias. If the studies were trials, the best this evidence could then reach is moderate, before the other four reasons are even checked.

```
4 minus 1 = 3
```

A strong suspicion could cost two levels: low.

What the claim does not establish. “Meta-analysis” names a way of pooling studies, not a level of certainty, and pooling cannot bring back studies that were never published. The suspicion is a suspicion: it does not prove that unpublished studies exist. And nothing here says how large the benefit really is.

10. Rate the experience, not the opinion. Every patient got the plan, and there is no comparison group. That is an uncontrolled case series, an observational design, so it starts at low.

Then risk of bias. With no comparison, weight could have changed for other reasons. Memory may keep the patients who did well and lose those who did not come back. Rate down one level.

```
2 minus 1 = 1
```

Very low. Balshem and colleagues rate a clinician’s report of his own experience the same way.

“Hundreds” does not help. A larger case series is still a case series: more patients shrink the chance error, not the bias.

What the rating does not establish. Very low certainty does not show that the plan fails. It shows that this experience cannot tell whether it works. A trial that compared the plan with usual care could.

8 · Certainty of the evidence is not the strength of a recommendation, and certainty has words

EXERCISE 1. High certainty: no hedge, a plain verb such as “reduces”. Moderate: “probably” or “likely”. Low: “may”, or “The evidence suggests”. Very low: “The evidence is very uncertain about the effect of X on outcome”.

Four things set the strength of a recommendation. The first is the certainty of the evidence. The second is the balance of good effects against bad ones. The third is what the people affected value. The fourth is whether the action is a wise use of money and staff.

EXERCISE 2. (a) Moderate certainty: the word is “probably”. The size is not marked as large or slight, so it is a moderate effect in Cochrane’s terms.

(b) High certainty: there is no hedge. The size is small but important: “slightly”.

(c) Very low certainty. At this level the sentence makes no claim about size at all.

(d) Low certainty: the word is “may”. The size is trivial or none: “little to no difference”. Note that (b) is more certain than (a), although its effect is smaller.

EXERCISE 3. There are three faults.

First, the verb. “Reduces”, with no hedge, is the wording for high certainty. The review rated the evidence low, so the sentence should say “may reduce”.

Second, the jump from review to recommendation. A systematic review rates certainty. It does not recommend; Cochrane asks review authors not to.

Third, the strength. Low certainty alone does not decide that a recommendation is weak, and the slide gives no other reason why it should be strong. A strong recommendation on low certainty needs a clear balance, with much to gain and little to lose. The slide says nothing about harms, about what people value, or about the cost to a district.

Rewritten: “Low-certainty evidence suggests that the programme may reduce readmissions.”

EXERCISE 4. Answer: not necessarily. It can be the right call, and the guideline should say why.

Certainty and strength answer different questions. Certainty says how sure we are of what the action does. Strength says how confident the panel is that the good outweighs the bad, for the people it is meant for. Strength also depends on harms, on what people value and on cost.

A strong recommendation on low certainty is reasonable when the balance is clear even with the uncertainty. GRADE’s own example is giving antibiotics fast in severe infection. No trial has tested it, yet the benefit so clearly outweighs the downside that panels would very likely recommend it strongly.

So ask the guideline two things. Which input made the recommendation strong? And does that input hold in our state: the same harms, the same costs, the same capacity to deliver?

9 · No evidence of an effect is not evidence of no effect

EXERCISE 1. “No evidence of an effect” means the study did not find a clear difference. It may have been too small to see one. “Evidence of no effect” means the data rule out any difference that would matter. They are different statements.

The two marks are zero, no difference, and the smallest difference that matters, fixed before you see the result.

Then the four cases.

- Zero is outside, and big differences are inside: a difference that may matter, or does.

- Zero is outside, and only small differences are inside: a real difference, too small to matter.
- Zero is inside, and big differences are inside too: the study cannot tell.
- Zero is inside, and only small differences are inside: little or no difference.

EXERCISE 2. (a) Zero is outside, and the interval reaches past 2. Evidence of a difference, which may matter: the true value could be under 2 or over it.

(b) Zero is inside, and the interval reaches past 2 on one side and past −2 on the other. Absence of evidence: the data allow a real benefit, a real harm or nothing.

(c) Zero is inside, and the whole interval lies between −2 and 2. Evidence of little or no difference. This is the only one of the five that supports that statement.

(d) Zero is outside, but the whole interval lies below 2. A real difference, too small to matter.

(e) Zero is outside, and the whole interval lies above 2. Evidence of a difference that matters.

EXERCISE 3. The trial found no detectable difference in weight between the advice groups. With P = 0.77 and no confidence interval reported for the difference, it cannot show that there is no effect. Rewritten: "In a 16-week trial of 309 adults trying to lose weight, advice to eat or skip breakfast made no detectable difference to weight."

The second sentence jumps from one trial to a recommendation. One trial's absence of evidence is not enough. The pooled trials found a small difference that favoured skipping. Their authors judged the trials mostly of low quality, so the better summary is that eating breakfast may not help weight loss. And what to tell patients is a recommendation, which also weighs harms, what people value and cost, as the previous section showed.

EXERCISE 4. Quotable sentence: "That trial could not detect an effect of breakfast advice on weight over 16 weeks. That is not the same as showing breakfast has no effect."

Then the rest. The trial was large for its field, and it found no sign that telling adults to eat breakfast helps them lose weight over 16 weeks. It reported no confidence interval for the difference, so it cannot rule out a small effect. When the randomised trials are pooled, the difference slightly favours skipping, but those trials were short and mostly of low quality. So: breakfast advice may not help weight loss, and nobody should promise that it does. Whether to eat breakfast for other reasons is a different question.

10 · Confidence as a probability: putting a number on how sure you are of one claim

EXERCISE 1. A degree of belief is a number from 0 to 1. It says how strongly one person holds one claim to be true, given what they know. Book 0's probability is a long-run frequency: how often an outcome turns up when a chance process is repeated many times. A one-off claim has no long run, so a number on it is a degree of belief. Both obey the same arithmetic.

80 per cent is 0.8, a chance of 4 in 5, and odds of 4 to 1 (0.8 divided by 0.2 = 4).

A stated 50 per cent says the speaker leans neither way: odds of 1 to 1. It does not say they are fairly sure. Nor does it say whether the 50 comes from evidence that points both ways, or from no evidence at all.

EXERCISE 2. What is right: Book 0's frequency reading describes a group or a repeated process. Read that way, a probability says nothing about one person, and a one-off claim such as "this trial's result will hold up" has no long run to count.

The flaw: the colleague assumes the frequency reading is the only one. "70 per cent sure" uses the other reading, a degree of belief. It describes how strongly the speaker believes, given what they know. It is not a rate in a group of people. The *Stanford Encyclopedia of Philosophy* calls this the subjective reading, in which probabilities are "degrees of confidence". The arithmetic is the same. Being 70 per cent sure it holds is being 30 per cent sure it does not, odds of 7 to 3.

The useful reply is to say which reading you meant. Then add that the number is a commitment. It can be checked across many such claims, though not by this one trial alone.

EXERCISE 3. Write the three forms on the board.

```
75 divided by 100 = 0.75
1 minus 0.75 = 0.25
0.75 divided by 0.25 = 3
```

You are 75 per cent sure it is this diagnosis, so 25 per cent sure it is not. That is a chance of 3 in 4, and odds of 3 to 1 that it is this diagnosis. Check by going back: 3 divided by (3 plus 1) is 0.75.

Now turn the odds round: 1 to 3 that it is this diagnosis, or “3 to 1 against”. If you say “3 to 1” to a colleague without saying which side, they may hear 25 per cent where you meant 75. So say the side first: “3 to 1 that it is this”, or say “75 per cent” instead.

Last, say what the 75 per cent is. It is your degree of belief, given what you have found so far. It is not a rate measured in patients like this one, and it should move when the next test comes back.

1.

```
90 divided by 100 = 0.9
25 divided by 100 = 0.25
60 divided by 100 = 0.6
5 divided by 100 = 0.05
```

0.9 is 9 in 10. 0.25 is 1 in 4. 0.6 is 3 in 5. 0.05 is 1 in 20.

2.

```
1 minus 0.85 = 0.15
1 minus 0.3 = 0.7
1 minus 0.5 = 0.5
1 minus 0.99 = 0.01
```

3.

```
0.75 divided by 0.25 = 3
0.2 divided by 0.8 = 0.25
0.9 divided by 0.1 = 9
0.6 divided by 0.4 = 1.5
```

0.75 is 3 to 1.

0.2 is 0.25 to 1, which is 1 to 4.

0.9 is 9 to 1.

0.6 is 1.5 to 1, which is 3 to 2.

4. Add the two sides to get all the chances, then divide the claim’s side by that total.

```
3 divided by (3 plus 2) = 0.6
1 divided by (1 plus 9) = 0.1
7 divided by (7 plus 3) = 0.7
1 divided by (1 plus 1) = 0.5
```

60 per cent, 10 per cent, 70 per cent and 50 per cent.

5. His side comes first: 65 for an attack, 35 against.

```
65 divided by (65 plus 35) = 0.65
1 minus 0.65 = 0.35
65 divided by 5 = 13
35 divided by 5 = 7
```

He was 65 per cent sure there would be an attack and 35 per cent sure there would not. In the smallest whole numbers his odds were 13 to 7 for an attack. Because his two sides add to 100, each side is already a per cent.

6.

```
20 divided by 100 = 0.2
1 minus 0.2 = 0.8
0.2 divided by 0.8 = 0.25
```

The event fails to happen with a chance of 0.8, or 80 in 100. The odds that it happens are 0.25 to 1, which is 1 to 4. The odds against it are 4 to 1.

The 20 per cent is the review's proposed meaning for the word, not a measured rate of any event.

7.

```
20 divided by (20 plus 80) = 0.2
80 divided by (80 plus 20) = 0.8
20 divided by 80 = 0.25
80 divided by 20 = 4
4 divided by 0.25 = 16
```

The low member was 20 per cent sure of an attack, odds of 1 to 4. The high member was 80 per cent sure, odds of 4 to 1. The high member's odds were 16 times the low member's.

In per cent, the gap looks like 60 points. In odds, it is a factor of 16. Both describe the same two men, who had agreed on one phrase.

8. The first step broke. The 30 per cent was for "false", and the writer used it as the degree of belief that the claim is true. Take the complement first.

```
1 minus 0.3 = 0.7
0.7 divided by 0.3 is about 2.33
7 divided by (7 plus 3) = 0.7
```

The writer is 70 per cent sure the claim is true. The odds that it is true are 7 to 3, not 3 to 7.

Why it looks reasonable: the division is correct, and 3 to 7 does go with 30 per cent. It just belongs to the other side of the claim. Naming the side before each number prevents it.

9. The first line broke. "1 in 4" is already a chance, not odds. It counts 1 out of all 4.

```
1 divided by 4 = 0.25
0.25 divided by (1 minus 0.25) is about 0.333
```

The chance is 25 per cent, and the odds are 1 to 3: one for, three against. The worked answer turned a 25 per cent chance into 20 per cent by reading "in" as "to".

Why it looks reasonable: both forms use the same two numbers, and even a published review lists "odds (one in three)" among its formats. Read the joining word before you convert.

10. Compute the odds.

```
0.5 divided by (1 minus 0.5) = 1
```

Fifty-fifty is odds of 1 to 1. The consultant leans neither way.

What it does not establish. It does not say the evidence is evenly balanced. A 50 can come from many studies pointing both ways, or from no studies at all, and the number looks the same. The IPCC's guidance for its authors says its middle phrase "should not be used to express a lack of knowledge" for exactly this reason. Ask the consultant: "Is that because the studies disagree, or because there are none?" It also says nothing about how large a fall in referrals might be, if there is one.

11. Turn the odds into a degree of belief.

```
3 divided by (3 plus 1) = 0.75
1 minus 0.75 = 0.25
```

Odds of 3 to 1 are a degree of belief of 75 per cent that the claim is true. That leaves 25 per cent that it is false.

What it does not establish. First, the "three" in the odds is not a size of effect. Odds of 3 to 1 say how sure the experts are that skipping causes some weight gain. They say nothing about how much; "three times as much weight" misreads a confidence as a quantity. Second, it is a degree of belief, so ask whose: which experts, and on what evidence? Third, 75 per cent also means a 1 in 4 chance, on the experts' own number, that skipping breakfast causes no gain at all.

11 · The same word read as different numbers

EXERCISE 1. A US intelligence estimate called an attack on Yugoslavia "a serious possibility". Kent meant odds of about 65 to 35 for an attack; the reader who asked him had taken the phrase to mean much lower odds. The board members who agreed on the phrase had held a wide range of odds. They ran from about 20 to 80 to about 80 to 20: from 20 to 80 per cent.

One measured finding will do. Wintle and colleagues tested 924 people. Readers shown no numbers gave ranges that overlapped the intended ranges by only 32 per cent on average. Another: readers disagreed more about "unlikely" than about "likely".

What readers agree on: the order of the words. "Very likely" is above "likely", above "possible", above "unlikely".

EXERCISE 2. What they show: the two groups read "possible" differently. The mothers' central figure is 62 per cent and the doctors' is 46, a gap of 16 percentage points (62 minus 46). The lower and upper figures sit 34 points apart for the mothers (79 minus 45) and 38 for the doctors (65 minus 27). For the mothers, "possible" sat above an even chance; for the doctors, just below it.

What they do not show. They do not show what any one mother or doctor meant. They do not show how people read the word now, or in India, or in another language. Nor do they show whether the gap changed anyone's decision. It is one small UK study from 1990, in a review its author calls generally low in quality. It also does not tell you what "possible" should mean. That is a matter of fixing a scale, which is the next section's subject.

EXERCISE 3. The flaw: “serious possibility” carries no number. Kent’s board agreed on that very phrase while holding anything from 20 to 80 per cent. The planning staff who read it took it as lower odds than the writer meant. The officer will fund or not fund on a number she has chosen herself.

The rewrite asks the writer for the number and puts it in. Suppose the writer’s degree of belief is 30 per cent. Then: “We judge the chance at about 30 per cent, roughly 1 in 3, that the programme will reduce overweight among the children screened.” The writer should also say why, which later sections teach. If the writer has no evidence either way, the rewrite says so in words rather than in a phrase that sounds like a judgement.

EXERCISE 4. It is clear about direction. It is not clear about size. Ask people to put a number on “unlikely” and different people give very different answers. In one British study, mothers read it as about 27 in 100, and doctors as about 16 in 100. In a large online study, people disagreed more about “unlikely” than about “likely”. Everyone agrees that “unlikely” is less than “likely”; they do not agree on how much less.

So when a decision depends on it, ask for the number: “How many in 100 people like me get this side effect?”

The sentence to quote: “‘Unlikely’ tells you the side effect is less likely than not; only a number tells you how unlikely.”

12 · Calibrated scales: fixing what the words mean, and putting the number beside the word

EXERCISE 1. “Likely” is 66 to 100 per cent on the IPCC scale and 55 to 80 per cent on ICD 203. The IPCC’s ranges overlap; ICD 203’s run end to end.

With the range in brackets, readers’ ranges overlapped the intended ones by 66 per cent on average, against 32 per cent with no range shown. A link to the table or a box under the mouse did little better than nothing.

Choose one scale. Name it the first time. Print the range, or your own number, in brackets beside the word every time.

EXERCISE 2. Three faults.

First, no scale is named. “Very likely” is 80 to 95 per cent on ICD 203 and 90 to 100 per cent on the IPCC scale. Kent’s chart has no “very likely” at all. The reader has to guess.

Second, “with high confidence” mixes two scales in one sentence. On the IPCC’s terms, high confidence is a judgement of the evidence and agreement, not a probability. ICD 203 forbids putting a confidence level and a likelihood in the same sentence, “to avoid confusion”.

Third, the sentence gives no number the reader can hold the writer to.

A rewrite with one scale and the grounds in their own sentence:

> *We judge it “likely” (55 to 80 per cent, on the scale of US Intelligence Community Directive 203) that the programme would work as well in other districts. We studied one district, so this rests on our judgement of how alike the districts are, not on a test in them.*

The writer may have meant a different number. The point of the rewrite is that the number is now on the page, where an examiner can ask about it.

EXERCISE 3. Answer: adopt one published word scale for all department reports, and print the number in brackets beside every probability word.

Why. A word like “likely” carries no fixed number. Readers turn it into very different numbers, which is how two districts read one report two ways. Published scales fix this on paper, but readers seldom look up a table. In a large on-

line experiment, printing the range beside the word roughly doubled the overlap of readers' ranges with the intended ones. It was 66 per cent against 32. Showing the table one click away barely helped.

Which scale. Two published scales are easy to borrow and they disagree. The US intelligence directive, ICD 203, has seven bands that do not overlap: "likely" is 55 to 80 per cent. The climate panel's scale (IPCC) has overlapping bands: "likely" is 66 to 100. Either can work. ICD 203 is easier, because each number has one word.

What it costs. One page of instructions, and a line in each report naming the scale. No new staff.

What it does not do. It does not make the department's judgements more accurate. It makes them readable, and so open to checking later.

1. The band runs from the centre minus the margin to the centre plus the margin.

```
60 minus 15 = 45 and 60 plus 15 = 75
85 minus 5 = 80 and 85 plus 5 = 90
10 minus 8 = 2 and 10 plus 8 = 18
45 minus 20 = 25 and 45 plus 20 = 65
```

45 to 75; 80 to 90; 2 to 18; 25 to 65.

2. Find the band whose lowest number is at or below the value and whose highest is above it.

number	band	letter
12	5 to 20	B
50	45 to 55	D
72	55 to 80	E
88	80 to 95	F
97	95 to 99	G
30	20 to 45	C

3. The shared part starts at the larger of the two lowest numbers and stops at the smaller of the two highest. If the start is above the stop, the ranges share nothing.

```
55 to 80 and 66 to 100: from 66 to 80, and 80 minus 66 = 14
20 to 45 and 0 to 33: from 20 to 33, and 33 minus 20 = 13
40 to 60 and 45 to 55: from 45 to 55, and 55 minus 45 = 10
63 to 87 and 80 to 95: from 80 to 87, and 87 minus 80 = 7
2 to 12 and 20 to 40: start 20, stop 12, nothing shared
```

4. Kent's bands are 2 to 12, 20 to 40, 40 to 60, 63 to 87 and 87 to 99. ICD 203's are 1 to 5, 5 to 20, 20 to 45, 45 to 55, 55 to 80, 80 to 95 and 95 to 99. On the IPCC scale, give the narrowest range that holds the number.

belief (%)	Kent	IPCC	ICD 203
70	probable (63 to 87)	likely (66 to 100)	likely (55 to 80)
85	probable (63 to 87)	likely (66 to 100)	very likely (80 to 95)
58	chances about even (40 to 60)	about as likely as not (33 to 66)	likely (55 to 80)

85 per cent is not "very likely" on the IPCC scale, which starts at 90. Only 70 per cent gets the same kind of word on all three.

5.

```
shared part: from 66 to 80
80 minus 66 = 14
width of ICD 203's band: 80 minus 55 = 25
width of the IPCC's band: 100 minus 66 = 34
14 divided by 25 = 0.56
14 divided by 34 is about 0.41
```

Only 66 to 80 per cent is "likely" on both. That is 56 per cent of ICD 203's band and 41 per cent of the IPCC's. A reader who knows only ICD 203 will misread most of the IPCC's "likely" (80 to 100 is 20 of its 34 points). One who knows only the IPCC will misread 55 to 66, nearly half of ICD 203's.

This is not Wintle's overlap, which divides by the whole stretch: 14 divided by 45 is about 0.31.

6.

```
44 minus 32 = 12
49 minus 32 = 17
66 minus 32 = 34
12 divided by 34 is about 0.35
17 divided by 34 = 0.5
```

Readers who used the link or the box gained 12 and 17 points. That is a third to a half of the 34 points from brackets. So the brackets' lead is not only that some readers never clicked: even readers who looked did less well.

7. The first line broke. It matched the word, not the number. On the IPCC scale "very likely" is 90 to 100 per cent. Find where that range falls on ICD 203.

```
90 to 95 falls in ICD 203's "very likely", 80 to 95
95 to 99 falls in ICD 203's "almost certain", 95 to 99
99 to 100 has no word on ICD 203
```

So no single ICD 203 word fits. Write the number: "90 to 100 per cent (IPCC: very likely)". The answer as given lowers the claim, letting in 80 to 90 per cent, which the IPCC authors had ruled out.

Why it looks reasonable: the two scales use the same words, so a word looks like a fixed thing that can be carried across.

8. The second line broke. On the IPCC scale the ranges overlap. "Likely" runs from 66 to 100, and "very likely" from 90 to 100. So 95 is both "likely" and "very likely".

It is not "virtually certain", which needs 99 to 100.

```
95 is at or above 90, so it is very likely
95 is below 99, so it is not virtually certain
```

The right word is "very likely (90 to 100 per cent)", the narrowest of the seven main terms that holds 95.

Why it looks reasonable: on Kent's chart and on ICD 203 each word stops where the next begins. Wintle's team notes that readers expect the same of the IPCC scale.

9. Compute the range of numbers a reader could take from "very likely", scale by scale.

```
ICD 203 “very likely”: 80 to 95
IPCC “very likely”: 90 to 100
Hashim’s proposal “very likely”: 80 to 95
lowest reading: 80, highest reading: 100
100 minus 80 = 20
```

Kent’s chart has no “very likely”. It lists “highly likely” under “almost certain”, 87 to 99. So a reader could take anything from 80 to 100 per cent, a spread of 20 percentage points. The report gives no way to say which.

What the sentence does not establish. It gives no grounds: no study, no past result, nothing to judge the “very likely” by. Even with a scale named, the number would be the writer’s judgement, not a measured chance. And “halve” is a separate claim about size. A scheme can very likely cut sales by some amount and still be unlikely to halve them.

A fair rewrite asks the writer for a number and a scale, and for the grounds in a second sentence.

10. Find the “doubles”. Wintle’s average overlap was 66 per cent with the range in brackets and 32 per cent with none. The paper also measured how often readers’ best single number fell inside the intended range. It was 59 per cent with none and 82 per cent with brackets.

```
66 divided by 32 is about 2.06
82 divided by 59 is about 1.39
```

On the overlap measure the gain is about two times. On the best-number measure it is about 1.4 times. “Doubles” holds for one of the two measures the paper used.

What the claim does not establish. The study measured how far readers’ ranges matched one scale’s ranges, ICD 203’s. That is not how well they understood a risk, or whether they made better decisions. The readers were members of the public, mostly in Australia, reading sentences from intelligence reports. The result says nothing yet about patients reading a consent form in India. And “putting a number beside the word” meant the scale’s range, not any number.

13 · Stating how confident you are, and saying why

EXERCISE 1. The claim, worded so it could be shown false. The confidence, as a word from a named scale with its number. The grounds: the kind of evidence and, where rated, its certainty. What would change it: a finding that could turn up, and which way it would move the number.

The IPCC calls the description of the grounds a traceable account.

A claim about an effect is about the average in a group. One patient may or may not have the outcome whatever the average does, because people in a group respond differently. For one patient, start from the frequency in people like them.

EXERCISE 2. The claim is loose. Reduce obesity in whom, by how much, compared with what, by when? As written, almost any result could be read as success.

The confidence is 100 per cent, which says no study could change it. Nothing that rests on evidence earns that, and no scale is named.

There are no grounds. “Everyone agrees” is not evidence about yoga and weight. A team that agrees may only have converged on one view in the room.

Nothing is said about what would change their minds.

A rewrite, with blanks for what is not known:

> *We judge it ____ (about ____ per cent, ICD 203 scale) that adults who join the programme will have a lower average weight after 12 months than similar adults who do not. Our grounds are ____ (the kind of study, and its certainty if rated). A trial in our district that found no difference at 12 months would lower our number.*

If the team has no study at all to put in the blank, the honest number is low, and the honest grounds sentence says "none yet".

EXERCISE 3. Start with what the trials saw: a small drop, from about 11 children in 100 to about 9. Then the range: the true effect could be a drop of about 5 in 100, or a rise of about 2. So the trials cannot tell a small benefit from none, or even from slight harm. Then why: few children had diarrhoea, and the trials had flaws, so the evidence is weak.

Give your number, and say it is yours. For example: "I'd put it at a bit better than even, perhaps 6 in 10, that it helps at all" (about 60 per cent; "likely" on ICD 203 would start at 55). Someone could reasonably say evens. What would change it: a bigger, better trial in older children.

The sentence to quote: "For older children the trials are too small and too flawed to say; probiotics might help a little, or not at all."

Well judged: the evidence is weak, and the sentence says so without becoming useless to her.

14 · The base rate: how often claims like this one hold up

EXERCISE 1. A base rate is how often claims of one kind have turned out true when they were tested. The reference class is the group of past claims, alike in what matters, that you count the present claim among. The base rate of that class is your starting confidence.

In the New England Journal of Medicine from 2001 to 2010, 363 articles tested an established practice. Of these, 146 (40.2 per cent) reversed it. Another 138 upheld it, and 79 gave no clear answer. The group is practices tested again and published in one leading journal, not medicine as a whole.

EXERCISE 2. The flaw: the confidence comes from how established the practice is, and being established is not evidence that it works. Established practices that were tested again were often found no better: 146 of 363 such tests in one leading journal reversed the practice. From that class, a fairer start is about 38 to 49 per cent that a tested practice holds up, not 99. Which end depends on whether you count the tests that settled nothing.

What to ask for: has this practice been tested against a lesser or earlier standard, in a trial? If yes, what did the trial find? If no, say that its standing rests on use, not on a test. Do not call the practice wrong either: the base rate is only where judgement starts.

EXERCISE 3. The quotable sentence: "When doctors test a treatment they already use, it often turns out no better than the simpler option. In one leading journal, it was about four times in ten."

Then the number with its group. In one leading journal, 363 articles over ten years tested a treatment doctors already used. In 146 of them, it did no better than the simpler or older option.

What it does not show: that four in ten of all treatments are wrong. A treatment is tested again mostly when someone doubts it, and such tests are rare. So the honest summary is this. Being in use is no guarantee. And when medicine finds its own errors, the system is working.

EXERCISE 4. A worked example, for "my paper will be accepted at the first journal I send it to".

The reference class: papers of this kind (an observational study from one Indian college) sent to journals of this standing. Where to find a count: the journal's own published acceptance rate, if it gives one; or your own department's record, counted from its last ten submissions.

If no count exists, say so. Then use the nearest count you can find, and name its group. “The journal says it accepts about one in five. I have no figure for papers like mine.” Then write your starting confidence beside the plan, dated, so you can check it later. Do not let the plan’s appeal set the number.

1.

```
14 divided by 48 is about 0.2917
20 divided by 48 is about 0.4167
14 plus 20 plus 14 = 48
```

A is 29.2 per cent, B 41.7 per cent and C 29.2 per cent. The per cents add to 100.1, not 100, because each was rounded. Check the counts, which add exactly.

2.

```
20 divided by 80 = 0.25
1 minus 0.25 = 0.75
0.25 divided by 0.75 = 0.3333
```

The base rate is 25 per cent. A claim of this class fails with a chance of 75 per cent, or 3 in 4. The odds that it holds up are 1 to 3.

3.

```
18 divided by 50 = 0.36
18 plus 22 = 40
18 divided by 40 = 0.45
```

36 per cent of all tests reversed the practice, and 45 per cent of the tests with a clear answer. The second is larger because its denominator is smaller: the 10 unclear tests are left out. Say which you used.

4.

```
146 divided by 363 is about 0.4022
138 divided by 363 is about 0.3802
79 divided by 363 is about 0.2176
146 plus 138 = 284
138 divided by 284 is about 0.4859
```

40.2 per cent reversed, 38.0 per cent upheld, 21.8 per cent unclear. Among the 284 clear answers, the odds of “upheld” are 138 to 146, which is 69 to 73: a little worse than evens. As a starting confidence, that is about 49 per cent that a clear answer upholds the practice.

The paper itself prints the 79 as 27.3 per cent in one place, 21.8 in another and 21.7 in a third. Work it out from the counts.

5.

```
396 divided by 3017 is about 0.1313
3017 minus 1373 = 1644
396 divided by 1644 is about 0.2409
```

13.1 per cent of all the trials, 24.1 per cent of the trials of an established practice. Use 24.1. The 13.1 per cent counts trials of new practices, and by definition a new practice cannot be reversed. The authors print 13 per cent; the 24.1 is your own arithmetic from their counts, so say so when you quote it.

6.

```
12 divided by 20 = 0.6
14 divided by 48 is about 0.2917
16 divided by 35 is about 0.4571
146 divided by 363 is about 0.4022
```

60.0, 29.2, 45.7 and 40.2 per cent. Single years run from about 29 to 60 per cent, because each rests on a few dozen tests. Use the decade's 40.2 per cent, which pools 363. The authors also found no trend across the decade, so pooling the years is fair.

7. The class broke. The 40.2 per cent is a share of tests of an established practice. This claim is a new drug's first positive trial, a different kind of claim. The 40.2 per cent is also a share of tests, not of all practices.

A closer class appears in the same paper, in the authors' own report of an earlier review. Of 45 highly cited studies that claimed a benefit, 7 were later contradicted.

```
7 divided by 45 is about 0.1556
1 minus 0.1556 = 0.8444
```

About 16 per cent were contradicted. So at most 84 per cent held up. That is a ceiling, not an estimate: studies that were weakened but not contradicted are inside the 84.

Why it looks reasonable: both numbers are about medical findings being overturned, and the subtraction is right. Only the group is wrong.

8. The second sentence broke. 53 divided by 396 is the share of the reversals that were supplements: about 13 per cent of the reversals. It is not the share of supplement trials that were reversals. That needs the number of supplement trials as its denominator, and the paper does not print it. With the counts given, no base rate for supplement trials can be worked out.

Why it looks reasonable: the division is right, and 13.4 per cent sits close to the study's headline 13 per cent. The two numbers answer different questions. Before you use any share, say what it is a share of.

9. Compute the reversal share in each group the study gives.

```
146 divided by 363 is about 0.4022
146 divided by 1344 is about 0.1086
```

Even in the narrowest group, tests of an established practice, 40.2 per cent were reversed. That is under half, so "most" is wrong on the study's own numbers. Of all the articles on a practice, 10.9 per cent were reversals.

What the claim does not establish. The study counted practices tested again and published in one leading journal. It did not count "medical advice" in general, and most practice is never tested again. And "not reversed" is not "upheld": 79 of the 363 tests gave no clear answer.

10. Check her arithmetic first.

```
756 divided by 981 is about 0.7706
```

77.1 per cent: her arithmetic is right.

What it does not establish. The class does not match her claim. The 981 were new practices that one leading journal, the New England Journal of Medicine, chose to publish. A thesis programme in a district school is not in that group, and no source this book holds counts her group.

So she has no measured base rate. What she can say: “No count exists for programmes like mine. The nearest one I know, from new practices published in one leading journal, is about 77 per cent. I take my start to be lower, because my class is less filtered.” Saying the class out loud is the part that matters.

15 · Updating: changing your confidence by as much as the evidence warrants

EXERCISE 1. Start from the base rate: of 1,000 cases, that many have the claim true, the rest false. Split each column by how often the evidence turns up in it: the first rate for the true column, the second for the false column. Take the row where the evidence turned up. Divide the count with the claim true by the row total. That is your new confidence.

Evidence is strong when it turns up far more often when the claim is true than when it is false. It is weak when it turns up about as often either way. Strong evidence moves you far; weak evidence moves you a little.

EXERCISE 2. What is right: a trial that finds no benefit is evidence, and confidence should move when it arrives.

The flaw: the resident swung to the latest result. Updating moves you from where you stood by as much as the trial’s strength allows.

How strong is it? A large, well-run trial would seldom miss a real benefit, so its null result moves you a long way. A small trial could miss a real benefit often, so it moves you a little. And “now we know it does not” claims more than any single result can show. The fair sentence is: “This trial lowers my confidence that the practice works, from about X per cent to about Y. A trial like it would usually have found a benefit if there were one.”

EXERCISE 3. Draw a two-way table and fill it with people. Take a disease in 1 of 1,000, and imagine 100,000 people: 100 with the disease, 99,900 without. A test that catches all 100 and flags 5 in 100 of the others flags 4,995 healthy people.

```
99900 times 0.05 = 4995
100 divided by (100 plus 4995) is about 0.0196
```

Of 5,095 positive results, 100 are right: about 2 per cent. Specificity is about the people without the disease. When those people outnumber the sick ones a thousand to one, even a small false-alarm rate makes most positive results false.

Then the habit to keep: before you read a positive result, say the base rate out loud. When a result comes in, change your view out loud too. “I started at 1 in 1,000. After this positive result I am at about 2 in 100, not 95.”

1.

```
1000 times 0.2 = 200
1000 minus 200 = 800
200 times 0.8 = 160
200 minus 160 = 40
800 times 0.1 = 80
800 minus 80 = 720
160 plus 80 = 240
40 plus 720 = 760
```

	claim true	claim false	total
evidence turns up	160	80	240
no evidence	40	720	760

	claim true	claim false	total
total	200	800	1000

2. Divide by the row you are in.

```
160 divided by 240 is about 0.6667
40 divided by 760 is about 0.0526
```

After the evidence turns up, about 67 per cent. After it does not, about 5 per cent. You started at 20 per cent, so the evidence moves you up a long way, and its absence moves you down.

3.

```
45 plus 5 = 50
50 divided by 1000 = 0.05
45 divided by 50 = 0.9
45 plus 95 = 140
45 divided by 140 is about 0.3214
```

The base rate is 5 per cent. The evidence shows in 90 per cent of true claims. But among cases with the evidence, the claim is true in only 32 per cent. The same top count, 45, is divided by two different totals.

4.

```
100000 divided by 1000 = 100
100000 minus 100 = 99900
99900 times 0.05 = 4995
100 plus 4995 = 5095
100 divided by 5095 is about 0.0196
99900 minus 4995 = 94905
0 divided by 94905 = 0
```

After a positive result, about 2 per cent. After a negative result, 0. The test catches every case, so no one with the disease tests negative. The practitioners' 95 per cent is the specificity. It answers another question. Their 5 per cent for a negative result reuses the false-alarm rate. That rate is about people without the disease, not about the negatives.

5.

```
1000 times 0.04 = 40
40 times 0.7 = 28
1000 minus 40 = 960
960 times 0.3 = 288
28 divided by (28 plus 288) is about 0.0886
1000 times 0.1 = 100
100 times 0.7 = 70
1000 minus 100 = 900
900 times 0.3 = 270
70 divided by (70 plus 270) is about 0.2059
```

From 4 per cent, about 8.9 per cent: inside the panel's range. From the practitioners' own start of 10 per cent, about 20.6 per cent. Even their too-high start, updated correctly, gives about 21 per cent, not 70. So they made both errors: they started too high, and they moved too far. The paper names the positive stress test as one of the results where the move was much too large.

6.

```
1000 times 0.3 = 300
300 times 0.8 = 240
1000 minus 300 = 700
700 times 0.2 = 140
240 divided by (240 plus 140) is about 0.6316
240 times 0.4 = 96
140 times 0.7 = 98
96 divided by (96 plus 98) is about 0.4948
```

After the first piece, about 63 per cent. For the second step keep only the 380 cases with the first piece of evidence. The second piece fails to appear in 40 of 100 true cases and 70 of 100 false ones. You end at about 49 per cent: close to evens. The second step started from 63 per cent, not from 30.

7. The last sentence broke. It skipped the base rate. It also took the chance of a positive result given the condition for the chance of the condition given a positive result. Fill in the table.

```
1000 times 0.02 = 20
20 times 0.9 = 18
1000 minus 20 = 980
980 times 0.1 = 98
18 divided by (18 plus 98) is about 0.1552
```

About 16 per cent, not 90. Of 116 positive results, 98 are false alarms.

Why it looks reasonable: 90 per cent is a real property of the test, and it is printed right there. It answers “of people with the condition, how many test positive?”, which is not the question.

8. The third line broke. The false-alarm rate, 10 in 100, applies only to people without the condition: 700 of the 1,000, not all 1,000.

```
1000 minus 300 = 700
700 times 0.1 = 70
270 divided by (270 plus 70) is about 0.7941
```

About 79 per cent, not 73. Why the wrong answer looks reasonable: “10 in 100 positive” sounds like a rate for everyone tested. Say which column each rate belongs to before you multiply.

9. Compute how far the positive result moved her chance, and where it left her.

```
3 divided by 0.3 = 10
9 divided by 0.2 = 45
1 minus 0.09 = 0.91
```

The positive result raised her chance roughly 10- to 45-fold. That is strong evidence. But from a start of 0.2 to 0.3 per cent, it leaves her at 3 to 9 per cent. Even at the top of that range, the chance that she has no cancer is 91 per cent. That is odds of about 10 to 1 against cancer.

What it does not establish. It does not establish that she is well: a 3 to 9 per cent chance calls for the next step of work-up, not for reassurance. The ranges are the panel’s judgement for one described patient in the United States. Strong evidence from a low start still ends low. “Certain” is wrong by a factor of more than ten.

10. Compute how strong the study would have to be to justify that move. Imagine 1,000 claims held at 30 per cent: 300 true, 700 false. To end at 90 per cent, the cases with such a finding must be 9 true for every 1 false.

```
300 divided by 9 is about 33.33
33.33 divided by 700 = 0.0476
```

Suppose every true claim produced such a finding, all 300. Then the finding could turn up in at most about 33 of the 700 false claims: fewer than 5 in 100. Would an association like this turn up in fewer than 5 in 100 cases where skipping does not cause weight gain? No. A third factor or the reverse direction, rivals named in section 3, would produce the association in many of the 700 false claims. Such a finding turns up often when the causal claim is false, so it is weak evidence for it. A small move up is warranted; 30 to 90 is not.

What this does not establish. It does not give a correct final number: nobody has measured how often such associations turn up when the causal claim is false. It shows only that the jump needs evidence far stronger than one observational study.

16 · Confidence is not accuracy: what being calibrated means

EXERCISE 1. Calibrated at 80 per cent: of all the claims you stated at 80 per cent, about 80 in 100 turn out true.

The tally. Write each claim before its outcome is known, with the date and a per cent. Write it on the side you lean to, so a claim at 30 per cent becomes its opposite at 70. When outcomes are known, group the claims by stated level. At each level, divide the number true by the number of claims. Compare that proportion with the level. Below it is over-confidence; above it is under-confidence.

Ten claims are too few because chance moves the tally. At 80 per cent the spread over ten claims is about 13 percentage points, more than the 10-point gap between 80 and 70. And one claim alone moves the proportion by 10 points.

EXERCISE 2. Work out the proportion true and the spread at each level.

```
15 divided by 20 = 0.75
0.6 times 0.4 = 0.24
0.4899 times 0.4899 = 0.24
4.472 times 4.472 = 19.999
0.4899 divided by 4.472 = 0.1095
12 divided by 16 = 0.75
0.75 times 0.25 = 0.1875
0.433 times 0.433 = 0.1875
0.433 divided by 4 = 0.1083
0.95 times 0.05 = 0.0475
0.2179 times 0.2179 = 0.0475
0.2179 divided by 4.472 = 0.0487
```

Every level came out at 75 per cent true.

At 60 per cent, 75 per cent came true: 15 points above her level, about 1.4 spreads of 11 points. That hints at under-confidence but does not show it.

At 75 per cent she matched her level exactly. That fits calibration, but with a spread of about 11 points, it would also fit a true level of 65 or 85.

At 95 per cent, 75 per cent came true: 20 points below, with a spread of about 4.9 points, so about four spreads. That row does show over-confidence. When she writes 95, she is right about 3 times in 4, which is the same as her other levels.

What it does not show. It cannot say whether her claims at each level were of the same kind. Nor can it say whether claims that share a cause made the counts scatter more than the spreads say.

EXERCISE 3. Start with what she has right. Feelings of being sure are real, and they carry some information. The question is how much, and nobody can answer it from the feeling.

Draw two columns on the board: “what I said” and “what happened”. Confidence lives in the first column, accuracy in the second. A feeling is a reading in the first column only. To check it, you need the second column, filled in later, line by line.

Then draw the made-up resident’s tally from this section: four levels, from 60 to 90 per cent, and about 6 in 10 true at every one. Ask her what the 90s were worth. They were worth no more than the 60s, and the resident who wrote them felt sure every time.

Give her the task. A notebook, starting today. One line per claim, before the result: date, claim, per cent, on the side she leans to. Tick or cross when the answer comes. Tally in three months, level by level. Warn her that ten claims at a level will not tell her much. She should look first at the row with the biggest gap and the most claims.

EXERCISE 4. A workable sheet has six columns.

- The date written.
- The claim, in words that can later be checked as true or false.
- The stated per cent.
- The date the outcome is expected.
- The outcome: true, false, or unknown.
- The date it was settled.

Rules. Write every claim on the side you lean to, so every per cent is 50 or more. Never edit a line after the outcome is known; a change of mind is a new dated line. Leave unknown outcomes out of the tally, and say how many you left out.

When to tally. At fixed dates, for example every three months, not when a striking result reminds you.

How many before acting. Use the spread: the square root of the level times (1 minus the level), divided by the square root of the count. Act on a row only when its gap is at least two spreads. As a guide, telling 80 from 70 needs about 64 claims at that level, and 90 from 80 about 36. Use a few levels only, such as 60, 70, 80 and 90, so each one fills faster.

1. Divide the true claims by all the claims, then compare with the level.

```
13 divided by 20 = 0.65
0.65 times 100 = 65
65 minus 70 = -5
7 divided by 10 = 0.7
0.7 times 100 = 70
70 minus 90 = -20
```

At 70 per cent: 65 per cent true, 5 points below. At 90 per cent: 70 per cent true, 20 points below. Both point to over-confidence, because each proportion true is below its level.

2. A claim below 50 per cent is rewritten as its opposite, at the complement.

```
100 minus 30 = 70
100 minus 10 = 90
```

A becomes “not A” at 70 per cent. A was false, so “not A” was true.

B becomes “not B” at 90 per cent. B was true, so “not B” was false.

C stays as it is, at 55 per cent. It turned out true.

D is at 50 per cent, which leans neither way. Either side gives 50, so keep it as written. It turned out false.

3. The spread of one score at 90 per cent is the square root of 0.9 times 0.1.

```
0.9 times 0.1 = 0.09
0.3 times 0.3 = 0.09
5 times 5 = 25
0.3 divided by 5 = 0.06
10 times 10 = 100
0.3 divided by 10 = 0.03
```

The spread of one score is 0.3. Over 25 claims the spread is 6 percentage points; over 100 claims it is 3. Four times as many claims halve the spread.

4. Multiply the number of patients by each chance.

```
50 times 0.8 = 40
50 times 0.25 = 12.5
50 times 0.42 = 21
80 minus 42 = 38
80 minus 25 = 55
```

The resident expects 40 of the 50 patients to have pneumonia. The evidence range gives 12.5 to 21, so about 13 to 21 patients. The gap is 38 to 55 percentage points, every bit of it in the direction of too high.

The 80 is a median across residents answering a written scenario, not one resident's tally of her own patients.

5. The gap you want to see is 10 points, so the spread must be at most 5 points, 0.05. The spread of one score at 90 per cent is 0.3.

```
0.3 divided by 0.05 = 6
6 times 6 = 36
```

About 36 claims, all at 90 per cent. Fewer are needed than at 80 per cent (64), because scores near the top of the scale scatter less.

6. First her overall figures, which are right as arithmetic.

```
8 plus 6 = 14
14 divided by 20 = 0.7
50 plus 90 = 140
140 divided by 2 = 70
```

Now each level, with its spread over 10 claims.

```
8 divided by 10 = 0.8
0.5 times 0.5 = 0.25
0.5 divided by 3.162 = 0.1581
6 divided by 10 = 0.6
0.3 divided by 3.162 = 0.0949
```

At 50 per cent, 80 per cent came true: 30 points above, about 15.8 points of spread, so nearly two spreads. At 90 per cent, 60 per cent came true: 30 points below, with a spread of about 9.5 points, so more than three spreads.

The tally shows a person who is under-confident at her low level and over-confident at her high one. Her levels point the wrong way. The overall 70 against 70 hid both, because the two errors cancel in the average. The 90 row is the one to act on first: its gap is several spreads.

7. The step that broke is the third. A claim at 30 per cent leans against itself. Rewrite each one on the side leaned to, "not X" at 70 per cent, before reading the tally.

```
100 minus 30 = 70
10 minus 4 = 6
6 divided by 10 = 0.6
60 minus 70 = -10
```

The rewritten claims were stated at 70 per cent and 60 per cent came true. That points to over-confidence in the writer's lean, not under-confidence. The fix runs the other way: move towards 50 per cent, say 40 rather than 30, not out to 20. With ten claims, though, a 10-point gap is less than one spread (about 14.5 points at 70 per cent), so the tally cannot yet tell.

Why the wrong answer looks reasonable. "More came true than I said" sounds like under-confidence, and for claims above 50 per cent it would be. Below 50, saying 30 per cent is a confident claim that the thing will not happen. Turning every claim to the side you lean to removes the trap.

8. The arithmetic in the first three lines is right. The step that broke is the fourth: it acts on a gap that ten claims cannot show. Work out the spread at 80 per cent over ten claims.

```
0.8 times 0.2 = 0.16
0.4 times 0.4 = 0.16
0.4 divided by 3.162 = 0.1265
```

About 12.6 points. The gap of 10 points is less than one spread. Seven out of ten is ordinary for a person who is truly calibrated at 80 per cent. The honest reading is "no sign yet either way"; keep writing and tally again near 64 claims.

Why the wrong answer looks reasonable. The tally does say 70, and acting on data feels responsible. But a proportion from ten claims moves 10 points with a single claim. So the size of the tally decides, not the gap.

9. What to compute: what a calibrated 90 per cent would look like, and how many recorded claims it would take to check.

```
100 times 0.1 = 10
0.3 divided by 0.05 = 6
6 times 6 = 36
```

A calibrated 90 per cent is wrong about 10 times in every 100. Thirty years of 90 per cent statements should hold many wrong ones. To tell a true 90 from a true 80, he would need about 36 claims at 90 per cent. Each would have to be written down before its outcome was known.

What this does not establish. It does not show he is over-confident. "Hardly remember being wrong" is memory, not a tally, and people may remember what happened more than what did not. If he truly is almost never wrong, his 90s are too low, which is under-confidence. The sentence cannot tell which, and neither can you, until the claims are written down. Being right often, even if it is true, is not the same as being calibrated.

10. What to compute: the factor in two of the study's scenarios before testing, and whether one rule such as halving fits both. In the survey, the median estimate for pneumonia was 80 per cent against evidence of 25 to 42 per cent. For cardiac ischaemia, reduced blood flow to the heart muscle, it was 10 per cent against 1 to 4.4 per cent.

```
80 divided by 42 = 1.904761905
80 divided by 25 = 3.2
10 divided by 4.4 = 2.273
10 divided by 1 = 10
80 divided by 2 = 40
10 divided by 2 = 5
```

The factor is about 1.9 to 3.2 for pneumonia and 2.3 to 10 for cardiac ischaemia, so "2 to 10 times" fits these. Halving gives 40 for pneumonia, inside the evidence range of 25 to 42, but 5 for cardiac ischaemia, still above the top of 4.4. No single factor corrects both.

What this does not establish. The study used written scenarios with US practitioners, not your doctor and your illness. Its evidence ranges were set by a panel from the literature, and the survey's reliability was not assessed. No held study has measured Indian doctors this way. And a fixed correction cannot repair a number nobody has tallied. What the study supports is narrower and still useful: ask your doctor how common the condition is in people like you, before any test.

17 · Changing a stated position out loud

EXERCISE 1.

1. What I said, to whom, and when.
2. What new evidence has appeared since.
3. What I now think, how confident I am, and on what grounds.
4. What I will do differently.

The rules. Say whether the old position was reasonable on the evidence of its time, and name separately any part that was not reasonable even then. And move only as far as the new evidence warrants, no further.

EXERCISE 2. Sentence 1 is too vague. "What I said about breakfast last year" does not say what was said, where or when, so a reader cannot tell which advice to drop.

Sentence 2 overstates. A meta-analysis of a few short trials, judged mostly low in quality by its own authors, does not prove anything. "Proved" turns low-certainty evidence into a fact.

Sentence 3 swings too far. The review found people assigned to breakfast ended up 0.44 kg heavier, from low-certainty evidence. That supports "breakfast may not help weight loss", not "skipping is better".

"Turns out" also says the evidence changed after he spoke. It did not: the review came out in 2019, before last year's talk. He claimed more than the evidence carried even then, and a later section of this book is about saying so.

It gives no confidence and no grounds, and no action. "My bad" is harmless, but it is not a reason anyone can check.

A rewrite:

> *At the journal club on [date] last year, I said eating breakfast helps people lose weight, and that we should advise it. That went beyond the evidence even then. A 2019 review had pooled seven trials. People assigned to eat breakfast weighed slightly more, by 0.44 kg, than those assigned to skip it. The trials were short, and their authors judged the quality mostly low.*

> *I think it "likely" (about 75 per cent, on the ICD 203 scale) that breakfast does not help weight loss. The evidence is too weak to say that skipping it helps. I will stop advising breakfast for weight loss and will check again when longer trials report.*

EXERCISE 3. The quotable sentences: "I advised breakfast for weight loss because many studies found that people who eat breakfast weigh less. The trials done since then suggest that eating breakfast does not help people lose weight, so I no longer advise it for that."

Then the honest part. "Was I wrong? About the cause, yes: those studies showed a link, and I treated it as a cause. About the link itself, no: it is still there."

Then how sure. "The trials are few and short, and none was done in India. So I would put it this way: breakfast may not help weight loss, and the evidence is too weak to say that skipping it helps. In numbers: about 75 per cent that it does not help weight loss,"likely" on the ICD 203 scale. If longer trials show otherwise, I will say so again, in your paper."

Do not let the sentence become "breakfast is bad for you". The trials measured weight, not health in general.

EXERCISE 4. Check your draft five ways.

- Sentence 1 names the place and the date, and quotes or closely states what you said, including how sure you sounded.
- Sentence 2 names the new evidence specifically enough to find: who, when, what kind of study, what it found, and how strong it is.
- Sentence 3 gives a confidence, as a number and a word from a named scale, with its grounds. If your new confidence is low, it says so.
- Sentence 4 names an action you will take and, if it applies, when you will look again.
- Your sentence on the old view separates what was fair then from what was not.

Then the audience. The people who heard the old version are the ones who need the new one: the same batch, the same group, the same readers. If what you changed is in a published paper, note it for the later section on how journals correct the record.

18 · How science corrects its record: correction, expression of concern, retraction, republication

EXERCISE 1. A correction: an honest mistake of fact that does not change what the paper concludes.

An expression of concern: a question is open. An inquiry may be running, or the editor may not yet know enough to decide.

A retraction: the findings can no longer be trusted. The cause may be an honest mistake or misconduct.

Retraction with republication (COPE calls it retraction with replacement): an honest mistake changes the result a lot, but the study itself is sound. The original is withdrawn, and the fixed paper comes out with a note on what changed.

None of them is used for debate, or for evidence that has moved on. Those go to letters to the editor or to new papers.

EXERCISE 2. (a) A correction. It is an honest error of fact, and it does not change the direction or implications of the conclusions.

(b) None of the four. ICMJE says matters of debate and evolving science are not errors. A letter to the editor, or a new publication, is the place for it. The section on belief beyond the evidence showed that causal words over an association overstate. That is a fault in the writing. It is not an error of fact in the data.

(c) Retraction with republication. An honest error leads to a major change in the results, and the underlying study appears valid. The corrected paper is published with an explanation, beside the retracted original.

(d) An expression of concern, if the editor judges one warranted. The question is open, and ICMJE allows one "pending the outcomes of those procedures".

(e) A retraction. The findings cannot be relied on. ICMJE says that if an investigation proves misconduct, the journal publishes a retraction. The notice should say why. COPE adds that it should stay factual and not attack the people.

EXERCISE 3. The paper was retracted in 2007. Fang and colleagues report that it was retracted "because of concerns about the cytokine's reported insulin mimetic properties". That is exactly the claim the slide cites it for. So the slide rests on the part of the paper that was withdrawn, and it does not say the paper was retracted. COPE: "Authors should not cite retracted papers without acknowledging the retraction."

A slide that can stand says what the record says:

> *Fukuhara et al. (Science 2005) reported visfatin as a product of fat tissue that mimics insulin. The paper was retracted in 2007 over the insulin-like claim.*

If the presenter wants a current claim about visfatin, they need a current source for it. A retracted paper cannot carry it.

EXERCISE 4. Start with the question that decides the tool. Does the wrong number change what the paper concludes? Draw two boxes: "conclusion stands" and "conclusion changes".

If the conclusion stands, the tool is a correction. ICMJE calls honest errors "a part of science and publishing". The journal publishes a short notice and a corrected version. Show her a real one: the 2013 correction to Fang and colleagues' paper on retractions. It is two sentences and a table.

If the conclusion changes but the study is sound, the tool is retraction with republication. The corrected paper comes out with an explanation. Retraction alone is for findings that cannot be relied on at all.

Then the fear. COPE says the purpose of retraction is to correct the literature, "not to punish the authors". A correction is not a retraction in any case.

Then what to do this week. Confirm the error, and tell her co-authors. Write to the editor: what is wrong, how she found it, the correct figure, and whether the conclusion changes. COPE says authors must raise a confirmed error. The editor decides the tool, not her.

19 · Being wrong in public is common, and owning it costs less than hiding it

EXERCISE 1. 146 of 363, about 40 per cent (New England Journal of Medicine, 2001-2010).

About a fifth (21.3 per cent) were for error. About two-thirds (67.4 per cent) were for misconduct: fraud or suspected fraud, duplicate publication and plagiarism.

Their earlier papers showed no statistically significant loss of citations. When authors did not report the error, each earlier paper lost 12.5 per cent of its citations a year, five years on. This is a link seen in the data, not proof that reporting protects you.

EXERCISE 2. The first row says the opposite. Five or more years on, the retracted paper itself lost 86.2 per cent of its yearly citations when the authors reported the error. It lost 81.5 per cent when they did not. Reporting your own error does not rescue the paper.

The second row is where the two differ. After an unreported retraction, each of the authors' earlier papers lost 12.5 per cent of its yearly citations five years on. After a self-reported one there was no statistically significant loss.

What it does not show: that reporting causes the protection. Authors chose whether to report, and Lu and colleagues say self-reported errors "may reflect more innocuous or explainable errors". And it measures citations only, not jobs, grants or reputation.

EXERCISE 3. First sentence. Fang and colleagues found 67.4 per cent of 2,047 retractions attributed to misconduct, which includes duplicate publication and plagiarism as well as fraud. Fraud or suspected fraud alone was 43.4 per cent. And these are shares of retractions, which are rare: about 0.14 per 1,000 papers in biology and medicine.

Second sentence. Prasad and colleagues looked at one journal over ten years. Of 363 articles that tested an established practice, 146 found it no better than a lesser or earlier standard. That is a share of practices someone chose to test, and tested practices may be the doubtful ones. It is not 40 per cent of "what doctors do", and "no better than a lesser standard" is not "useless" in every case.

Third sentence. The conclusion does not follow. The figures show that medicine finds and corrects its errors: the reversals were found mostly by trials, and the retractions by editors. A fair version: "Medical practice is often overturned by better trials, and a small number of papers are withdrawn, most of them for misconduct. That is how errors get found."

EXERCISE 4. Answer first: No. Retraction is rare, and medicine's usual way of being wrong is the one that research itself corrects.

The numbers. About 0.14 papers in every 1,000 in biology and medicine journals are retracted (Lu and colleagues, 2013). Of retracted biomedical papers up to 2012, two-thirds were withdrawn for misconduct and a fifth for honest error (Fang and colleagues, 2012).

The commoner kind of wrong is different. Prasad and colleagues (2013) read ten years of one leading journal. About four in ten of its articles that tested an established practice found it no better than a lesser or earlier standard. Most of those reversals came from randomised trials.

What to do. Rely on reviews of many studies rather than single papers. Ask that each policy note says when its evidence was last checked. Ask that it be checked again when a new review appears.

The limit. These counts come from journals and databases outside India. They show that errors are found and corrected. They do not measure how often Indian research is wrong.

20 · Saying "I was wrong about that"

EXERCISE 1. Five steps: find, sort, choose the instrument, say it, date it.

Three things that can have happened. The evidence moved after you spoke. You claimed more than the evidence of the time could carry. Or you made a mistake in your own work.

Four sentences:

1. What I said, to whom, and when.
2. What new evidence has appeared since (or, for an overstatement or an error, what I now see).
3. What I now think, how confident I am, and on what grounds.
4. What I will do differently.

EXERCISE 2. Check your work against six tests.

- The claim is your own and was public: a paper, thesis chapter, lecture, ward teaching, poster or post. A claim you only thought does not count.
- You checked it against a source newer than the claim, preferably a systematic review, and you can name that source.
- Your sort is honest. If studies that already undercut the claim were published before you made it, you wrote "I overstated", not "the evidence changed".
- The correction reaches the people who met the claim. A spoken claim goes back to the same class or ward. An error in your published work goes to the editor. A published claim that later evidence has overtaken gets a letter or a new paper.
- Sentence 3 names a word, its scale and a number ("unlikely", ICD 203, about 25 per cent), and its grounds.
- The dated line has the date you said it, not the date you wrote the draft.

The statement contains no apology for honest work and no claim that you always knew.

EXERCISE 3. (a) The statement abases the speaker and gives the listener nothing to use. It does not say what exactly was claimed or when (sentence 1). It names no evidence (sentence 2). "Rubbish" is not a new position with a confidence (sentence 3). "Ignore my lectures" throws out every claim, right or wrong, instead of changing this one (sentence 4). An honest error does not call for shame.

(b) The statement hides who was wrong. It has no "I", no claim and no date (sentence 1). "Evolving research" names no study (sentence 2) and quietly says the evidence changed, which for a 2016 slide was not the case. There is no new position and no confidence (sentence 3), and nothing will be done differently (sentence 4).

A version that works is the four sentences in this section's first illustration.

EXERCISE 4. Start with what the faculty already saw. The slide is out there either way. The question is whether the correction comes from him or from someone else.

Then the evidence about cost. Being wrong in public is common: in one leading journal over ten years, 146 of 363 tests of an established practice reversed it.

In Lu and colleagues' study, when authors reported their own retraction, the loss of citations to their earlier work disappeared. When others reported it, their earlier work lost citations. That is an association, and it is about papers, not seminars. Rohrer and colleagues argue that admitting an error is not a reliable sign of being error-prone.

Then the procedure, on the whiteboard. Find: he has. Sort: he did not check the reference's status (or it was retracted after his talk; the dates will tell).

Instrument: the same audience, so a short note to the department, or two minutes at the next seminar. Then the four sentences. What he said, and when. That the study was retracted, and why. What he now thinks, with a confidence. That he now checks each reference's status on the day he presents. Then date it.

End with the habit from the section on correcting the record: check a paper's page on the day you cite it.

21 · The dated belief register: what you believe, how sure you are, and when you said it

EXERCISE 1.

1. The number and the date.

2. The claim, worded so that someone else could judge it: what, in whom, measured how, by when.
3. The confidence: a word from one named scale, with its number.
4. The grounds: the kind of evidence, and, where it has been rated, its certainty; where it has not, how directly it bears on the claim.
5. What would change your mind, named in advance.
6. A review date.

The rule: a line is never edited. A changed view, or an outcome, is a new dated line that names the line it refers to. It exists because the tally of stated confidence against outcome needs the level you held before you knew the outcome. An edited line no longer shows that.

EXERCISE 2. Five faults.

- "Works" for what, in whom, over what time? Nobody could judge the claim true or false.
- "Probably" and "very confident" are two words from no scale, and they may not agree. There is no number.
- No grounds: no kind of evidence and no certainty.
- Nothing says what would change her mind.
- "Sometime next year" is not a date.

A rewrite, with gaps marked:

no.	date	claim	confidence (ICD 203)	grounds and their certainty	what would change my mind	review
1	2026-10-15	In adults with obesity, [a named fasting schedule] produces more weight loss at 12 months than [the comparison]	[word] ([number]%)	[kind of evidence, and its certainty, from a review she has read]	[a named result, such as a large trial at 12 months showing no difference]	[a date]

The gaps are hers to fill from the evidence. Do not fill the grounds from memory.

EXERCISE 3. Draw two columns: "what I thought then" and "what happened".

Start with her own example. Ask her how sure she was, on her first day, of something she has since learned. Whatever number she gives now is a number she gives knowing the answer. There is no way to show it was the number she held then. That is the gap the register closes.

Then show one line, with its date and number. Show a second line that revises it, and point out that the first line stays. Explain the goal. After a few years she can count how often her "70 per cent" lines came true. That count is the only way to learn whether her 70 means 70. Memory cannot give her that count, because it holds what she thinks now, not what she wrote then.

EXERCISE 4. Check your page against this list.

- Every line has a number and a date, and the page names one scale.
- Every claim could be judged true or false by a stranger. Read each aloud and ask: what result would make this false? If you cannot say, rewrite it.

- Every confidence has a word and a number of 50 per cent or more, and the number falls in that word's band on your scale.
- Every line's grounds name a kind of evidence and its certainty. None says only "I have seen it" or "everyone knows".
- At least one line has grounds you judge weak, and its number says so. A register of five certainties is a register that has not looked at its evidence.
- At least two lines close on a date.
- Nothing has been crossed out. Any change is a new dated line.

Your register meets the standard when all seven hold.

EXERCISE 5. A workable design:

- **Where.** One file or bound notebook per resident, with numbered lines. If it is a shared spreadsheet, keep its version history switched on, so any edit to an old line can be seen.
- **Who changes what.** Only the owner adds lines. Nobody, the owner included, edits an old line. A revision or an outcome is a new line that names the old one.
- **One scale.** The department picks one scale, names it at the top of every register, and keeps it for the two years. Changing scales midway makes the old lines hard to count.
- **Reviews.** Twice a year, each resident closes the lines whose date has passed and adds the outcomes. At the review meeting, the group reads closed lines without blame.
- **Before discussions.** At journal club, each resident writes a line before the discussion, as the Intergovernmental Panel on Climate Change (IPCC) suggests its authors do.

Say what the design cannot do. Six residents over two years will close too few lines to tell one person's 60 from their 70. That count needs years of lines and a scoring method beyond this book. The design makes the count possible later.

22 · One claim, from a belief to a dated line: "eating breakfast helps you lose weight"

EXERCISE 1.

1. Sort it: hypothesis, finding or belief? Is the verb causal?
2. What does it rest on, and what other explanations fit that evidence?
3. The two legs: is there a measured difference between groups made by chance, and a shown mechanism? Which is weak?
4. Base rate: how often do claims of this kind hold up, and how well does the class fit?
5. Each trial: what did it find, and what could it not show?
6. The body of trials: what is the estimate, and what is its certainty, in words?
7. Update: how far should this evidence move me, counted once, in a two-way table?
8. The confidence statement: claim, word and number on a named scale, grounds and certainty, what would change my mind.

9. The register: a dated line, never edited.
10. Out loud: did I teach or publish the old view, and to whom do I say so?
11. What does it not show: people, place, time, outcome, cause?

EXERCISE 2. Four faults.

- "Proves": evidence supports a claim more or less strongly; it never proves one. Here the certainty is low.
- "Skipping breakfast is the secret to weight loss": the trials compared eating breakfast with skipping it. Those who ate it ended a mean of 0.44 kg heavier, with an interval of 0.07 to 0.82 kg. That does not show skipping causes weight loss of any useful size.
- The few weeks behind it: the trials averaged seven weeks and none ran past 16. A column read by people planning for a year is reading a few weeks of evidence.
- It says nothing of who was studied: adults in the UK and the US (one trial also in Denmark), none in India.

A sentence it could have printed:

> *Short trials in adults in rich countries did not find that eating breakfast helped people lose weight. The evidence is weak, and none of the trials was done in India.*

EXERCISE 3. Correct the question gently. The trials compared eating breakfast with skipping it. Skippers ended a little lighter over a few weeks, at low certainty, which is not evidence that skipping is a way to lose weight. Then give her the sentence.

> *The trials suggest that eating breakfast does not help adults lose weight. But skipping it is not a proven way to lose weight either.*

Then how sure, in plain words: "I'd put it at about one in four that breakfast helps. The trials are short, a few weeks, and none was done in India, so I hold that loosely." If she asks what would change your mind, say: a well-run trial lasting a year or more.

EXERCISE 4. Judge your sheet by four tests.

- Did you set the starting confidence before reading the trials, and say how well the reference class fits?
- Is every judgement marked as one, with its number, so another reader could change it and redo the table?
- Did you count each piece of evidence once? Check whether any trial you read is inside a review you also used.
- Does your "does not show" list name the people, the place, the time, the outcome and the cause?

If your claim has no trials at all, steps 5 to 7 have nothing to work on. Then the line should say so. Give the starting confidence from the base rate and its certainty. Put the trial that would settle it under "what would change my mind".

Glossary

Every term of art this book teaches, in the plain words it gets where it is first used, and the section that teaches it.

Term	Plain words	Section
absence of evidence (of a difference)	a comparison that finds no clear difference between two groups; it has not shown evidence that there is no difference	9
accepted (stance towards a claim)	taken as the best available description of what it describes (Barseghyan et al.)	2
association	a finding that two things go together: one is found more often, or in larger amounts, where the other is present, as excess weight is found more often among people who skip breakfast	1, 3
base rate	how often claims of one kind have turned out true when they were tested; the evidence particular to a new claim of that kind then moves you from there	14
belief	a statement a person holds to be true, held with some strength, on whatever grounds that person has	1
blinding	the people in a study, or those measuring them, do not know who got which treatment	6
body of evidence	all the studies that address one question for one outcome	6
calibrated (of a person's confidence)	see calibration (2): of all the claims a person has stated at one level, the proportion that turn out true is about that level; of their 80 per cent claims, about 80 in 100	16
calibrated scale (of probability words)	a published table that fixes, for each probability word, the range of degrees of belief it stands for; two bodies can fix the same word differently	12
calibration	(1) measuring a machine against something already known, so its readings mean something (Book 0); (2) of a person: of all the claims they stated at one level of confidence, about that proportion turn out true, checked with a tally over many claims (Book 10, section 16)	16
case series	a study that describes one group of people who share something, such as weight lost and kept off, with no comparison group	4
causal claim	a claim that changing one thing, the exposure, changes another, the outcome	5
certainty of the evidence (GRADE)	GRADE's rating of how sure one can be that an estimate of an effect is close to the true effect, for one outcome across all the studies that measured it: high, moderate, low or very low; older texts call it the quality of evidence	7

Term	Plain words	Section
claim about an effect; claim about a single event	what something does on average in a group of people, against what will happen once	13
confidence	(1) this book's word for a person's degree of belief that a claim is true, stated as a word from a named scale with its number; never GRADE's certainty, and not a confidence interval (Book 10, section 10); (2) in the IPCC's and ICD 203's usage, a judgement of the evidence and agreement behind a finding, in words from very low to very high; not a probability (Book 10, section 12)	10, 12
confidence statement	how sure you are of one claim, and why, in a form someone else can check, in four parts: the claim, the confidence (a word from a named scale with its number), the grounds with their certainty, and what would change it	13
confounder	a third factor: something linked to both things, which produces the association although neither causes the other	3
correction	(1) subtracting a known systematic error, such as a zero error, from the reading (VIM 2.53) (Book 0); (2) a notice a journal publishes for an honest error that should have been recognised at the time of publication and does not change the direction or implications of the results; citations go to the most recent version (Book 10, section 18)	18
dated belief register	a written list of claims one person holds, one numbered, dated line each, kept so each claim can be checked later against what happened; each line gives the claim, the confidence (word and number on one named scale), the grounds and their certainty, what would change the person's mind and a review date; only added to, never edited	21
degree of belief	a number from 0 to 1 that states how strongly one person holds one claim to be true, given what that person knows; it obeys the same arithmetic as a probability	10
difference between groups (evidence of)	compare groups that differ in the exposure and ask whether the outcome differs between them; other things equal, strongest when chance decided which group each person joined (Howick's "direct evidence")	5
dose-response gradient	more of the exposure goes with more of the effect	6
erratum	an older word for a correction (2)	18
expression of concern	a notice an editor may publish while a question about a paper is open: while an investigation runs, or while the editor lacks the information to decide	18
falsifiable	some possible result would count against it	2

Term	Plain words	Section
finding	what a particular study observed, in its own people and setting, with its uncertainty	1
GRADE	Grading of Recommendations Assessment, Development and Evaluation: the GRADE Working Group's scheme for rating a body of evidence	6, 7
guideline (clinical practice)	a document in which a group of experts says what clinicians, patients or health services should do; not the same as guideline (NMC)	8
hypothesis	a proposed explanation, stated so that a study can be designed to support or refute it	1
ICD 203	US Intelligence Community Directive 203 (2015, amended 2022), whose seven likelihood bands run end to end, from 1 to 99 per cent	12
imprecision	few people or few events in the studies, so the range of values the data allow is wide	6
inconsistency	the studies give widely different results that nobody can explain	6
indirectness	the studies answer a nearby question: other people, another version of the treatment, another comparison, or a stand-in outcome	6
inductive reasoning	reasoning from particular observations to a general conclusion; it supports a conclusion more or less strongly, never with certainty	6
judgement on evidence (Kent's second kind)	a statement the evidence makes likely but cannot make certain	1
likelihood scale (IPCC)	the IPCC's seven terms for a probability, from exceptionally unlikely (0 to 1 per cent) to virtually certain (99 to 100), whose ranges overlap by design	12
loss of confidence (in one's own finding)	a researcher no longer believing a result they published (Rohrer et al.)	19
mechanistic evidence	evidence for the chain of steps by which the exposure is said to produce the outcome	5
medical reversal	an established practice found by a later study to be no better than a lesser or earlier standard of care	14
natural frequencies	whole counts of imagined cases, out of 1,000, or out of 100,000 when the base rate is rare, so every count stays whole	15
number needed to treat (NNT)	how many people must be treated for one more of them to do well	6
observational study	a study that watches people as they are, without changing anything: its groups are observed as they came, not made by chance	1, 6

Term	Plain words	Section
over-confidence (at a level)	at a level above 50 per cent, the proportion of claims that turn out true is below the level the person stated: they were right less often than they said	16
parallel evidence	other studies, replications and similar ones, that give similar results (Howick et al.)	5
prediction	a statement of the form "if the hypothesis is true, then this observable result will follow", fixed before the result is seen	2
probability	(1) a number from 0 to 1 that states how often an outcome would turn up if the same chance process were repeated a great many times (Book 0); (2) a degree of belief in one claim, which obeys the same arithmetic (Book 10, section 10)	10
publication bias	studies missing because their results were never published	6
pursued (stance towards a claim)	thought worth developing and testing further	2
rate down, rate up (GRADE)	lower or raise a certainty rating by one or two levels, for a reason stated in a footnote	7
recommendation (in a guideline)	a statement of what should be done; it answers "what should be done?", where the certainty of the evidence answers "how sure can we be that the estimate of the effect is close to the true effect?"	8
reference class	the group of past claims, alike in what matters, among which you count the present claim	14
retraction	a notice published when a paper's findings and conclusions can no longer be relied on; the cause may be honest error, a naive mistake or misconduct	18
retraction with republication	withdrawing a paper whose honest error changed its result substantially, and publishing the corrected version beside it (COPE: retraction with replacement)	18
reverse direction	the second thing causes the first, not the first the second	3
review date (of a register line)	the date written on a line of the register for looking at it again, or for judging it	21
risk of bias	flaws in how the studies were run that push the estimate one way	6
self-reported retraction	a retraction in which the authors themselves reported the error (Lu et al.)	19
silent update	changing what you teach, such as your slides, without saying so, which leaves the old position with everyone who heard it; not the same as silent change (Book 6)	17
smallest difference that matters	the smallest difference that would change what someone does, decided before the result is seen	9

Term	Plain words	Section
strength of a recommendation	GRADE's grade for a recommendation, strong or weak, set by four things: the certainty of the evidence, the balance of desirable and undesirable effects, the values and preferences of the people affected, and whether the action is a wise use of resources	8
Summary of findings table	the table in a Cochrane review that gives, for each important outcome, the effect and its GRADE certainty, with the reasons in footnotes	7
surrogate outcome	a measurement that stands in for the outcome that matters to patients	6
tally (of claims by confidence level)	claims grouped by the level stated for them; at each level, the number of claims, the number true, and the proportion true	16
testable (hypothesis)	a prediction follows from it: "if it is true, then this result"	2
traceable account	the IPCC's name for a description of the evidence and agreement behind each finding	13
under-confidence (at a level)	the proportion of claims that turn out true is above the level the person stated; at 50 per cent, any proportion true far from 50, high or low	16
updating	changing your confidence in a claim when new evidence arrives, by as much as that evidence warrants and no more	15
used (stance towards a claim)	taken as an adequate tool for a practical job	2
values and preferences	what the people affected value, and how they weigh the good and bad effects of an action; one of the four things that set the strength of a recommendation	8
verbal probability	a word or phrase that states a degree of belief without a number: "likely", "unlikely", "possible", "a serious possibility"	11
weak recommendation	most people would choose the action, but many would not, so decide with the person; some panels write "conditional" or "discretionary" for weak	8

The series: all 196 books

In build order: Book 0 first, then every Introductory rung, then every Intermediate rung, and so on; within a level, a subject comes after the subjects it depends on. "done" means that book's version 1 is finished. This is Book 10.

GROUND FLOOR		
0	Ground floor	done
INTRODUCTORY		
1	Energy balance and dynamic weight-change modelling · Rung 1	done
2	Mathematics for modelling and inference · Rung 1	done
3	Qualitative and mixed methods · Rung 1	done
4	Food systems, value chains and Indian food policy · Rung 1	done
5	Policy process, framing and advocacy craft · Rung 1	done
6	R, reproducibility engineering and data management · Rung 1	done
7	Question selection and research strategy · Rung 1	done
8	Teaching, mentorship and coalition building · Rung 1	done
9	Scientific writing, visualisation and public communication · Rung 1	done
10	**Epistemics: philosophy of science and calibration · Rung 1 ◀ you are here**	done
11	Adjacent specialist domains: recognise and route · Rung 1	
12	Probability and statistical inference · Rung 1	
13	Integrative metabolism, insulin resistance and energy expenditure regulation · Rung 1	
14	Commercial determinants and corporate political activity · Rung 1	
15	System dynamics and complexity concepts · Rung 1	
16	Indian regulatory architecture and the live policy agenda · Rung 1	
17	Comparative and international policy · Rung 1	
18	Python for simulation, scraping and tooling · Rung 1	
19	Regression modelling for health outcomes · Rung 1	
20	Potential outcomes, identification and target trial emulation · Rung 1	
21	Adipose tissue biology, ectopic fat and the South Asian phenotype · Rung 1	
22	Appetite, reward and the determinants of intake · Rung 1	
23	Nutrient science for obesity and cardiometabolic risk · Rung 1	
24	Physical activity and sedentary behaviour science · Rung 1	
25	Sleep science and obstructive sleep apnoea · Rung 1	
26	Participatory systems mapping and group model building · Rung 1	
27	Agent-based and network modelling · Rung 1	
28	Clustered, longitudinal, hierarchical and small-area models · Rung 1	
29	Evidence synthesis and meta-analysis · Rung 1	
30	Prediction modelling and machine learning · Rung 1	
31	Measurement error theory and correction · Rung 1	
32	Causal diagrams, structural bias and sensitivity analysis · Rung 1	
33	Study and trial design, power and missing data · Rung 1	
34	Clinical assessment, staging and secondary causes · Rung 1	
35	Indian dietary patterns, guidelines, food composition and diet cost · Rung 1	
36	Circadian biology and chrononutrition · Rung 1	

37	Behavioural theory, choice architecture and intervention design · Rung 1
38	Measurement instruments: body composition, expenditure, diet, activity, sleep · Rung 1
39	G-methods, time-varying treatment and mediation · Rung 1
40	Instrumental variables and Mendelian randomisation · Rung 1
41	Quasi-experimental policy evaluation · Rung 1
42	Genetics, developmental origins and life-course epidemiology · Rung 1
43	Social epidemiology, equity measurement and surveillance in India · Rung 1
44	Nutritional epidemiology and its pathologies · Rung 1
45	Lifestyle, dietary and behavioural treatment · Rung 1
46	Pharmacotherapy of obesity · Rung 1
47	Metabolic and bariatric surgery: selection and co-management · Rung 1
48	Eating behaviour, mental health and weight stigma · Rung 1
49	Food environment measurement, retail and digital delivery · Rung 1
50	Food demand, elasticity and the economics of externality · Rung 1
51	Implementation science and programme craft · Rung 1
52	Protocol, preregistration, reporting standards and research ethics · Rung 1
53	Life-stage obesity: paediatric, adolescent and pregnancy · Rung 1
54	Ultra-processed food and the effects of processing · Rung 1
55	Activity measurement and compositional 24-hour analysis · Rung 1
56	Fiscal instruments: taxation and subsidy design in India · Rung 1
57	Burden of disease, economic evaluation and decision-analytic modelling · Rung 1
58	The Indian health system and its financing mechanics · Rung 1
59	India's data infrastructure, access and complex survey analysis · Rung 1
60	Ethics of obesity intervention and industry engagement · Rung 1
61	Health financing, HTA and pharmaceutical access in India · Rung 1
INTERMEDIATE	
62	Energy balance and dynamic weight-change modelling · Rung 2
63	Mathematics for modelling and inference · Rung 2
64	Qualitative and mixed methods · Rung 2
65	Food systems, value chains and Indian food policy · Rung 2
66	Policy process, framing and advocacy craft · Rung 2
67	R, reproducibility engineering and data management · Rung 2
68	Question selection and research strategy · Rung 2
69	Teaching, mentorship and coalition building · Rung 2
70	Scientific writing, visualisation and public communication · Rung 2
71	Epistemics: philosophy of science and calibration · Rung 2
72	Probability and statistical inference · Rung 2
73	Integrative metabolism, insulin resistance and energy expenditure regulation · Rung 2
74	Commercial determinants and corporate political activity · Rung 2
75	System dynamics and complexity concepts · Rung 2
76	Indian regulatory architecture and the live policy agenda · Rung 2
77	Comparative and international policy · Rung 2
78	Python for simulation, scraping and tooling · Rung 2
79	Regression modelling for health outcomes · Rung 2
80	Potential outcomes, identification and target trial emulation · Rung 2
81	Adipose tissue biology, ectopic fat and the South Asian phenotype · Rung 2

82	Appetite, reward and the determinants of intake · Rung 2
83	Nutrient science for obesity and cardiometabolic risk · Rung 2
84	Physical activity and sedentary behaviour science · Rung 2
85	Sleep science and obstructive sleep apnoea · Rung 2
86	Participatory systems mapping and group model building · Rung 2
87	Agent-based and network modelling · Rung 2
88	Clustered, longitudinal, hierarchical and small-area models · Rung 2
89	Evidence synthesis and meta-analysis · Rung 2
90	Prediction modelling and machine learning · Rung 2
91	Measurement error theory and correction · Rung 2
92	Causal diagrams, structural bias and sensitivity analysis · Rung 2
93	Study and trial design, power and missing data · Rung 2
94	Clinical assessment, staging and secondary causes · Rung 2
95	Indian dietary patterns, guidelines, food composition and diet cost · Rung 2
96	Circadian biology and chrononutrition · Rung 2
97	Behavioural theory, choice architecture and intervention design · Rung 2
98	Measurement instruments: body composition, expenditure, diet, activity, sleep · Rung 2
99	G-methods, time-varying treatment and mediation · Rung 2
100	Instrumental variables and Mendelian randomisation · Rung 2
101	Quasi-experimental policy evaluation · Rung 2
102	Genetics, developmental origins and life-course epidemiology · Rung 2
103	Social epidemiology, equity measurement and surveillance in India · Rung 2
104	Nutritional epidemiology and its pathologies · Rung 2
105	Lifestyle, dietary and behavioural treatment · Rung 2
106	Pharmacotherapy of obesity · Rung 2
107	Metabolic and bariatric surgery: selection and co-management · Rung 2
108	Eating behaviour, mental health and weight stigma · Rung 2
109	Food environment measurement, retail and digital delivery · Rung 2
110	Food demand, elasticity and the economics of externality · Rung 2
111	Implementation science and programme craft · Rung 2
112	Protocol, preregistration, reporting standards and research ethics · Rung 2
113	Life-stage obesity: paediatric, adolescent and pregnancy · Rung 2
114	Ultra-processed food and the effects of processing · Rung 2
115	Activity measurement and compositional 24-hour analysis · Rung 2
116	Fiscal instruments: taxation and subsidy design in India · Rung 2
117	Burden of disease, economic evaluation and decision-analytic modelling · Rung 2
118	The Indian health system and its financing mechanics · Rung 2
119	India's data infrastructure, access and complex survey analysis · Rung 2
120	Ethics of obesity intervention and industry engagement · Rung 2
121	Health financing, HTA and pharmaceutical access in India · Rung 2
ADVANCED	
122	Energy balance and dynamic weight-change modelling · Rung 3
123	Mathematics for modelling and inference · Rung 3
124	Qualitative and mixed methods · Rung 3
125	Food systems, value chains and Indian food policy · Rung 3
126	Policy process, framing and advocacy craft · Rung 3

127	R, reproducibility engineering and data management · Rung 3
128	Question selection and research strategy · Rung 3
129	Teaching, mentorship and coalition building · Rung 3
130	Scientific writing, visualisation and public communication · Rung 3
131	Epistemics: philosophy of science and calibration · Rung 3
132	Probability and statistical inference · Rung 3
133	Integrative metabolism, insulin resistance and energy expenditure regulation · Rung 3
134	Commercial determinants and corporate political activity · Rung 3
135	System dynamics and complexity concepts · Rung 3
136	Indian regulatory architecture and the live policy agenda · Rung 3
137	Regression modelling for health outcomes · Rung 3
138	Potential outcomes, identification and target trial emulation · Rung 3
139	Adipose tissue biology, ectopic fat and the South Asian phenotype · Rung 3
140	Appetite, reward and the determinants of intake · Rung 3
141	Nutrient science for obesity and cardiometabolic risk · Rung 3
142	Physical activity and sedentary behaviour science · Rung 3
143	Sleep science and obstructive sleep apnoea · Rung 3
144	Participatory systems mapping and group model building · Rung 3
145	Clustered, longitudinal, hierarchical and small-area models · Rung 3
146	Evidence synthesis and meta-analysis · Rung 3
147	Measurement error theory and correction · Rung 3
148	Causal diagrams, structural bias and sensitivity analysis · Rung 3
149	Study and trial design, power and missing data · Rung 3
150	Clinical assessment, staging and secondary causes · Rung 3
151	Indian dietary patterns, guidelines, food composition and diet cost · Rung 3
152	Circadian biology and chrononutrition · Rung 3
153	Behavioural theory, choice architecture and intervention design · Rung 3
154	Measurement instruments: body composition, expenditure, diet, activity, sleep · Rung 3
155	G-methods, time-varying treatment and mediation · Rung 3
156	Instrumental variables and Mendelian randomisation · Rung 3
157	Quasi-experimental policy evaluation · Rung 3
158	Genetics, developmental origins and life-course epidemiology · Rung 3
159	Social epidemiology, equity measurement and surveillance in India · Rung 3
160	Nutritional epidemiology and its pathologies · Rung 3
161	Lifestyle, dietary and behavioural treatment · Rung 3
162	Pharmacotherapy of obesity · Rung 3
163	Metabolic and bariatric surgery: selection and co-management · Rung 3
164	Eating behaviour, mental health and weight stigma · Rung 3
165	Food environment measurement, retail and digital delivery · Rung 3
166	Food demand, elasticity and the economics of externality · Rung 3
167	Implementation science and programme craft · Rung 3
168	Protocol, preregistration, reporting standards and research ethics · Rung 3
169	Life-stage obesity: paediatric, adolescent and pregnancy · Rung 3
170	Ultra-processed food and the effects of processing · Rung 3
171	Activity measurement and compositional 24-hour analysis · Rung 3
172	Fiscal instruments: taxation and subsidy design in India · Rung 3
173	Burden of disease, economic evaluation and decision-analytic modelling · Rung 3

174	The Indian health system and its financing mechanics · Rung 3
175	India's data infrastructure, access and complex survey analysis · Rung 3
176	Ethics of obesity intervention and industry engagement · Rung 3

EXPERT

177	Energy balance and dynamic weight-change modelling · Rung 4
178	R, reproducibility engineering and data management · Rung 4
179	Question selection and research strategy · Rung 4
180	Probability and statistical inference · Rung 4
181	Indian regulatory architecture and the live policy agenda · Rung 4
182	Regression modelling for health outcomes · Rung 4
183	Potential outcomes, identification and target trial emulation · Rung 4
184	Adipose tissue biology, ectopic fat and the South Asian phenotype · Rung 4
185	Appetite, reward and the determinants of intake · Rung 4
186	Measurement error theory and correction · Rung 4
187	Causal diagrams, structural bias and sensitivity analysis · Rung 4
188	Circadian biology and chrononutrition · Rung 4
189	Measurement instruments: body composition, expenditure, diet, activity, sleep · Rung 4
190	Genetics, developmental origins and life-course epidemiology · Rung 4
191	Nutritional epidemiology and its pathologies · Rung 4
192	Eating behaviour, mental health and weight stigma · Rung 4
193	Activity measurement and compositional 24-hour analysis · Rung 4
194	India's data infrastructure, access and complex survey analysis · Rung 4
195	Ethics of obesity intervention and industry engagement · Rung 4

OBESITY EXPERTISE · BOOK 10

Epistemics: philosophy of science and calibration · Rung 1

A book on how much to believe. It covers why evidence comes in strengths, how a hypothesis differs from a finding and both from a belief, how to state a confidence in words that mean the same thing to every reader, how to change a position when the evidence changes, and why saying "I was wrong about that" is part of the work.

Book 10 of the series' 196 books, free to read at drharshmaheshwari.com.



Dr. Harsh Maheshwari
MD Community Medicine Resident, AIIMS Raipur
contact@drharshmaheshwari.com
ORCID 0000-0001-8848-1918
drharshmaheshwari.com
Version 1.0 · 14 October 2026

CC BY-NC-SA 4.0 · free to share, not to sell