UNEQUAL

Also by Eugenia Cheng

How to Bake Pi

Beyond Infinity

The Art of Logic

x + y

Is Maths Real?

UNEQUAL

The Maths of When Things Do (and Don't) Add Up

EUGENIA CHENG

Profile Books

First published in Great Britain in 2025 by
Profile Books Ltd
29 Cloth Fair
London
EC1A 7JQ

www.profilebooks.com

1 3 5 7 9 10 8 6 4 2

Typeset by seagulls.net

Printed and bound in Great Britain by
CPI Group (UK) Ltd, Croydon, CR0 4YY

A CIP catalogue record for this book is
available from the British Library.

ISBN 978 1 80522 306 1
eISBN 978 1 80522 308 5

CONTENTS

INTRODUCTION

What's the first thing you think of when you think of maths? It might be arithmetic and times tables – things with clear answers like '$2 + 2 = 4$' and '$6 \times 9 = 54$'. It might be equations where you have to 'find x', by a series of manipulations that you may or may not remember or understand. It might be equations for a straight line in the form $y = mx + c$, or quadratic equations, or equations involving trigonometry. Or maybe some famous equations pop into your head, like Pythagoras's $a^2 + b^2 = c^2$, or Einstein's $E = mc^2$, or, for those who have gone a bit further into maths, Euler's mysterious-looking equation $e^{i\pi} = -1$.

Maths is famous for its equations, but it's about so much more than that. And even the equations themselves are much more than they seem. At first sight, equations seem very clear-cut: they are just situations where something is the same as something else.

Or are they?

When it comes to the kind of maths that is further along the usual education process – abstract mathematics, for example – things start to look different. Deep down, equations are hiding much more beyond just something being the same as something else. It's true that 1 does equal 1, and all that '$1 = 1$' means is that 1 is the same as 1, and likewise $2 = 2$ and $3 = 3$, and so on. But this is not enlightening. In fact, it's rather boring. No new understanding emerges if we declare $1 = 1$; no lightbulbs go off, no doors open, no new possibilities arise, no connections are made. So equations like $1 = 1$, while true, are not the point of maths.

What *is* the point of maths?

The point of maths is to gain new understanding and illuminate different points of view. When things are very obviously the same, it doesn't get us anywhere to know that or state it. It's when things are *not* obviously the same that the situation gets interesting. And this is the secret behind equations: they are not just about things being the same, they are also about things being *different*. They're about the senses in which things are both equal and not equal, at the same time. And things can be equal and unequal in many different ways. Equations are not supposed to be obvious, but they appear obvious to us if we only see the ways in which the two sides are equal and forget about the ways in which they're not. If we think the equation is obvious, we might proceed more quickly to a straightforward answer, but we will miss some rich opportunities for understanding along the way.

When equations are about numbers, 'equality' does have quite a straightforward meaning, but maths is about many more complicated, interesting, fun and illuminating things than numbers. Maths is also about shapes, symmetries, logical ideas, and more. If we start considering those other objects and the different worlds they live in, 'sameness' and 'difference' open up a whole new world of understanding. As mathematical ideas get more complex, so does the concept of 'equality', and the boundary between things being equal and unequal becomes more of a grey area, or perhaps a rainbow of beautiful, vibrant, subtly nuanced colours. A big part of my research field of category theory is about finding more nuanced and insightful ways to say when things can count as the same even though they're different, and this involves making choices about how we want to view the things around us. It's a delicate balance, a dance that is too often hidden in our haste to find clear right answers. But the dance is beautiful, it's fascinating, and it can help us think much more clearly about not just mathematics but our lives and the world.

In this book I am going to introduce increasingly complex and subtle mathematical ideas, and the increasingly complex and subtle approaches to sameness and equality that go with them. Maths is an abstract thinking ground, a place to practise developing our thought processes, a controlled environment for brain work. By contrast, life is far more complex and nuanced than any mathematical ideas, and so our approaches to sameness and equality in life need even more nuance. Understanding these subtleties in mathematics is a powerful starting point for us to understand them in life too, so that instead of oversimplifying into black-and-white arguments, we can look for and embrace the grey areas of equality where different potentially valid points of view can live.

Maths is about much more than an equation on a page, and even an equation on a page is more than it seems. It's much more interesting, and much more powerful, when we see that, in fact, = does not just mean =. We see that mathematical thinking is not just about rigid answers, but about flexibility, and the discipline of seeing different points of view and holding them in our heads at the same time. We see that almost everything can be considered equal and unequal at the same time, whether it's numbers, shapes, patterns, transformations, structures, words, meanings or people. And it's up to us what we do about it.

Mathematics isn't a series of rules, facts or answers. It's an invitation into a way of thinking. Welcome in! I'm glad you're here.

1

WHEN ARE THINGS THE SAME?

In maths it might seem that equations and equalities are very clear-cut and unambiguous. However, in this book I am going to argue that this is an oversimplified view of what maths is, and that, actually, there is plenty of grey area around the concept of equality, even in maths. Much of the power, the efficacy and the fun of maths comes from the very fact that sameness and difference are hazier notions than they might first appear. We are probably more used to having grey areas around what equality means in life, but if the idea seems strange in maths I think that's a sign that we're presenting maths as too separate from normal life. Unfortunately, approaches to fixing that problem often involve shoehorning maths problems into very contrived 'word problems' about life; it is less common to look at how we intuitively think about things in life and then relate that to maths. But abstract maths is often about investigating our intuition and giving it a rigorous framework, and equality or 'sameness' is a key example of that.

Quite often in normal life we say that things are 'the same' but we don't mean it entirely rigidly, and we mean different things in different contexts. We are actually extraordinarily flexible about what 'the same' means, while also being remarkably precise in our contextual understanding and interpretation.

Take cats and dogs, for example. Most children learn to recognise cats and dogs when they're very young. We can point out a cat to a small child, and they will quite reliably be able to point out other cats afterwards, not mistaking them for dogs. So they know which animals are 'the same' animal, and which ones are different. But what is it that makes some animals the same animal and some different? It's quite hard to articulate that. You could say it's about the sound they make, but children can tell cats and dogs apart even if they're not meowing or barking. Perhaps it's the way they move; but children can tell them apart even if the animals are asleep or are just in pictures. In fact, children can distinguish cats and dogs even in cartoon-like drawings, which are not at all realistic but somehow capture the essential cat-ness and dog-ness. That is, after all, one of the great skills of a cartoonist or any kind of visual artist: to capture the essence of something without necessarily depicting all its details.

I think that's the key point: that we as humans are able to see beyond details, deep into the essence of things. This means we can say things are 'the same' when they are the same in essence, even if they differ in some details. Importantly, we're also able to be flexible about what level of detail we count as important in any given context, so that the boundary between 'surface details' and 'deep essence' is itself movable. It's not that we're being ambiguous, it's that we're recognising different things as important in different situations. If you see posters about a lost cat, you'll be able to be much more specific about whether another cat you see is 'the same' as the cat in the picture. But at some level of detail it becomes much more like a learnt skill. I'm not at all a cat expert, so I might think a cat looks like the one in the picture if it has the same general colouring, but someone who is much more attuned to the exact features of cats might be able to tell immediately that those are different cats.

The same general idea causes difficulty in learning new languages, when you're trying to tune in to sounds that you're not used to. You might think you've pronounced something 'the same' as your teacher, but perhaps there is some nuance that you're not detecting, and so to a native speaker your pronunciation would sound completely different.

I once attempted to learn Russian but gave up after a particularly frustrating lesson during which I was attempting to pronounce the difficult letter щ. The session with my teacher went something like this:

Me: sh
Teacher: no
Me: sh
Teacher: no
Me: sh
Teacher: yes!
Me: sh
Teacher: no
Me: sh
Teacher: yes!
Me: sh
Teacher: yes!
Me: sh
Teacher: no
Me: sh
Teacher: no
Me: sh
Teacher: yes!

At least, that's how it sounded to my ears: as far as I could discern, I was saying it exactly the same each time, but evidently to her I was saying it differently in some critical way that I just didn't know how to detect.

This is a particularly noticeable issue with tonal languages, for someone who is trying to learn them but whose first language is not tonal. In English, for example, the word 'soup' means the same thing whether we say it low, high or with a scooping up tone because we're indicating a question. However, in Cantonese the word for soup sounds something like 'tong', but you have to say it with the right pitch because otherwise you might be saying sugar, or iron, or candy, or fillet. In one language 'same' just means the vowels and consonants, but in another it also means the pitch; conversely in Cantonese the consonants 'l' and 'n' are more or less interchangeable, so 'lay' and 'nay' both mean 'you' (if said with the right pitch) but in English those are different words.*

These subtleties around sameness and difference come to the fore if we think about asking computers or artificial intelligence (AI) to do certain things that we take very much for granted, such as recognising cats and dogs; AI was very bad at that for a long time, needing to be trained on thousands and thousands of images rather than the one or two that children need. Understanding handwriting is also very difficult for AI. It's a bit of a mystery to me how we're able to recognise so many different renditions of a letter as that letter. For example, you can probably recognise all the following as a letter 'f', but perhaps you'd be hard pushed to describe in words how you do that:

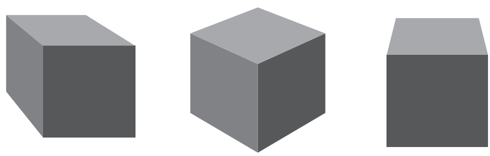
Context definitely helps: we can recognise a letter in a word in a sentence more easily than out of context. This makes shopping lists harder to decipher than a paragraph in a letter – though perhaps nobody writes either of those by hand any more. It does make it harder for children to learn to write, when they haven't yet understood what the tolerances are for something to count as 'the same' as the image they're trying to reproduce as the letter 'a'. So a crudely drawn 'a' counts, but a perfect reproduction of a typed 'a' in mirror image does not. On the other hand, if you look at yourself in a mirror, you can still identify yourself as you even though the image is flipped. You can also probably recognise yourself upside-

* However, apparently 'nay' is more refined-sounding than 'lay', so in that sense they're not interchangeable.

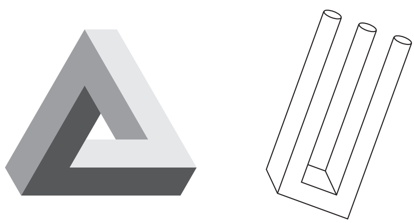
down, or from the back, but if you're trying to recognise someone you know less well that might not work.

Changing literal point of view

When we are looking at things that are less complicated and more predictable than humans, we are adept at extrapolating from what we see, to work out what's going on where we can't see. For example, we can recognise a cube from many different angles:



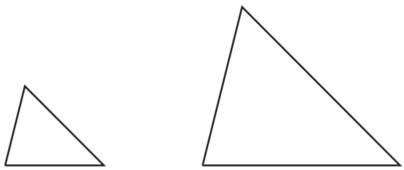
However, we can be tricked by optical illusions such as so-called 'impossible objects'. These are objects we can draw in two dimensions and which our brain spontaneously tries to interpret in three dimensions, except that the object in three dimensions would be impossible. My two favourites are the impossible triangle and the impossible fork.



Impossible structures were made famous by the artist M. C. Escher, but they were pioneered much earlier in 1934 by Swedish graphic artist Oscar Reutersvärd when he was just eighteen years old and still at school.

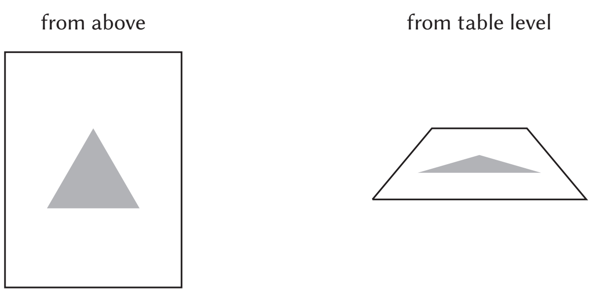
If we just think about what an ordinary (possible) triangle could look like from different directions, we might think that the triangle could be rotated but all its lengths will stay the same. However, if we move further and further away from the triangle, it will look smaller and smaller. Our brains may be able to adjust for the distance and understand that the triangle hasn't actually shrunk, but only if there is some context to help us.

When the triangle looks smaller it will still have the same angles, it's just that the lengths of the sides will have shrunk.



Triangles with the same angles as each other but different overall sizes are called 'similar' in maths, whereas those that have exactly the same size and angles are called 'congruent'.

If we also think about looking at a triangle obliquely, rather than straight on, we will find that the angles change as well. An equilateral triangle flat on a table will only look equilateral if we look at it from above; if we put our eye close to the table and look across it, the triangle will appear wide and short.



This principle is invoked when warning words are painted on roads. Drivers who see them aren't actually eye-level with these markings, but if they're driving at speed, then, effectively, the surface shrinks in the direction of travel according to how fast they're moving, so the words need to be elongated in order to be legible. The word below looks very elongated, but if it were on the road and you were driving over it, it might just look like normal text.

SLOW

We might find that all triangles can look like each other if we view them from different angles, so perhaps all triangles are in some sense the same. We'll come back to this – and we'll find that in some mathematical contexts all triangles really do count as the same.

Projective geometry is a field of maths that studies what shapes look like when projected in different directions. It grew out of the study of perspective by architect Filippo Brunelleschi in the fifteenth century. Artists had been coming to realise that in order to make their two-dimensional paintings of the three-dimensional world look more realistic, they needed to adjust the angles, so that things with right-angles in the world would not necessarily end up having right-angles in the painting. Brunelleschi and his contemporary Leon Battista Alberti were the first people to undertake a systematic study of this. The techniques can be counterintuitive, but once you get used to it you know that if you're drawing a cube in perspective, then the back square face needs to be smaller on the page than the front square face, even though they are representing squares that are the same size: things can be the same in one sense but very different in another.

Changing mathematical point of view

Understanding how the same thing looks different from different points of view is an important part of maths. This is very different from the sadly widespread view of maths as something fixed and rigid. Take the idea of the 'commutativity of addition', for example, which says that adding numbers together gives the same answer whichever way round we do it; for example:

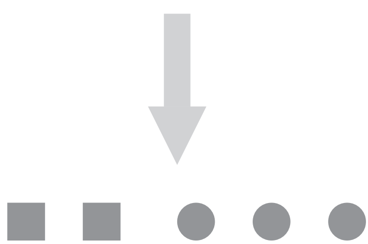
$$2 + 3 \quad = \quad 3 + 2$$

This can be thought of as just a 'fact' or a fundamental truth about numbers, but I prefer to understand it as seeing things from a different point of view – quite literally, if you are doing your arithmetic using counting blocks.

The left-hand side of that equation is 2 + 3 and it might look like this in counting blocks:

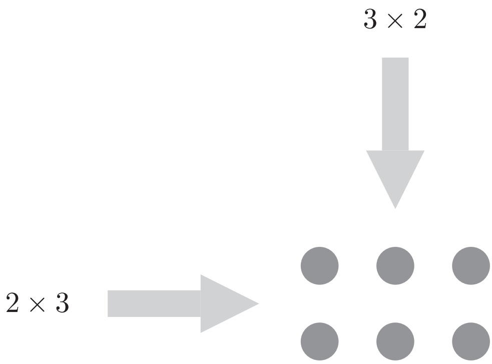
whereas 3 + 2 looks like this:

At first glance, the two look different, but the second version is just what you would see if you were to walk round to the other side of the first set-up and see it from there:



I like demonstrating this by really walking around, to make vivid the point that we are changing our point of view while also being sure that the configuration isn't changing.

We can do something similar to think about multiplication too: 2×3 and 3×2 are different views of this grid:



If you understand the change of point of view that has happened, then you can get a better understanding of the sense in which the two things in question are the same; and this is true in life as well. We could do with more skill at seeing things from other people's points of view. That's not usually thought of as something to do with maths, but I think it really is, because of how carefully maths deals with different views of the same concept. However, sometimes we might get so used to the different views counting as 'the same' that we forget that the shift in viewpoint is happening, such as if we are so sure that 2×3 and 3×2 give the same answer that we forget how they're different. This sameness and difference comes up if we think about the prime factorisations of numbers.

Prime factorisations

Prime numbers are the basic building blocks of numbers, using multiplication. This means they are defined so that 'every number can be expressed as a product of prime numbers in exactly one way'. However, in order for 'exactly one way' to make sense, we have to declare that changing the order in which we write down the product doesn't count as different. For example, 6 can be expressed as a product of prime numbers as 2×3 or 3×2. But those aren't *interestingly* different, so we declare that they don't count as different.

It's up to us what we count as the same and different, and under some circumstances 2×3 and 3×2 really do count as different. I mean, they are in fact different expressions, for a start: they use symbols in a different order. They also mean different things: if we are counting packets of cookies, then 2×3 is two packs of cookies with three cookies each, whereas 3×2 is three packs of cookies with two cookies each. Those are different, even though the total number of cookies is the same. (One uses more packaging than the other.)

But when we're talking about prime factorisation, we're just trying to understand, abstractly, how numbers break down into smaller parts. Worldly issues such as packaging don't come into it. Calling 2×3 and 3×2 different wouldn't be very illuminating from that point of view. Moreover, it would mean that we can't make the key statement about 'unique prime factorisation', and that statement is interesting and illuminating. So we invoke the more forgiving notion of sameness in order to say something insightful. This idea of choosing a more forgiving and therefore insightful sameness comes increasingly into play with mathematical concepts that are more complicated than numbers, such as patterns.

Sameness of patterns

Here are two 4-by-4 grids filled in with a particular pattern. Can you see the pattern, and do you consider that the two grids have the same pattern?

0	1	2	3
1	2	3	0
2	3	0	1
3	0	1	2

a	b	c	d
b	c	d	a
c	d	a	b
d	a	b	c

They are visibly not *exactly* the same because one has numbers and the other has letters. But the general pattern is the same, because each one has diagonal stripes running from top right to bottom left. One way of expressing this idea more formally is to make a direct translation from the left-hand pattern to the right-hand one by replacing the numbers with letters according to the following scheme:

$$\begin{array}{ccc} 0 & \longrightarrow & a \\ 1 & \longrightarrow & b \\ 2 & \longrightarrow & c \\ 3 & \longrightarrow & d \end{array}$$

This idea of being able to make some direct superficial replacements to 'translate' one thing into another is a key notion of mathematical sameness that we'll come back to.

That's all very well, but do you think this next pattern is also the same?

a	b	c	d
d	a	b	c
c	d	a	b
b	c	d	a

This one still has diagonal stripes, but they are now pointing the other way, from top left to bottom right.

There is no correct answer to whether or not this is the same pattern as the previous one because it really depends what you care about in this particular moment. What we count as 'the same' involves choices: it's not something fixed and absolute. Mathematics is about making precise arguments, so instead of just going 'I think this is the same' only for someone else to say 'I think this isn't', we try to come up with precise definitions of what we mean by 'the same' at any given time. This means it is at least incontrovertible whether two things count as the same *according to that particular definition*, even though there could well be other definitions according to which we conclude something else. In any given context we choose how much of the sameness to focus on, as well as how much of the difference. This is very explicitly the case when we are dealing with approximation.

Approximation

Approximation is one approach to things being the same in a more nuanced and forgiving way than just equality. The idea of a grey area around equality might seem strange in a mathematical context, but approximation is something we do all the time in our daily lives, possibly without really thinking about how much we're doing it. It is forced on us by the messy and complicated nature of the world, by our limitations as observers, and by the inaccuracy of our measuring equipment or our brains. This makes it sound like a constraint, but it's also a superpower: the idea of approximation is made possible by our brains' amazing capacity for accepting difference as part of sameness, even if we don't realise we're doing it.

As with many mathematical concepts we can start by thinking about it for numbers, and come up with ways to say that a number is 'approximately the same as' another. But we can then apply it to other things as well, to use this way of thinking far beyond numbers. Is a biscuit approximately the same as a cookie? It depends what country you're in; in the US a biscuit is approximately the same as a British scone, but not exactly the same, and British people are liable to get upset about things masquerading as scones if they're not exactly how a scone 'should' be. Still, some people are more forgiving about scones than others. (I'm not very forgiving at all.)

Here are some pictures of hand-drawn shapes. The first one is my good-faith attempt to draw a free-hand circle. As the shapes progress they become less and less like perfect circles. Which ones would you be prepared to accept as circles?

It probably depends on the context. If I'm cutting out a circle of baking parchment to line a cake tin, I want it to be fairly accurate because it bothers me if there are parts of the circle sticking up getting in the way of the batter, but it doesn't bother me so much if there are small gaps at the edge of the base. However, if I'm giving a talk about symmetry, then it's better if the circles I have on my slides are genuine circles, so that they all have the symmetry I'm trying to describe. But even here I am being slightly approximate – there can't be a *genuinely* genuine circle on my computer screen as the pictures are made of pixels, and if you zoom in far enough you'll see that the picture is full of jagged right angles. However, since we can't see them when zoomed out, that doesn't bother me, though it might bother other people – we all have different tolerances, and it doesn't mean some of us are right and some are wrong.

Sometimes we can get fixated on particular features or details of difference for reasons that aren't entirely rational, but that doesn't always mean the reasons are invalid. I get upset with American 'scones' because they're often not round, but triangular. I see that cutting scones into triangles is more efficient, because when you cut out circles you end up with bits left over in between the circles. If you don't want to waste that dough, you then have to re-form and re-roll those leftover parts, and the second and third rollings are never as good as the first. However, I prefer my scones to be delicious rather than efficient, and I personally don't like the tough, crunchy corners of a wedge-shaped scone. Furthermore, a scone is not just food to me but an emotional connection, and if it's the wrong shape it doesn't make that emotional connection in my head. Food is of course a much more expansive and multidimensional concept than numbers, so it's easier to come up with a rigorous concept of approximation for numbers than for food.

There are various reasons we might want to approximate numbers. Perhaps we don't know the exact number. Perhaps we can't measure the exact number beyond a certain accuracy. Perhaps we don't *need* to know the exact number beyond a certain accuracy, so it's not worth the effort of being any more accurate than that. It all depends on context. If I'm making a cake, my measurements don't need to be that exact, despite the idea that baking has to be very accurate. I try to measure my ingredients to within about 5 grams, but there have been times when I've done it entirely by feel because I was somewhere with no scales, and the cake was completely fine. By contrast, if you're a medical professional measuring a dose of drugs for a patient, then you will need to be a lot more accurate than that, especially if the drug is potent.

What could it mean for two numbers to be approximately the same? There are two main possibilities that we humans use. Each has different benefits and different drawbacks, so when we use them we should be aware of the choices that we're making and the consequence of those choices. This is a lesson we keep in mind as the maths becomes more complicated and we are making ever more choices about what to count as the same.

Approximation by tolerance interval

One approach to approximation is to say that two numbers will count as roughly the same as each other if they're within a certain distance of each other. We can then pick what the tolerable distance is, depending on how accurate we want to be. We might say someone is about the same age as us if they're no more than two or three years different; however, if we're talking about babies we might need them to be at most two or three months apart, or even two or three days if they're very young indeed. I am going to refer to this sort of method as approximation by 'tolerance interval' because we start by specifying what size of difference we are going to 'tolerate' within our concept of 'approximately the same'. We might pick it to be 10 grams if we're weighing ingredients for a cake, or 2 grams if we're weighing ingredients for macarons, which are much more sensitive than cake.

This is fine if you're just baking something or talking about two people. However, if you try to build longer arguments with this concept you start running into trouble as the distances pile up. You might think you're roughly the same age as someone who's three years older than you, and they think someone is roughly the same age who's three years older than them, and so on, and in a few steps there's a twenty-year gap.

The question of whether it makes sense to build up long strings of 'sameness' is something we will keep in mind throughout this book: it's called the question of *transitivity*. If we can make a long string of things that are 'roughly the same' and at the end we still have something that's 'roughly the same', the concept of sameness is said to satisfy transitivity, or we say it is transitive.

I think this issue is why it's so easy to overeat by just eating a tiny bit more, repeatedly. Because if you eat a tiny bit more it feels like it's not *really* more, but rather, the total you've eaten is more or less the same as before you ate that tiny bit. And then you eat a tiny bit more, which doesn't *really* make a different total from the previous amount, and then you keep going until you've eaten twice as much as your body actually needed. At least, I do. Some people don't seem to be prone to this at all.

This is also how people can steal large amounts of money without their victims noticing – they just skim a tiny bit off a large number of accounts, or overcharge by a tiny amount. If the amount is small enough, it won't be noticeable to the victim, but the criminal can accrue a substantial amount of money because those small parts add up to something large.

While those are negative outcomes, I think we can also turn this concept to our advantage to change our lifestyle for the *better*, by changing it just a tiny bit repeatedly. Each tiny change can feel like it's not really that different, but if you keep making tiny changes repeatedly you eventually get to a lifestyle that's dramatically different from the one you started with. I did this to make my lifestyle healthier, gradually getting used to eating smaller portion sizes and a larger proportion of vegetables, so that now I fill almost my entire plate with vegetables. Twenty-five years ago that would have horrified me.

There are many cases where we implicitly have different tolerance intervals around bigger numbers. A thief may be able to skim more money off a bank account with a large amount in it without someone noticing, rather than from someone who is counting pennies (although some very rich people have become rich in part by taking great care of the pennies). Similarly, eating an extra bite or two of dessert will make much less difference to someone who weighs more or who has a very high metabolism or uses up a lot of energy as an athlete, than to someone who is tiny or has a very slow metabolism.

Perhaps the tolerance interval we allow in an approximation should be a proportion rather than a fixed amount. This method is sometimes used in the rules for election recounts. A recount may occur if the totals for the two leading candidates are close enough together that an inaccuracy in counting could result in a different winner. In the UK there does not seem to be a blanket rule for when there must be a recount, but several US states do have such a rule. In most states with an automatic recount rule, the threshold is 0.5 per cent, that is, there must be a recount if the difference between the two leading candidates is less than 0.5 per cent of votes cast. In some states there's a fixed number that can trigger a recount as well, such as in Delaware where the threshold is 1,000 votes or 0.5 per cent of votes cast, whichever is smaller.

One of my favourite ever election stories was the Winchester by-election after the UK general election in 1997. At the general election, the Liberal Democrat Mark Oaten was declared the winner by just two votes (26,100 versus 26,098) following repeated recounts and arguments about spoilt ballots. However, the candidate who came second was the Conservative Gerry Malone, the sitting MP. He challenged the result in court and the difference of two votes was deemed not certain enough, given that fifty-four ballots had been declared invalid. So there was a re-run in the form of a by-election. In the re-run, Mark Oaten not only still won, but he won by a rather hilariously large 21,556 vote margin. Perhaps this demonstrates the rationale behind voting systems in which the leading two candidates proceed to a second-round run-off election.

Using proportions instead of a fixed amount to measure differences addresses the need for a sliding scale across values, but it doesn't fix transitivity: suppose we fix the acceptable difference to be 1 per cent, then we can still find A, B and C where A and B are roughly the same, B and C are roughly the same, but A and C aren't. We can fix some of the issues of tolerance intervals by using rounding instead.

Approximation by rounding

When we use rounding, instead of saying two numbers are roughly the same if they are within a certain distance, we round every number up or down to fixed reference points, and then everything that rounds to the same reference point counts as the same.

For example, we could round everybody's age to the nearest multiple of 5. So everyone from 18 to 22 we could call 'roughly 20', and everyone from 23 to 27 is 'roughly 25', and so on:

18 to 22	roughly 20
23 to 27	roughly 25
28 to 32	roughly 30
33 to 37	roughly 35
38 to 42	roughly 40

⋮

We are actually already doing that when we refer to everyone's age by a whole number instead of saying 'They are 33 years 4 months 13 days 5 hours 3 minutes and 41 seconds old. Oh wait, it's 46 seconds now.'

I always find it endearing when small children give their age more precisely than adults do, because when you're four-and-a-half, that feels *very* different from merely being four. We adults can use more accuracy when it matters too. For decimal numbers we can round to any number of decimal places we want. If we are rounding to one decimal place, we are declaring that we really don't need to know a number more accurately than to the nearest tenth. When I'm converting between a round cake recipe and a square one, I usually take π to be 3.14 because that's plenty of accuracy for a cake. Actually I could probably take π to be just 3 and it would be fine.

We can round to any chosen number of decimal places, which means we look at the next decimal place after that one to tell us if it should round up or down. If we don't want this to be centred around whole numbers or decimal places we can set the bands to be anywhere we want, such as when you're asked for your age on a demographic survey and it gives you bands to choose from like this:

20–24
25–29
30–34
35–39
and so on.

It always feels a bit galling when I've just crossed over into a higher one: this is one of the anomalies that we incur by switching from tolerance intervals to rounding.

This anomaly causes great frustration to students when we decide on grade boundaries for exams. When I was last grading with grades, we were allowed to set our boundaries wherever we wanted, so it might be something like this:

A	=	90–100
A–	=	80–89
B+	=	75–79
B	=	70–74
B–	=	65–69
C+	=	60–64
C	=	55–59

The last time I was marking undergraduate exams in the UK the boundaries were:

First	70–100
2:1	60–69
2:2	50–59
Third	40–49

Then, of course, it becomes contentious around the boundary. If someone gets 69 they feel very hard done by and think it's unfair that they were only 1 away from a first but didn't get a first. How can we be so sure that they didn't deserve a first but someone with one more mark did?

And herein lies the issue with this way of attempting to fix the problem we had with tolerance intervals. Yes, we do fix the problem of transitivity, as this form of 'more or less the same' is now transitive. But now we have ended up with a different peculiar anomaly, where things that are actually very close together end up counting as far apart where they're right on a boundary. For example, 2.49 rounds to 2, but 2.5 rounds to 3, even though they are only one hundredth apart. Meanwhile, 1.5 and 2.49 round to the same thing even though they are almost 1 apart.

This leads to all sorts of issues at the boundaries, including students being cross about missing an A by 1 point, or someone's disability benefit being cut off when they work 1 more minute per week, or people being ineligible for social support because they earned 1 more pound per year than someone else who was eligible.

One way we fix it for more innocuous situations like age brackets is to be more vague in our terminology. We might use the terms 'early thirties', 'mid thirties' and 'late thirties' without being specific about exactly where the cut-offs are. The Association for Women in Mathematics has produced a lovely pack of mathematical playing cards, called EvenQuads, with each card featuring a picture of a woman mathematician. I'm honoured to be one of them, but am amused that my vague birth date is listed as 'late twentieth century' whereas a friend of mine who is only a year older is listed as 'mid twentieth century'. Funnily enough, I think 'late twentieth century' actually makes me sound older than I am, but I can't really put my finger on why that is, except perhaps my general paranoia about ageing, together with the rational fear of how society judges and dismisses ageing women.

We tried to fix this issue for exam grade boundaries at my job in the UK by holding epic faculty meetings where we discussed each individual case, and if someone was within 2 marks of the boundary we usually bumped them up over it. However, this just created a new anomaly: people who got 68 were bumped up to a first, while people who got 67 weren't, but they were only 1 away from being bumped up so they felt hard done by. We felt like we were making the boundary more forgiving, but in effect we were just moving it. One might then argue that the people who were within 2 of the new effective boundary also deserved to be bumped up, but then where would it end?

When I was grading exams in the US and had more autonomy over how I did it, I tended to just make sure nobody ended up with a mark that was close to a boundary, so that they wouldn't feel frustrated and I wouldn't have to get into arguments with them about it. But then I came up with an entirely different system which was to decide who deserved what grade based on descriptors rather than accruing points. The descriptors were things like 'Has a strong grasp of logic and can build rigorous formal proofs', 'Has a reasonable grasp of logic but writes proofs that have small logical gaps in' and so on. I would pick descriptors to assign grades, and then assign a numerical mark after assigning the grade (so basically, backwards from the usual way) and made sure that everyone was in the low to mid range so that they didn't feel there was a near miss with the next grade up.

I now have an even better method than that: I teach at the School of the Art Institute of Chicago, which has no grades at all so I don't even have to deal with this issue any more.

Another issue that can happen with rounding is if you round too early in the process and then start substituting your answer into further calculations. Substitution is one of the fundamental principles about equality that we will discuss in the next chapter. It is different from the question of transitivity, and is more about whether your notion of sameness is 'preserved' by doing various operations on it.

Preservation

When we are talking about approximate sameness of numbers rather than exact sameness, we have to be more careful about whether it will be maintained by various operations.

For example, take two numbers that are within 1 of each other; if we add the same amount to both of them they will still be within 1 of each other. But what if we multiply? For equality this is not a problem, but it is a problem if we take two numbers that are only *approximately* the same, say 2 and 2.5. If we multiply them both by 10 we get 20 and 25, which are no longer within 1 of each other.

We say that this notion of 'approximate sameness' is not *preserved* by multiplication by 10. That is, we can start with two things that *are* approximately the same (according to the definition in question), but if we then multiply them both by 10 the results are no longer approximately the same (according to the same definition in question).

This issue still arises if we do rounding instead of tolerance intervals. For example, 2.4 and 1.8 both round to 2, but if we multiply each number by 10 we get 24 and 18, which do not round to the same whole number.

We also run into trouble if we round some numbers before adding them together. For example, if you do 2.4 + 2.4 you get 4.8, which rounds to 5. This means that if you round too soon in your calculation you'll find yourself doing 2 + 2 and getting 4, whereas if you had rounded later you'll get the answer 5.

This anomaly is known in some circles as 'WeightWatchers maths'. The famous (but also vilified) weight-loss programme calculates a number of points for any given food, based on its nutritional content. The idea is to simplify a large amount of information into a simpler single number, so that you don't get overwhelmed when trying to decide what to eat each day to fit within a certain quota. Regardless of what you think of this system, the rounding errors make for some amusing 'maths'. Each calculation is rounded to the nearest whole number, so perhaps one slice of bread should have been 2.4 points but gets rounded down to 2. Then in the app if you put in two pieces of bread it logs 5 points, and everyone feels hard done by and says '2 + 2 = 5 in WeightWatchers maths'.

I remember being frustrated in physics lessons at school by the whole issue of decimal places, and trying to remember that if you're aiming for an answer that's correct to, say, three decimal places, you have to use earlier answers that are correct to perhaps four or even five decimal places. But I could never quite remember how accurate you needed to be and when, and I still can't; I was much better at maths when it had letters instead of numbers and we didn't have to deal with these rounding issues. Many people feel horrified when the 'numbers turn into letters' in maths, but I am relieved and even delighted.

Choices

There are no right and wrong answers when it comes to approximation. Perhaps I should say this more carefully – there are possible wrong answers, but only if you go against logic, or say you're doing something you're not (or measure a dangerous drug too inaccurately). It can turn out 'wrong' if you say you're doing something to a certain accuracy when you're not, for example. It would be wrong to claim that some notion of sameness is transitive when it isn't, or to claim that it's preserved by an operation when it isn't, or that it works with substitution when it doesn't. But within those logical constraints there aren't right and wrong ways to approximate things, nor are there right and wrong answers to what can count as the same as something else or not.

Instead, there are just choices we can make, and consequences to those choices. It is up to us to decide what we want to count as the same and not the same, but we will then have to put up with whatever consequences come from those choices, so it's better to be aware of what those consequences are, whether it's non-transitivity or weird boundary behaviour. In different situations different strange behaviours are more tolerable than others. Understanding the differences in the much simpler context of numbers can help us be more ready for them when they come up in life situations that are much more complex.

For example, I wish we had a better way to deal with anomalies about the thresholds for benefits and exam grades. I think if we were all better at dealing with sliding scales we could do that. Even if we think students need to be assessed, it's not clear why we have to divide them into four distinct categories rather than just putting everyone on a more continuous scale. I'm generally not in favour of the American SAT system, but one thing in its favour is that it does have a much larger range of possible scores (400–1,600) so it's more like a sliding scale.

Another more complicated scenario involving dividing people into distinct categories comes into play when we think about political parties. Even though this is no longer about numbers, similar phenomena arise from our attempt to group people into separate parties according to them having 'similar political beliefs', particularly when there are very few parties to choose from, as in two-party systems.

One way to group people is similar to the tolerance intervals for numbers – maybe we say that people have similar political beliefs if they are within a certain distance of each other, politically. The first problem is that it's hard to know how to measure that as there are so many different dimensions, and people's beliefs may be very similar about, say, healthcare, but be very different about university tuition fees or foreign policy.

But the next problem is the non-transitivity: person A might have similar beliefs to person B, who has similar beliefs to person C, who has similar beliefs to person D, and so on, but if you keep going in one direction, then by the time you get to person *Z* they're very different from person A. If you then extend that across thousands of people or millions of people, everything can – and does – become extremely incoherent. It can result in different factions of one political party potentially being even more opposed to each other than they are to the other party, especially when the furthest-right portion of a left-wing party is closer to the furthest-left portion of a right-wing party.

An unfortunately much more coherent approach to party politics is to unite around one core common cause or a common enemy and ignore other issues. This is more like rounding to the nearest whole number. If you form a party around the principle of being anti-abortion, or anti-Europe, or anti-immigration, then you automatically achieve transitivity. That is, you might define this:

Person A and Person B have similar political views
if they are both anti-immigration.

This relation is then automatically transitive, and so it can unite a party much more easily. This is often something that right-wing parties do more successfully, whereas left-wing parties try to find similarities by something more like a 'tolerance interval', and end up fighting each other and unable to unite.

The trouble with the approach of 'uniting around one core principle' is then trying to get anything done that isn't related to that core uniting principle. This might seem a long way from maths, but the principles at work are the same, and maths really is about understanding principles so that we can apply them very broadly.

Perhaps, like with exam grades, it would be more sensible just to stop trying to divide people into clearly delineated parties like that, or at least to have more parties so that the scale can become a bit more sliding and less discrete; this is more the case in some countries other than the ones I know best (the UK and the US). However, this would require a huge change to both the political systems and to the general mentality of all the citizens.

All this discussion of approximation is a key example of how we can look at the senses in which things are both the same and not the same. We might examine a situation and decide that some things are within tolerances of being similar, but then it turns out we're looking at the wrong criteria. This happens when non-chess-players think a chess configuration is similar to another because it is physically similar, but it's actually strategically completely unalike.

Thinking about sameness always involves choices about what we're going to see as the same and what we're going to see as different. It involves choices about which differences we're going to see and which we're going to disregard. This is something we do all the time in our daily lives, to greater and lesser extents in different contexts, depending on our personalities, our preferences and our tolerances; honing our ability to do this in mathematical settings helps us to notice the choices we're making in life too.

Once we start seeing the nuances of 'sameness' and 'difference' we realise that there are grey areas everywhere, and there are important lessons we can learn from that about the world around us, especially if we acknowledge that sameness involves choice rather than treating equality and inequality as immutable facts. This can enrich our lives in many ways that aren't obviously to do with maths. For example, we can listen to the same piece of music but have different experiences from hearing different performances, which is why I love going to live concerts so much more than listening to recordings.

If we all listen to the same piece of music, look at the same painting, eat the same food or swim in the same sea, there is a shared experience, but if we describe it in words we will all come up with different things. There might be some overlap, but some people might have found it thrilling whereas others found it boring. Some people might want to do it again as soon as possible, and some might want to avoid it for the rest of their lives. However, it was still a shared experience. On the other hand, sometimes I have found nothing quite so alienating as being in the same place as a group of people, doing the same thing as them but feeling completely differently about it from everyone else.

I don't have to describe my feelings the same way as my friends in order to value sharing an experience with them. Perhaps the very fact that we can end up with completely different words to describe the same experience makes our lives rich and rewarding. I think this is what makes maths rich and rewarding too: that there is always an interplay between sameness and difference. That interplay is a whole world to be explored, and that's what I'm going to do in this book.

We will think about the interplay between sameness and difference in maths, and we will go on to consider sameness and difference in life, and when differences do and don't matter. When does it matter what someone looks like? If you need to cover for someone at work it only matters that you can fulfil all the roles they do; it doesn't matter if you look like them or not (unless you're working as a look-alike or a stunt double). When should people count as having the same qualifications as someone else? When should they count as having the same experience? What should gender pay equity mean? What does it mean to treat people fairly?

We can think about whether it makes a difference if we buy the same thing at a local family-run shop or from a giant chain where everything is cheaper; we can consider what kind of shops we choose to spend our money at, and weigh this up against our desire to spend less of our own money. And we can also choose the ways in which we treat people the same. It is entirely true that no two people are exactly the same – nobody is the same as anybody except themselves. But it's not very helpful to take either of the extreme views 'everyone is entirely different' or 'we're all entirely the same'. In maths one extreme would be to declare that no equations hold except ones of the form $5 = 5$; the other extreme would be to declare that everything is equal because everything is 'a thing'. If mathematics did either of those things, it would be rather less interesting, productive or useful than it is. In fact, I hazard that it would be completely boring, unproductive and useless.

Our exploration of a more productive approach to sameness in maths is going to be a journey, starting from concepts that I hope are familiar, but progressing into deep mathematics including open research. I hope that as the journey advances you will rest assured that nobody is expecting you to understand everything. Even mathematicians don't typically feel that we understand anything fully, but rather than this putting us off, it is what drives us forwards.

I don't believe we should keep maths hidden because it is 'too hard'. I hope that even when it gets complicated you will appreciate seeing glimpses of it, or enjoy gazing at it like abstract art. It's in contemplating things we don't understand that we expand our minds. Mine has definitely expanded in the process of me writing about these things and pondering all the things I don't understand either. I will always keep trying to understand more, and I will always keep sharing what I've understood as widely as I can.

2

EQUATIONS

The tension or flexibility between sameness and difference is more obvious when we're thinking about the human experience, or shapes, or patterns. But it is also crucially present when we talk about the most straightforward form of sameness in maths: equations.

Equations might seem like the most clear-cut and inflexible part of maths; in fact, if you think maths is all about equations, and you think equations are rigid black-and-white facts, then you probably think that maths is all rigid and black-and-white. However, throughout this book we're going to see that there is far more to equations than right and wrong answers.

First of all I want to think about the subtly different roles that the little equals sign can play for us. It can be hiding depths that we might not notice because we've become so used to the idea of equations supposedly saying something unambiguous. So, what are the different things that equations do?

Declaring what the answer is

First of all, equations are often used simply to say what the answer is to something. This happens when we're just calculating the result of an operation, like

$$2 + 2 \quad = \quad ?$$