Long-term principles for meaningful teaching and learning of mathematics

David Tall
University of Warwick, UK
david.tall@warwick.ac.uk

Abstract

As mathematicians reflect on their teaching of students, they have their own personal experience of mathematics that they seek to teach. This chapter offers an overall framework to consider how different mathematical specialisms may require different approaches depending on the nature of the specialism and the needs of the students. It involves not only problematic transitions for learners but also fundamental theoretical differences between specialisms. At university and college level, mathematicians may have sophisticated knowledge that they believe will offer enlightenment to their students, but these insights may not be shared either by learners in their own community, nor by experts in other communities.

The proposed framework reveals simple insights related to mathematical thinking that are visible to both teachers and learners, offering new insights into long-term meaningful growth from practical to theoretical mathematics and on to formal definition and proof. It invites readers to challenge their own beliefs to make an informed choice of strategies that respects the needs and choices of different communities that are essential parts of a complex society.

1. Introduction

Every one of us has our own personal history and different forms of expertise. To formulate a long-term overview of the ways in which different individuals make sense of mathematics in different ways requires us to be prepared to question our own beliefs, at least to the extent of gaining insight into the reasoning of others.

The mathematics required for a pure mathematician, an engineer, an economist, or a participant in any other profession, depends on the needs and purposes of each specialism. The last half century or so has seen the development of technological tools that enable us to operate in ways that were unthinkable a century ago. While the tools have changed, the essential structure of the human brain has had little time to develop in a Darwinian evolutionary sense. It is reasonable to assume that human brain structure has remained essentially the same over the last five or ten thousand years which covers the main development of modern mathematics. To make sense of the differing approaches, it is therefore of value to focus on fundamental aspects that underpin human thinking. It is even more valuable to identify features that are easily observed by teachers and learners that may lead to a deeper understanding of long-term development of sophistication in mathematics. Aspects considered here will include:

- how we *speak* mathematical expressions to make precise sense of them,
- how we *hear* someone else speak mathematically,
- how we *interpret* expressions flexibly as operations in time or as mental objects,
- how we *see* static and moving objects as constants and variables in the calculus,
- how we *read* mathematical proofs to make sense of them.

In *How Humans Learn to Think Mathematically* (Tall, 2013), three strands of development termed *conceptual embodiment, operational symbolism* and *axiomatic formalism* were formulated, which develop in sophistication in the long-term. The first focuses on (physical and mental) *objects* and their properties, the second on *operations* on objects and the third on formally defined *properties*. These develop long-term from *practical mathematics* interacting with the world we live in, to *theoretical mathematics* used in society to model and predict outcomes, and *axiomatic formal* mathematics using axiomatic definitions and formal proof.

Prepared for *Mathematicians' Reflections on Teaching: A Symbiosis with Mathematics Education Theories*, (Springer), edited by Sepideh Stewart. Submitted May 22 2019, edited Feb 7 2020.

Our species uses tools to enhance our abilities and means of expression – tools to paint pictures, marks on clay tablets to represent numbers, ruler and compass in Euclidean geometry, quill pens to write books, printing, telescopes, microscopes, logarithm tables, slide rules, and now the development of digital technology, with arithmetic calculation, dynamic visual representations, symbol manipulators, and smart phones with retina displays that we can control by moving our fingers or using our voice.

This chapter will show how features that can be observed by teachers and learners offer new insight into the meaning of operational symbolism and interpretation of visual information.

How we speak and hear mathematical expressions lead to new ways of giving precise meaning to operational symbolism and its communication. This, together with how we interpret expressions as operations or objects, offers a comprehensive theory of meaning of mathematical expressions throughout the whole of mathematics. How we see static and moving objects allows us to imagine points on a number line where static points are constants and moving objects are variables. At a more advanced level of axiomatic formalism, we can prove theorems that provide more sophisticated levels of embodiment and symbolism.

Mathematical thinking has been enhanced by digital technology, performing algorithms in arithmetic and algebra and providing dynamic graphical displays. The recent development of retina displays and new knowledge of how our eyes and brain interpret visual information offer an insightful resolution of foundational difficulties that have confounded our understanding of the calculus for three and a half centuries.

Mathematical thinking changes in different contexts and what may work well in one context may become problematic in another, causing conflict that can impede progress.

The usual approach to the long-term curriculum is to break it down into sub-stages which may be designed by different communities with specified criteria to move from one stage to the next. Summative assessment of successive stages can be helpful. (In my own case I welcomed the need to review what I had studied, to link the whole together in a more coherent framework.) However, it can also cause the teacher or learner to focus on rote-learnt procedures, to pass the test in a short-term manner that may cause increasingly problematic aspects in new contexts over the longer term.

To provide a supportive approach that deals with the difficulties, I propose a long-term *principle of long-term meaningful learning* in which 'the teacher is aware of those ideas that remain supportive through several changes of context, to give confidence to the learner, and to make explicit those ideas that are problematic so that they can be addressed meaningfully.' I name this the 'Simon Principle' in honour of my grandson whose insight led to its formulation.

Our current mathematical knowledge is the product of development of previous generations in different cultural contexts over the centuries, yet we expect our children to grasp sufficient ideas for their own purposes in society in their own lifetime.

We live in a complex society that requires different individuals to use different kinds of mathematics in productive ways. Some may require practical mathematics for everyday situations, some may require specific kinds of technical mathematics in different professions, some may require more theoretical mathematics that enables them to model real life situations to predict possible outcomes, some may go on to more formal aspects of pure mathematics and logic involving set-theoretic axioms, definitions and formal proof.

As professional mathematicians reflect on their own teaching practice, it is important to take account not only the current learning of their students, but also the knowledge structures their students have developed from previous experience. This chapter will reveal details of student difficulties at college that arise from learning at all levels in school and will offer new ways of approach that can be of value to improve long-term learning at all levels.

Different communities may have differing beliefs and needs, to such an extent that their views may be considered to be incompatible, leading not only to so-called 'math wars' but also to disrespect of one community for another. In a complex society it is essential to have an interconnected web of differing talents.

To address the social aspect of differing communities, it is necessary to compare different communities of practice, which Lave & Wenger (1991) describe as 'group[s] of people who share a craft or a profession.' Individuals in a community may have differing personal viewpoints but, overall, they agree (or believe that they agree) to certain shared principles. Each community has 'experts', well-versed in the practices, and 'novices' who are being introduced to them as learners developing over time. Some learners may focus on copying routine practices of the experts, others may develop sophisticated personal knowledge structures that go beyond the practices of the community. It is also important for experts themselves to reflect on their experiences and act as learners seeking new insights.

2. Enlightenment, transgression and multi-contextual overview

To make sense of the developments that occur in a community of practice, it is essential to consider the views of both individual experts and individual learners. Experts are likely to see their role as offering enlightenment into the practices of the community. While some learners may make personal sense of this enlightenment, others may find the situation problematic as if they are faced by a conceptual boundary that they are unable to cross.

The same phenomenon may also occur between communities of practice with radically differing belief structures, each of which makes sense in its own context. For example, a pure mathematician may regard the real line as a complete ordered field which cannot include infinitesimal elements while an applied mathematician may work with 'arbitrarily small' variable quantities which they imagine in their own version of 'infinitesimal calculus'. The beginning of the twentieth century saw a splintering of mathematical beliefs into different communities of practice, such as intuitionism, formalism, logicism and various later versions of standard, non-standard and constructivist views of mathematics.

To formulate a framework that incorporates these phenomena within the development of individuals in a single community and also between different communities, it is useful to consider what happens when one or more individuals from a given community are faced with the possibility of transition to a second community with radically different views.

Initially there is a boundary which acts as an *impediment* to those fixed in the first community, unable to comprehend the views of the other. However, if one or more individuals shift between communities, those in the community that the individuals leave may see the move as a *transgression*, while those in the community that is moved to may see it as an *enlightenment*. This may involve strong differences of opinion between the communities. However, these differences may be considered in a more cooperative light in a *multi-contextual overview* where each community is aware of the values shared between the two, to build confidence based on their communalities while respecting and addressing their differences (Tall, 2019). They may still continue to hold strong opinions about their differences but may now be in a position to listen to each other. This framework applies not only to the differences between communities, it also applies to a learner in a given community faced with a problematic change in meaning from one context to another (figure 1). Some fortunate learners may benefit from the enlightening insight provided for them, while others are faced with an impenetrable boundary.

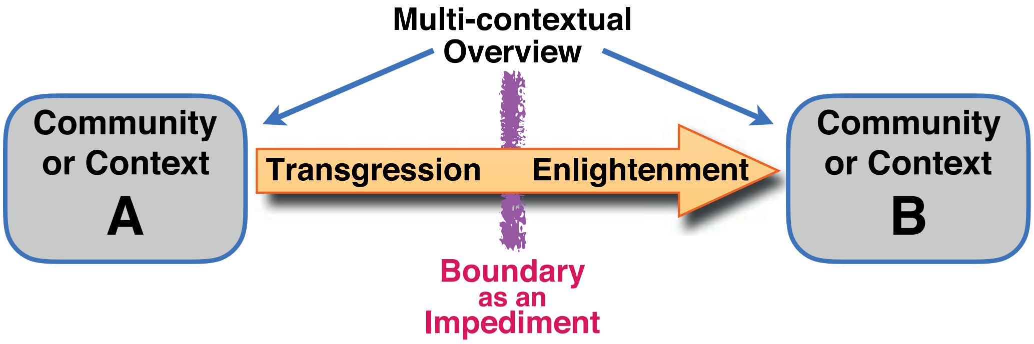


Figure 1: transition over a boundary

In the case of differences between one community and another, the possibilities may be characterised as:

Impediment: *inability* to leave the current community to cross over a boundary
Transgression: crossing *out of* the current community over a boundary
Enlightenment: crossing *into* a new community over a boundary
Overview: encouraging communication *between* communities.

Examples include differences between communities of pure and applied mathematicians, between mathematicians and educators, between politicians who prescribe the curriculum, curriculum designers, teachers and assessors, between different levels of teaching in early learning, primary, secondary, university, and different forms of expertise in mathematics.

In the case of an individual seeking to make a change in context within a single community, the possibilities are:

Impediment: *inability* to change context
Transgression: *unwillingness* to change context
Enlightenment: *ability* to change context
Overview: ability to switch *between* contexts.

Examples include generalising number systems from counting numbers to fractions, to signed numbers, to rational numbers, reals, complex numbers, from arithmetic to algebra, from practical drawing to Euclidean proof, through changes in meaning in geometry (van Hiele, 1986), from school mathematics to university, and so on.

2.1 Questioning personal beliefs

It is natural for academic mathematicians to focus on aspects relevant to their own research and teaching, but this is not sufficient to deal with the broader long-term issues that arise. Mathematics at all levels builds on what was experienced before and affects what happens after. Making sense of university mathematics depends on students' previous experience in school and mathematicians affect what happens in school through their participation in mathematics education. To grasp an overview of the wider picture requires communication between the various participants.

To be able to participate in such a discourse it is important for all participants, *including the reader,* to identify our own sources of personal knowledge and beliefs. *After reading this section, it would be useful for the reader to pause and question his or her personal beliefs*. It is helpful if this includes the mathematical topics where we have expertise, the level at which we have familiarity of learning, teaching and research, and to contrast these with aspects where we have had little engagement to seek to identify our own personal experiences and prejudices.

In my own case, I grew up in a working-class family, encouraged to be competitive in a boys-only Grammar School selected by an 11-plus examination, obtaining a scholarship to Oxford University, being awarded a university prize in finals as an undergraduate, then obtaining a Doctorate with Fields medallist, Sir Michael Atiyah. This gave me a bias towards high-level pure mathematics which colours my views. It includes not only the pleasure of achievement, but also the tensions arising from struggling to make sense of boundaries, some of which I still consider to be problematic.

My first position as a university lecturer made me realise how much I enjoyed thinking about mathematics in a way that could make sense to students and I moved to a position as ‘Lecturer in Mathematics with Special Interests in Education’ at Warwick University in 1969, where I continued as a mathematician and began to research undergraduate mathematical thinking, leading to a second PhD in Psychology of Mathematics Education with Richard Skemp in 1986. I spent a year as a school teacher with 8 to 12-year-olds, and then, as I shared ideas with other colleagues with varied professional experience and took on PhD students from around the world, I gained a sense of differing practices in a wide range of countries.

Over the years, I worked on building a long-term framework covering the development of the individual from birth to adulthood (Tall, 2013). Instead of performing a detailed comparison of many available theoretical frameworks, I took these theories into account by seeking underlying fundamentals that contribute to longer-term development of sophistication in mathematical thinking. This proved to have relevance both in the growth of the individual child and also over the generations in history.

However, the analysis implicitly accepted the broad sequence of curriculum development designed to build ideas over the years with summative assessment that set standards to be attained at key points in transition from one stage of schooling to the next. Could it be that the chosen sequence has an adverse effect on the long-term outcome? More seriously, could the approach chosen by a particular community of practice actually *impede* long-term growth of sophistication? To investigate this possibility, the next step is to study mathematical growth in a manner that is not bound by today’s curriculum decisions. This begins by considering long-term historical evolution before turning to the lifetime development of individuals.

2.2 Long-term historical evolution of mathematical thinking

The historical evolution of our species has developed at an accelerating pace over the last two million years or so, with many aspects that are shared with other species. These relate to how we use our perception to input information, our action for output, and internal processes in our brain to make immediate, often unconscious, decisions. I refer to this as *conceptual embodiment*. It is initially independent of language, though language is essential to formulate increasingly subtle levels of sophistication.

The development of language itself is difficult to date as it leaves no physical evidence. However, some forms of proto-language using vocalised sounds evolved over time, possibly in species of Homo two million years or so ago, continuing to function in various ways in many other species today.

Mathematical symbolism evolved in Homo Sapiens in the last fifty thousand years or so (which coincides approximately with the appearance of constructed artefacts), with the earliest known tally marks for counting dating back around twenty-five to thirty-five thousand years[1]. As archaeological evidence continues to be discovered, these dates may be modified to some extent. Arithmetic symbols proliferated in various communities in Egypt, Babylon, India, China around five thousand years ago. The first flowering of mathematical proof arose in Greek geometry two and a half thousand years ago.

Meanwhile arithmetic problems for unsigned numbers were described verbally in various cultures leading to al-Khwarizmi’s book on *Al-jabr* in 720 AD which solved problems related

[1] For more detail, look up ‘Tally marks’ and ‘History of mathematical notation’ in Wikipedia.

to areas of rectangles and squares expressed as linguistic equations. On each side of the equation was a verbal expression to add, subtract and square unsigned numbers and two methods of operation were introduced to give a solution: one to move subtracted numbers to the other side to become an addition, the other to perform the same operation on both sides of the equation.

When Descartes linked geometry to algebra to deal with curves in the plane in algebraic form in 1635, his quantities were still unsigned lengths. Algebraic methods that led to negative and complex solutions of equations were rejected. They continued to be regarded with suspicion even when Argand (1806) and Gauss interpreted them as points in the plane at the turn of the nineteenth century as a visual representation of solutions of polynomial equations. Eventually they became so useful in the complex analysis of Cauchy (1821) that they could no longer be resisted.

Then, at the turn of the twentieth century, new approaches using quantified set-theoretic axioms and definitions introduced a more sophisticated *axiomatic formal* approach that applies not just to 'naturally occurring' real world problems, but also to any, as yet undiscovered, contexts that satisfy the specified set-theoretic axioms and definitions. In this sense, 'axiomatic formal mathematics' is 'future-proofed' and can be used to build up new mathematical theories, even though there may be subtle implicit meanings that later require more careful consideration.

The evolution of sophistication in mathematical thinking continues today. The need for a formal proof to be given in a finite number of steps led to limitations in working with infinite systems including the counting numbers: Gödel's incompleteness theorem shows that there are quantified statements that may be true for all whole numbers yet may not have a finite proof.

New forms of logical systems have been proposed such as the hyperreal numbers and the theory of non-standard analysis. Again, these new structures involve problematic boundaries to be crossed and they evoke a sense of transgression in some pure mathematicians, who prefer to stay in the context of standard analysis where they feel comfortable because it still works for them. Meanwhile, applied mathematicians, who find it useful to think in terms of 'arbitrarily small quantities', continue to use ideas that they refer to as 'infinitesimal calculus'.

In my own personal transition from the Mathematics Department at Warwick University to the Mathematics Education Research Centre in the Science Education Department, I experienced strong differences between the two cultures. I was contracted to continue to teach one mathematics course and one mathematics education course in the Mathematics Department each year. The mathematics education involved two courses which alternated in successive years so that second- and third-year students could study both options. One was a 'problem solving' course following the approach of Mason et al (1982) and the other was a course on 'development of mathematical concepts' in which I compared long-term cognitive, historical and logical sequences of development. The undergraduates attending the mathematical concepts course, including a young David Pimm, persuaded me to run a course on Keisler's approach to infinitesimal calculus based on his book for instructors (Keisler, 1976). I translated the logical formulation into set theory and offered a new mathematics option called 'Infinitesimal Calculus'. This ran for one year but was rejected for the next by a vote of the Mathematics Department (whose expertise at the time focused on standard analysis, algebra and topology) claiming that the course was not pure mathematics. Nevertheless, it was accepted as an 'education option' and was followed by large numbers of mathematics students who had previously studied standard mathematical analysis and found the ideas supportive (Tall, 1980a).

The framework of set-theoretic definition and formal proof also provided a context to introduce infinitesimals in any proper ordered field extension of the real numbers by proving properties that have embodied and symbolic interpretations that can be pictured visually and manipulated algebraically (Tall, 2013, Stewart & Tall, 2014, 2018).

The general approach taken is widely applicable in other axiomatic formal contexts. It involves proving a *structure theorem* that an axiomatic system has properties that can be interpreted in embodied and symbolic ways, thus evolving more sophisticated levels of embodied, symbolic and formal modes of thinking.

For example, in mathematical analysis, the real numbers, as a complete ordered field, can be proved to have a visual representation as the number line and a symbolic representation as infinite decimal expansions. In the theory of vector spaces, the elements of a finite dimensional vector space over a field F can be represented using coordinates $(x_1, \ldots, x_n)$ in F^n, that could be operated upon symbolically, or imagined visually (for instance when F is the field of real or rational numbers and $n = 2$ or 3). The theory of finite groups can be interpreted as operations on a visual figure with n vertices where the operations are permutations of a set of n objects. Stewart & Tall (2014) show how these different structure theorems lead to different kinds of formal theory, including one with a single structure unique up to isomorphism (the real numbers in mathematical analysis), another involving specific generic cases (a finite dimensional vector space, F^n), and others seeking the classification of different cases (finite groups).

This gives a spiralling evolution of embodiment, symbolism and formalism where embodiment and symbolism inspire axiomatic formal theories which in turn give sophisticated forms of embodiment and symbolism, leading to more sophisticated formal structures.

2.3 Long-term development of mathematical thinking in the individual over a lifetime

How Humans Learn to Think Mathematically (Tall, 2013) focused on the long-term cognitive and affective development of the individual. This considered the development of embodiment, symbolism and formalism in topic areas including arithmetic, algebra, geometry, trigonometry, calculus, in school, and on to undergraduate and research mathematics at university.

The three long-term threads of development (that I termed ‘three worlds of mathematics’) reveal distinct roles played by *objects* (initially physical, later as mental representations), *operations* on objects, (initially involving actions on physical objects, then on mental objects) and *properties* (of objects and of operations). Each thread incorporates all three aspects in different ways. *Conceptual embodiment* focuses on objects and their properties, with operations on (physical and mental) embodied objects performed to reveal more sophisticated properties of the objects. As more sophisticated contexts are encountered, the language used at one level may change in meaning at the next. *Operational symbolism* focuses more on operations and the symbols specifying operations can be used as objects to manipulate at a higher level. *Formalism* is seen to arise at two distinct levels, one involving *theoretical* mathematics, to apply to objects that can be imagined in a thought experiment to deduce that one property implies another, as occurs in Euclidean geometry. The other is a higher *axiomatic formal* level where specific properties are selected as a basis for a theory using quantified set-theoretic definitions and all other properties are deduced by manipulating the quantified statements using formal proof.

As noted in the historical development, formal proof is not the final summit of mathematical thinking. When axiomatic formal proof is attained, the individual may prove a wider range of embodiment, symbolism and formalism using structure theorems.

The original study (Tall, 2013) looked in detail at the growth of mathematical thinking. In attempting to assess progress of individuals, not only do different individuals diverge in their ability to succeed in successive stages of the curriculum, the same individual can operate at different levels in different aspects at a given time. To gain insight into the broad development of mathematical thinking in individuals and in communities over time, figure 2 distils the broad outline of development of mathematical thinking through practical, theoretical and axiomatic formal levels of development in the three worlds of mathematics.

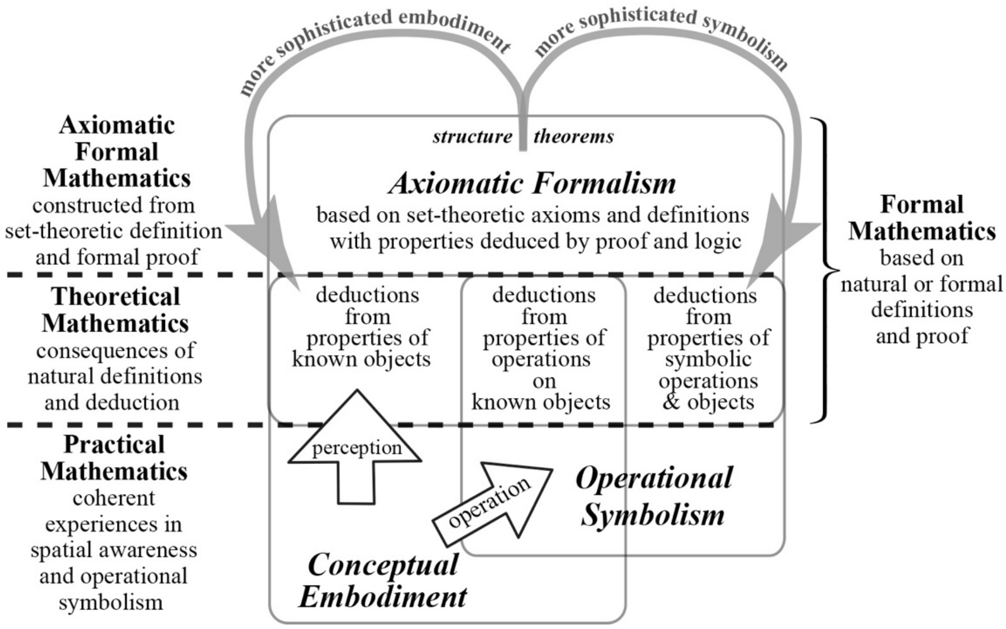


Figure 2: Long-term development of mathematical thinking (based on Tall, 2013, p. 403)

Our purpose now is to reflect on this overall picture, to compare the coherence of practical mathematics with the consequence of definition and deduction in theoretical mathematics, and with set-theoretic definition and deduction in axiomatic formal mathematics.

Different readers may see this framework in differing ways depending on their insight into various aspects. Since this book is entitled *Mathematicians' Reflections on Teaching: A Symbiosis with Mathematics Education Theories*, the plan is to focus on more advanced mathematics in college and university. Even so, students (and mathematicians) base their mathematical thinking on knowledge, beliefs and attitudes built up over years of previous experience: fundamental aspects of human thinking over a lifetime are relevant at all levels.

It is essential to realise that the framework does not represent a strict sequential development. Although embodiment begins before operational symbolism, practical mathematics begins before theoretical mathematics, which precedes axiomatic formal mathematics, as the individual's knowledge structure becomes more sophisticated, at a given time, *all aspects in the framework may be relevant in any order*.

In the embodied world, practical mathematics involves not only perception and operation with physical objects, but also imaginative thinking related to coherent *recognition* and *description* of properties of objects, such as those arising in ruler and compass constructions in Euclidean geometry. By contrast, theoretical mathematics involves *definition* and *proof* using carefully chosen definitions of naturally occurring objects and properties that follow as a consequence of a deductive argument.

In arithmetic and algebra, practical mathematics involves observed recognition of operations in arithmetic such as the fact that adding a list of numbers gives the same total regardless of the order of addition. Theoretical mathematics selects specific rules such as the commutative, associative and distributive laws of addition and subtraction in their simplest forms and deduces general properties. This involves minimal definitions but requires more sophisticated proofs. For example, general properties familiar from practical mathematics may require theoretical proof by induction in a *potentially infinite* form, with a starting statement and a succession of deductions that if the property holds at one stage, then it holds at the next, so the statement can be proved *for any specific stage* after a finite number of steps.

In axiomatic formal mathematics, induction uses the Peano postulates and takes the form of a finite proof: prove the first stage, then prove the general deduction that *if* it is true at one stage it is true at the next, then quote the induction axiom that asserts the truth of *all* stages. Dealing with the infinite reveals a crucial distinction between practical, theoretical and axiomatic formal levels of thinking.

There is a boundary between practical and theoretical mathematics in the transition from geometry and algebra to the calculus. In Euclidean geometry, the tangent to a circle is a practical construction that can be described as 'a straight line that touches the curve at precisely one point.' It can be constructed by drawing the radius from the centre of the circle to a point on the circle and the tangent is the line at right-angles to the radius. But in the calculus, the tangent to a more general curve is no longer given by this definition. This is a van-Hiele-type change in context where the language in geometry no longer applies in the context of calculus.

Meanwhile, the practical algebraic definition of the tangent at $y = f(x)$ involves calculating the limit of the expression $(f(x+h) - f(x))/h$ as h tends to zero. This is a potentially infinite process where *h* gets as small as desired but cannot be taken equal to zero. This causes a conflict because the quotient only exists if *h* is non-zero but it is calculated by putting *h* equal to zero.

A visual resolution of this difficulty can be seen by plotting points

$\left(x, \dfrac{f(x+h) - f(x)}{h}\right)$, for fixed *h* and variable *x* (Tall, 1985).

These build up points on a graph that can be called the *practical slope function*. For small values of *h* the graph stabilises on the graph that can be called the *theoretical slope function*, which is the derivative of *f*. The infinite limit *process* arises by varying *h* in the practical slope function, which, visually, stabilises on the derivative as a graphical *object*.

The difference between practical and formal mathematics arises in the capacity of the learner to think of the expression

$$\lim_{h \to 0}\left(\frac{f(x+h) - f(x)}{h}\right)$$

flexibly as an infinite process or a limit concept. Some individuals find this a powerful insight, often implicit in their use of symbolism, others find it to be an impediment causing conflict.

A further transition occurs between calculus and analysis using the quantified set-theoretic definition of limit. The definition is still a potentially infinite *process* ('given an epsilon, however small, there exists a delta such that …') but the derivative is now *defined* to be the limit *object*. The devil lies in the detail. Axiomatic formal mathematics involves multi-quantified definitions that are highly problematic for many learners.

The same framework of practical, theoretical and formal development in embodiment, symbolism and formalism applies in other areas of mathematics. For example, the arithmetic of whole numbers begins with practical mathematics including embodied aspects that allow visualisation of properties of triangular, square, rectangular numbers and non-rectangular prime numbers that have no proper factors. Theoretical mathematics includes the uniqueness of prime factorisation and the potential infinity of primes. Formal mathematics develops set-theoretic definitions, distinguishing prime and irreducible numbers in algebraic extensions of the whole number system and the vast machinery required to address Fermat's last theorem. (Stewart & Tall, 2015).

We now underpin this framework to provide evidence that can be observed by teachers and learners and trust that experts will be willing to consider evidence that comes from a different area of expertise.

3. Long-term supportive principles and resolution of problematic transitions

In this section we focus on simple observations that are visible to the teacher and learner to seek supportive principles that extend over several changes in context to provide a stable foundation to give learners confidence to deal explicitly with problematic aspects. While the focus of our study will be on more advanced mathematics studied in college and university, many problematic aspects at this level arise from experiences in school relating to how we *speak*, *hear*, *see*, *read*, *interpret* and *communicate* mathematical ideas.

The plan is to use this information to formulate a strategy to support meaningful development of mathematical thinking for a wide range of teachers and learners:

> **The principle of long-term meaningful learning:** it is essential for the teacher to be consciously aware of those ideas that remain supportive through several changes of context, to give confidence to the learner, and to make explicit those ideas that are problematic so that they can be addressed meaningfully.

I name this the 'Simon Principle' as it arose from conversations with my grandson whose ways of thinking fundamentally enhanced my personal view of the long-term development of mathematics. It is related to a small number of general principles that apply throughout the full extent of mathematics.

3.1 Conservation of counting number and general principles of arithmetic

The foundational idea of counting lies in Piaget's Principle of Conservation which gives:

> **The principle of conservation of number**: The number of elements in a collection of objects is independent of the way it is counted.

Gray & Tall (1991, 1994) studied the development of children aged 5-12 and revealed how various methods of counting (count-all, count-on, known facts, derived facts) compress different procedures for counting into the concepts of number and arithmetic, revealing the way in which a symbol such as $3 + 2$ operates dually as a process in time (addition) or as a mental concept (the sum). The dual use of such a symbol flexibly as process or concept was named a 'procept' in which many different operations give the same mental object. The notion of procept is a foundational idea throughout operational symbolism. For example, the principle of conservation of number underpins the idea that whatever operation is used to add a collection of numbers, the result is always the same. This can be formulated as:

> **The principle of conservation of addition**: The sum of any collection of numbers is independent of the order of addition.

This applies not only to whole numbers, but also to fractions, signed numbers, rational numbers, real numbers and complex numbers. For example,

$3 + \frac{1}{4} + (-2) + \sqrt{2}$ is the same as $\sqrt{2} + (-2) + \frac{1}{4} + 3$.

The Simon principle operates over several changes of context to provide the learner with a stable foundation for long-term learning. The situation will seem very different for various learners, who may be impeded by their current experience with number systems, compared to experts, who already have different experiences of more sophisticated ideas. In practice, learners will spend long periods working with whole numbers and encounter other operations such as subtraction and multiplication which present them with a range of problematic aspects. For example, learners may read the expression

$2 + 2 \times 2$

from left to right in the normal reading order to get

2 + 2 is 4 and 4 × 2 is 8

yet they are told they must respect a convention that says, 'multiplication takes precedence over addition', so they must first calculate

2 × 2, is 4, then add 2 + 4 to get the 'correct' answer 6.

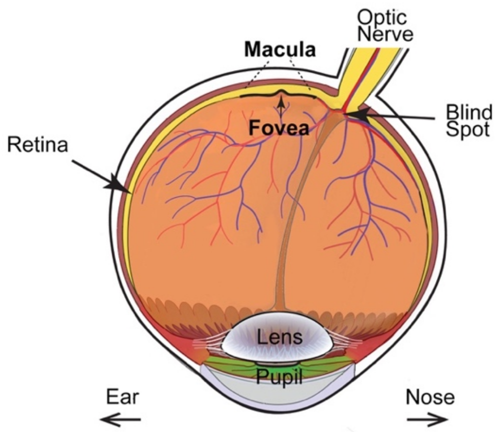
This violates their fundamental experience of reading from left to right and can often lead to rote-learning of the rule of precedence which is not understood and is likely to lead to more and more errors over the long term. The literature is full of student errors and 'misconceptions' with sophisticated theories about what is going wrong. The Simon principle not only considers why such errors occur, it also offers principles to improve learning at all levels.

3.2 How humans make sense reading and speaking text

To give meaning to mathematical expressions, it is useful to understand how humans make sense when reading and speaking. This may be done by becoming aware of what happens as we read a line of text. The reader should select any paragraph on the page and notice what is happening as you read. *Do this now*, so that you may gain a personal sense of how it works ...

You will find that your eye does not move smoothly over the text as you read, instead it moves in a sequence of jumps (called 'saccades'). *Read any paragraph again to make sure you are aware of this*.

This happens because the central area in the retina of the eye (the fovea) that focuses on detail is only around 1.5 mm in diameter and only takes in a few characters of text at a time. A larger area (the macula) around 5.5 mm in diameter surrounds the fovea and gives surrounding detail that is not as clearly in focus while the blind spot which links to the optic nerve gives none (Figure 3)[2].



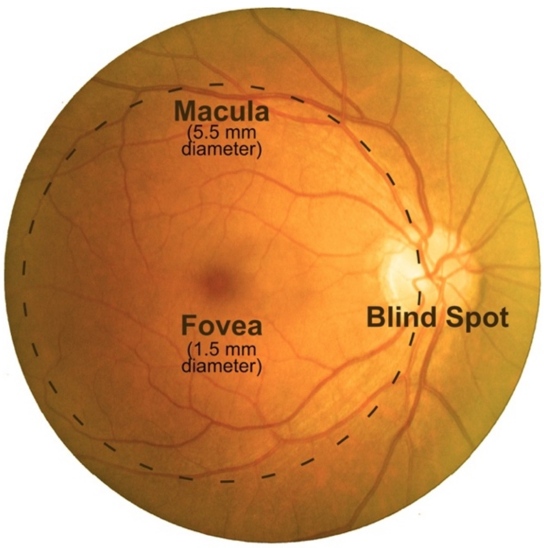


Figure 3: The central part of the retina seen through the pupil of the eye

The brain interprets each chunk of information in turn and fits them together to give sequential meaning to the text. Each chunk is interpreted as a whole, so short stretches such as the number 123 in German may be processed as 'ein hundert, drei und zwanzig', reading the digits in the order 1, 3, 2, while the chunks as a whole are read successively in a direction appropriate for the language (left to right in Western languages, right to left in Hebrew or Arabic, top to bottom

[2] See: https://en.wikipedia.org/wiki/Macula_of_retina

in columns written from right to left in Chinese). If the text is spoken out loud, then it will be heard by another person in the same spoken sequence in time.

However, when text is spoken, there are other aspects that communicate additional meaning, such as the tone of voice or the way in which the language is articulated. This enables us to give more precise meaning to mathematical expressions.

2.2 Giving meaning to expressions through spoken articulation

The manner in which we speak mathematics can radically clarify its meaning. We can give different meanings to an expression depending on how we say it. In particular, by leaving short gaps in speech we can distinguish between two different meanings of $2 + 2 \times 2$ as:

'Two plus [gap] two times two' gives $2 + 4$, which is 6

'Two plus two [gap] times two' gives 4×2, which is 8.

I joked about this with my 11-year-old grandson, who contacted me two days later using FaceTime on his iPad (Tall, Tall & Tall, 2017). He asked, 'What is the square root of nine times nine?' I knew he was aware of the properties of negative numbers and thought he was trying to test me, so I replied, 'It can be plus or minus nine.' He smiled and said, 'No, its twenty-seven!'

I had no idea what he was talking about until he explained:

'It is the square root of nine [gap] times nine.'

Over the weeks that followed, I steadily realized that this offered a simpler interpretation of the difficulties with symbolism expressed in the wider literature and also in my own publications. For example, it now gives a *meaning* to the need to introduce brackets.

The reader is invited to give different meanings to the following by speaking of them in different ways:

five minus four plus one,

two plus three times four,

minus two squared.

Does 'five minus four plus one' mean

'five minus four [gap] plus one,' which is $5 – 4$ plus 1, giving $1 + 1$, which is 2,

or is it

'five [gap] minus four plus one,' which is 5 minus $4 + 1$, giving $5 – 5$, which is 0.

The reader is invited to reflect on the meaning of the other expressions given above.
This led me to propose:

The Articulation Principle: The meaning of a sequence of operations can be expressed by the manner in which the sequence is articulated. (Tall, 2019)

This is not a definition in a mathematical sense. It is a principle to encourage us to reflect on the meaning of mathematical expressions as they develop in sophistication over the long term. More importantly, it leads to a more natural way of expressing meaning through the use of brackets (or 'parentheses' in American English). For instance,

$2 + 3$ [gap] $\times 4$ may be written as $(2 + 3) \times 4$

$2 +$ [gap] 3×4 may be written as $2 + (3 \times 4)$.

The use of brackets leads us to consider the distributive law involving the product of a number times a sum in brackets, such as the expression $3 \times (2 + 4)$ where the sum $2 + 4$ must be performed first. This may be related to the expression $3 \times 2 + 3 \times 4$ using embodied pictures, one calculating with whole numbers, the other calculating areas where lengths need not be whole numbers.

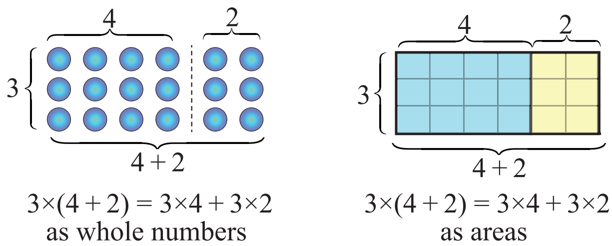


Figure 4: Embodied examples of the distributive law: counting and measuring

In the absence of brackets, it is natural to perform operations in the sequence spoken in time or read in the standard order from left to right. For instance, the expression 5 – 4 + 1 gives 2. When an expression just adds numbers together, the order of addition does not matter. If the operations are a mixture of addition and subtraction, there is a possible problematic aspect. If the terms in 5 – 4 + 1 are reordered as 1 – 4 + 5 then this works if negative numbers are allowed, but not with whole numbers, because you can't take four objects away if you only start with one.

Over the longer term, when contexts arise that include negative numbers (such as bank balances or temperatures above or below zero) then an extended general principle arises that applies not only to whole numbers but to any quantities in arithmetic or algebra:

> **The General Principle of Addition and Subtraction**: A finite sequence of additions and subtractions of quantities is independent of the order of calculation.

For an expert with sophisticated experience, this may appear as an enlightenment. For the learner encountering successive number systems including signed numbers, real numbers and complex numbers, it may (and often does) involve problematic transitions.

A similar principle holds for multiplication and division, though, in this case, division has different properties for whole numbers (in terms of quotients and remainders) than for real and complex numbers, where division by a non-zero number is always possible. In the latter case, for fractions, rational, real and complex numbers, we have:

> **The General Principle of Multiplication and Division**: A finite sequence of multiplications and divisions of quantities is independent of the order of calculation.

These two principles need to be put together using the distributive law. In practical mathematics, the expression in brackets may be a finite sequence of additions and subtractions and there may be occasions when several brackets are multiplied together. Learning arithmetic involves both specific knowledge of number bonds and general principles of relationships. Both aspects need to be made explicit so that the learner not only has technical facility but also is able to apply general principles meaningfully in a variety of situations.

As the expressions become more complicated, more complex issues arise, such as the order of precedence of operations, involving Parentheses (Brackets), Exponents (Indices), Multiplication, Division, Addition, Subtraction. This is often given by a rote-learned mnemonic, such as PEMDAS (Please Excuse My Dear Aunt Sally) in the USA or BIDMAS (Brackets, Indices, Division, Multiplication, Addition and Subtraction) in the UK, with the added complication that the order of precedence is

P > E > M = D > A = S, or equivalently, B > I > D = M > A = S.

Frankly, few people make sense of these conventions. To make sense of mathematical ideas requires more than rote-learning rules, as explained by my doctoral supervisor:

> I always want to try to understand why things work. I'm not interested in getting a formula without knowing what it means. I always try to dig behind the scenes, so if I have a formula, I understand why it's there […] to understand why it works, *you*

have to have a kind of gut reaction to the thing. You've got to *feel* it.

(Michael Atiyah, quoted in Roberts, 2016, with italics added.)

To *feel* the different strengths of bonds in mathematical expressions relates to being able to realise that sub-expressions can be seen flexibly as operations or as objects and also that an operation of higher order is bound more strongly than one of lower order. For instance, the expression $3x^2$ has the power 2 bound more strongly to the variable x than 3 is to x. It is therefore seen as '3 times the object x^2' rather than 'the square of the object $3x$.'

4. Interpreting the duality and flexibility of expression as operation or object

To aid a learner to *see* an expression built up in various ways with some parts as operations and other parts as objects, Tall (2019) proposed drawing (or imagining) an object in an expression being placed in a box. For example, the expression 2 + 3 can be expressed as:

$\boxed{2}+\boxed{3}$ as the operation of addition of two mental objects (numbers)

or

$\boxed{2+3}$ as the single mental object (the sum of the two numbers).

In the case of the expression $3x^2$ (also written as $3\times x\,\hat{}\,2$), the stronger bond for the power allows it to be written as

$\boxed{3}\times\boxed{x\,\hat{}\,2}$ or as $3\boxed{x^2}$.

It is important for the learner to have a sense of the relative strength of different operations. In an expression such as

$$3x^2+2x+1$$

it may be sensed that the expression x^2 is bound together more strongly than the other terms, and the whole expression is a sum of three objects, $3x^2$, $2x$ and 1, where $3x^2$ and $2x$ can be seen flexibly as objects or operations.

The use of boxes supports the initial interpretation of simple expressions. At a later stage, it can help to interpret more general expressions written spatially, such as

$$\frac{3x^2+2x+1}{x^2+2}$$

This can also be interpreted in the same way:

as an object $\boxed{\dfrac{3x^2+2x+1}{x^2+2}}$ or as an operation $\dfrac{\boxed{3x^2+2x+1}}{\boxed{x^2+2}}$

where the sub-expressions as objects can recursively be seen as processes. Ultimately what matters is that the learner develops the sense to interpret expressions in a flexible way, taking account of the precedence of the operations.

4.1 Making sense of equations and the equals sign

The study of equations features widely in mathematics education research. What may be less well understood is that the meaning of equations depends very much on whether the expressions are interpreted as operations (processes) or objects (concepts), or dually as procepts which can be flexibly interpreted as either.

Initially an arithmetic equation such as 4 + 3 = 7 is seen as an addition giving the answer as a number, so the left-hand side is an operation and the right-hand side is an object. In this interpretation, the equation 7 = 4 + 3 may be problematic because 7 does not 'make' 4 + 3.

Research has shown that an equation of the form 'expression = number' such as $4x - 1 = 7$ is more easily solved than an equation with expressions on both sides such as $3x + 2 = 4x - 1$ (Filloy & Rojano, 1989). The first equation can be interpreted as an operation giving a number which can be 'undone' by reversing the steps of the operation $4x - 1$ by adding 1 to the result 7 to get $4x$ is 8, so, dividing by 4 gives $x = 2$. The second equation involves difficulties of a higher order, catalogued widely in the literature that often involve rote-learnt procedures applied in inappropriate ways.

In terms of the framework offered here, if the two sides are seen to give the same numerical object, then applying the same operation to both sides will maintain the equality and a suitably chosen sequence of operations can lead to a solution. There are two distinct possibilities: either the equation is true for *all* values of x, which gives an *identity*, or it is not. The latter case leads to the study of techniques to solve equations, while an identity is the notion of 'procept' where different operations represent a single object. Visually, different algebraic expressions that represent the same function are pictured as the same graph, offering an embodied 'gut reaction' that they *feel* like a single mental object.

The equals sign can also represent the definition of a quantity. For instance, if a variable is written as being equal to an expression, such as $y = x^2$, then it expresses the variable y as a dependent variable given in terms of the independent variable x, or, in set-theoretic terms, it expresses the notion of a function.

In more advanced mathematics, a power series such as

$$\sin(x) = x - \frac{x^3}{3!} + \ldots + \frac{x^{2n-1}}{(2n-1)!} + \ldots$$

offers a practical definition of a function as a potential process of calculation to any desired accuracy. Drawn over a finite interval, the approximations stabilise on the limit graph. Again, there is a sense that the infinite process stabilises on the limit object.

In all these cases it is possible to focus either on the potentially infinite process that gives a *practical limit*, as close as is desired, or switch attention to the *theoretical limit*, which is the limit object itself.

Axiomatic formal mathematics takes the theory to a more sophisticated level by formulating the definition of the limit as an object in terms of a quantified set-theoretic definition which can be used to prove relationships using formal proof.

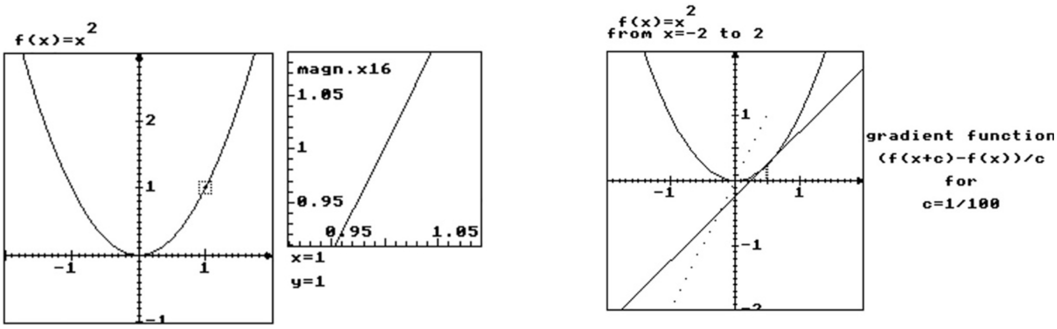
4.2 Practical and theoretical limits in the calculus

The arrival of graphical displays in the late seventies and early eighties offered new possibilities to distinguish between practical and theoretical limits by drawing practical approximations and considering what happens as they go through the process of approaching the limit object.

The *Graphic Calculus* software that I programmed in the mid-1980s offered a number of facilities. For differentiation the student is encouraged to magnify a graph to see that, under sufficiently high magnification, a differentiable function will look 'locally straight', giving (within the limitations of the primitive graphics) a straight line. Then, looking along the graph of the function $y = f(x)$ it becomes possible to imagine the changing slope and to draw the *practical slope function*

$$\frac{f(x+c) - f(x)}{c}$$

for a small, fixed values of c. The limiting process can be studied by consider what happens to the practical slope function as c gets small (Figure 5).



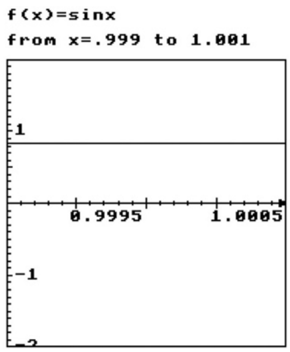
A **differentiable** function magnified to look **locally straight**

The **practical slope function**

Figure 5: magnifying a differentiable function and its practical slope

In the picture the practical slope function of $f(x)=x^2$ is $2x + c$ and, for small values of c, this stabilizes on the *theoretical slope function* $f'(x)=2x$. This technique can be used to investigate all the derivatives of standard functions, using visualisation.

The software also allows the user to stretch the graph horizontally while maintaining the vertical scale. If this is done for a *continuous* function, then the graph 'pulls flat' (Figure 6).



A **continuous** function stretched horizontally to look **locally flat**

Figure 6: A continuous graph pulled flat

This can be translated into the definition of continuity. If the pixel height represents a scaled value of $\pm\varepsilon$ for a real number $\varepsilon > 0$, then the graph of $y = f(x)$ 'pulls flat' around a point x_0, means a real number $\delta > 0$ can be found such that for points within δ of x_0, the points $(x, f(x))$ lie in the horizontal line of pixels. If this is true for any positive $\varepsilon > 0$, then this is the theoretical $\varepsilon - \delta$ definition of continuity.

As a pure mathematician intent on giving meaning to learners, I used the embodied ideas of local straightness and pulling flat to build a practical theory of differentiability and continuity that includes an embodied version of the fundamental theorem. This was designed to be shared by beginners yet has the capacity to form the initial introduction to a wide variety of approaches to calculus and analysis appropriate for different communities. However, in reality, technology was evolving so fast that curriculum change could not keep up with the technological advances.

As we will see in §6, the latest developments in enactive retina displays offer new tools to make sense of sophisticated embodied ideas. The direction taken in the future will depend on whether various communities get to know of these ideas and whether they consider them to be an enlightenment or an impediment.

4. Making sense of constants and variables

In our discussion so far, we have often spoken about quantities that are 'variable'. This builds on how our eyes and brain interpret the dynamic movement of objects. Again, this arises as our eyes jump in saccades to focus on specific detail, as can be seen by holding a finger at a distance from the eye and moving it to the left and right. In this case there is an initial saccade to fix on the finger and then the eye follows the finger, keeping it in focus. The notion of 'variable' is therefore a natural feature of our human perception and has been with us for centuries. It underpins our fundamental mathematical thinking of variable quantities.

On a number line we can imagine two different kinds of point. One remaining in a fixed place can be considered a *constant* and another, which moves around, can be called a *variable*. This supports the interpretation of graphs and calculus in terms of constants and variables. Such an idea can be imagined in our mind's eye, but with a dynamic picture on a retinal display, we can now offer a visual representation of a variable approaching a fixed point (Figure 5).

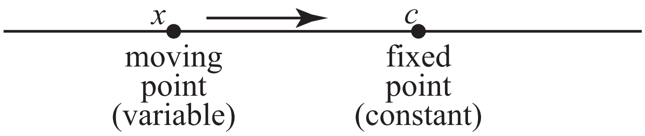


Figure 5: Embodied constant and variable

Although the modern approach to analysis speaks in terms of sets and relations in which points are represented as being fixed elements on a number line, thinking of quantities as constants or variables is deeply embedded in our human thought processes and is a more 'natural' way of thinking about the variation of quantities in the calculus.

5.1 Variables as infinitesimals

A variable that tends to zero can be visualised dynamically as an infinitesimal. In figure 6, points are marked on the vertical line $x = t$, where it is crossed by the constant line $y = k$ (for $k > 0$), the straight line, $y - x$, and the parabola, $y = x^2$. The points are marked on the line $x = t$ as k, t, t^2 respectively.

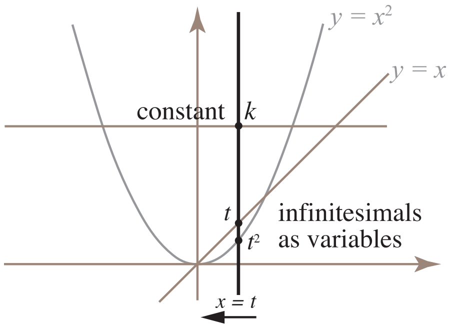


Figure 6: variables as infinitesimals (Stewart & Tall, 2013 p. 347)

The order of the points on the vertical line is determined by their relative height as t moves down to zero. In this case, once t is less than k, the height of the points is $0 < t^2 < t < k$. This gives a linear order on the field $\mathbb{R}(t)$ of rational expressions in an indeterminate t in which $0 < t^2 < t < k$ for any positive real number k and so t and t^2 are *infinitesimal*. Furthermore, t^2 tends to zero faster than t, so it is, in this sense, an infinitesimal of higher order. Given two non-zero elements u, v, we say that u is 'a higher order infinitesimal than v' if u/v is infinitesimal and 'of the same order' if u/v is finite. In this sense, t^2 is of higher order than t. Their inverses $1/t$ and $1/t^2$ are larger than any real number and so are *infinite*, with t^2 a higher order of infinity than t.

This new way of conceptualising infinitesimals may be problematic for some students and even for some mathematicians. However, the idea of an infinitesimal as a variable tending to zero arose at various times in history, including the approach to real and complex analysis introduced by Cauchy (1821). (See Tall & Katz, 2014.)

5.2 Infinitesimals as fixed points on a number line

Conceptualising infinitesimals be problematic for some. However, for an individual familiar with formal axiomatic mathematics, it is possible to use the completeness of $\mathbb{R}$ to prove the following:

> **Structure Theorem for ordered extensions of the real numbers:**
> If K is an ordered extension field of $\mathbb{R}$, then every finite element x in K (meaning $a < x < b$ for some real numbers a, b) is uniquely of the form $x = c + \varepsilon$ where $c \in \mathbb{R}$ and ε is infinitesimal or zero.

The proof is elementary. Let L be the set of real numbers less than x, then L is a non-empty subset of real numbers (because $a \in K$) and is bounded above by b, so it has a unique least upper bound $c \in \mathbb{R}$, and it is straightforward to show that $\varepsilon = x - c$ is zero or infinitesimal.

If K is any ordered extension field of the real numbers, we may call the elements of K, 'quantities', and elements of $\mathbb{R}$ may be called 'constants'. The structure theorem now says:

> ***Every finite quantity is either a constant or a constant plus an infinitesimal.***

For a finite quantity x, the unique constant c is called the *standard part of x* and written as $c = \mathrm{st}(x)$.

The map $m: K \to K$ given by $m(x) = (x - c)/\varepsilon$ for any $c, \varepsilon \in K$, $(c \neq 0)$ is called *the ε-lens pointed at c*. This can be defined for any $c, \varepsilon \in K$ for $\varepsilon \neq 0$, but it is of particular interest when ε is a positive infinitesimal: in this case, m is said to be the *ε-microscope pointed at c*.

The subset V given by

$$V = \{x \in K \mid (x - c)/\varepsilon \text{ is finite}\}$$

is called the *field of view* of m and

$$\mu : V \to \mathbb{R} \text{ given by } \mu(x) = \mathrm{st}\left(\frac{x - c}{\varepsilon}\right)$$

maps the field of view to the real numbers. This is called the *optical ε-lens pointed at c*. If ε is a positive infinitesimal, then μ is said to be the *optical ε-microscope pointed at c* (Tall, 1980b).

This allows us to *see* infinitesimals as fixed points on an extended number line.

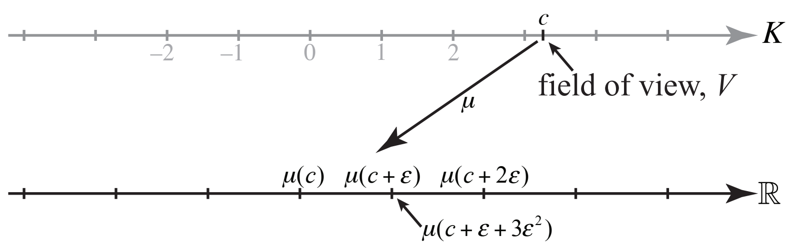


Figure 7: Optical microscope mapping infinitesimal detail onto a real line

For any $\lambda \in \mathbb{R}$, the image $\mu(c + \lambda\varepsilon) = \lambda$, so μ maps onto the whole of $\mathbb{R}$. If two points in V differ by a value that is infinitesimal and of the same order as ε, then they will be distinguished in the real image, but if they differ by a higher order of infinitesimal, then they will be mapped to the same point.

The same representations generalise to two or more dimensions. All that is necessary is to operate with an optical lens on each axis. It is possible to use the same scale on the axes to see local straightness, or different scales to see local flatness.

With these formal possibilities in mind, we now return to reflect on how we humans interpret visual information using enactive retinal quality graphics. We will find that it enables us to bridge the transition between practical mathematics and theoretical mathematics in calculus and, if required, to transition further to axiomatic formal analysis.

5. Making sense of the calculus using new technological tools

The early software available for drawing pictures on a computer or on a graphic calculator had relatively large pixels, so that the picture of a straight line looks like a line of visible blocks, as shown earlier in figure 5. The increasing pixel density on a modern retina display allows the human eye to imagine zooming in to ‘see’ local straightness (Figure 8).

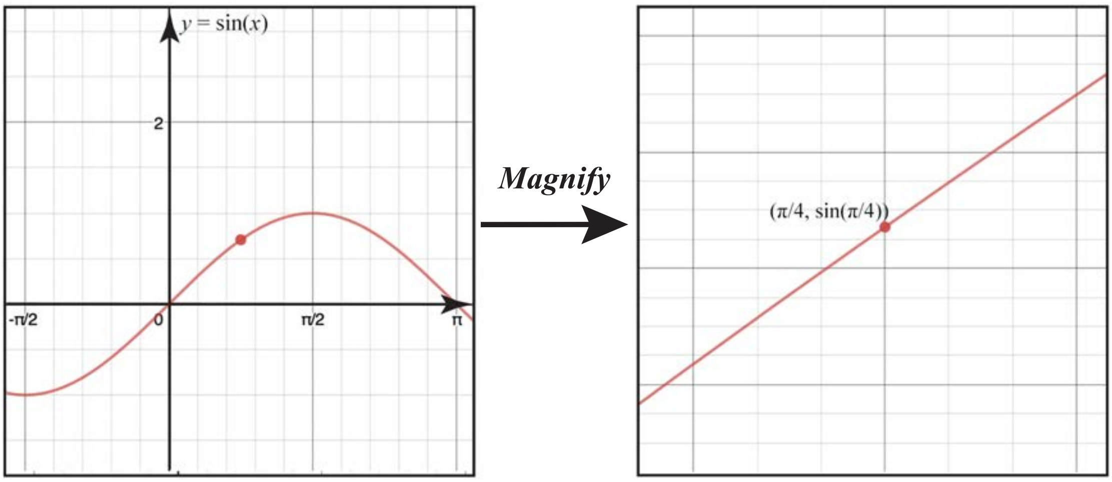
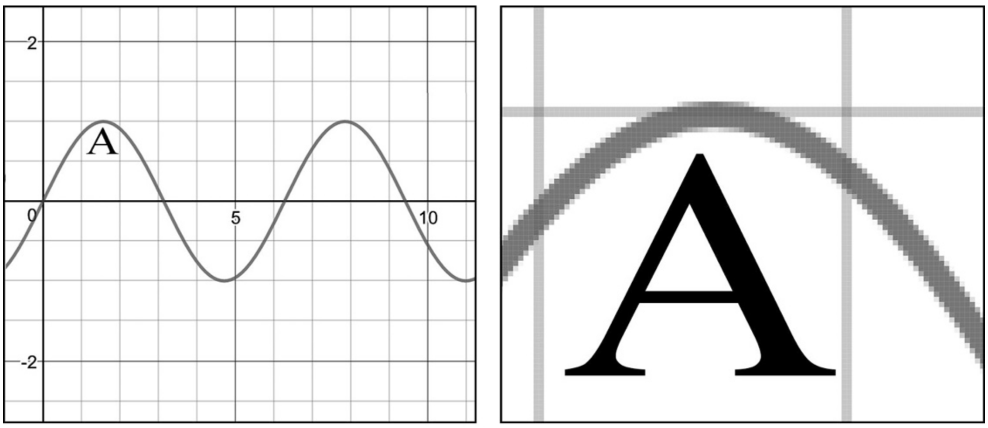


Figure 8: Magnifying a graph on a retina screen

This is because the magnified graph is redrawn to the same thickness, maintaining the illusion that the magnification is zooming in on the graph itself, rather than magnifying a picture of a static graph on a page in a book where magnification would thicken the graph by the same magnification factor. Figure 9 shows a graph drawn on a smart phone using the software *Desmos* (2011): beside it is part of this printed picture reprinted at a higher resolution. This reveals that the graph is drawn with pixels at the edges coloured in lighter shades so that it appears smooth to the human eye when viewed from an appropriate distance.



Graph drawn with *Desmos* software — Part of picture enlarged

Figure 9: Graph drawn in *Desmos* on an iPhone, with part of the picture enlarged

The illusion of smoothness is related to the structure of the eye as illustrated in figure 3. The part of the eye that detects the highest resolution detail lies in the fovea, which is a circular disk containing approximately 200,000 cones (Kolb, 2007). From this, the diameter of the fovea can be calculated as approximately 250 cones across. This places a severe restriction on the amount of detail that the eye can perceive. Holding a smart phone at a comfortable distance sees a graph as in figure 8, but presented on a large high-resolution screen, close up, it may be possible to see individual pixels (as in figure 9), or even tiny arrays of red, green and blue dots that at a distance are seen as a combined colour.

The result is that, when seen highly magnified, a differentiable function may be seen as being locally straight and the observer may look along the graph to see its changing slope to give a visual meaning to the derivative.

5.1 Embodying integration and the fundamental theorem of calculus

Visually, the area A between a graph $y = f(x)$ and the horizontal x-axis from $x = a$ to $x = b$ can be *seen* (Figure 10). On the left the area A is being approximated by the sum of rectangular strips and on the right, it is being extended by an extra strip width dx.

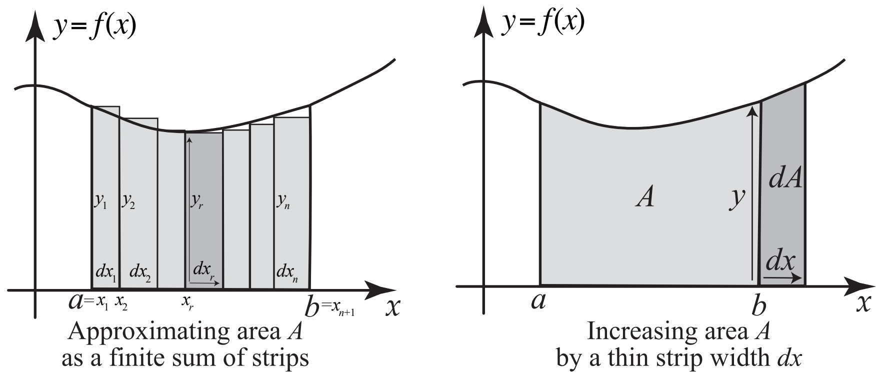


Figure 10: The area A between a graph and the x-axis

In calculus courses, the Riemann integral is usually introduced by using a partition

$$a = x_1 < x_2 < \ldots < x_r < \ldots < x_{n+1} = b$$

and calculating the approximate area as

$$\sum_{r=1}^{n} y_r dx_r \text{, where } dx_r = x_{r+1} - x_r .$$

The important information here is not the number of strips, but the endpoints a, b and the maximum width m of the strips (called the *mesh* of the partition). Denoting the width of a strip by the symbol dx and interpreting y, dx as variable quantities defined by the partition, we can denote the Riemann approximation by the notation

$$\sum_{a}^{b} y\, dx .$$

For the expert in analysis, figure 10 requires two formal proofs for a continuous function:

(1) The limit A of $\Sigma_a^b\, y dx$ exists as the mesh tends to zero

(2) $dA/dx = y$.

For the learner encountering the calculus at a practical or theoretical level, it is not necessary to prove the existence of the area A as it can be *seen* as a physical object: it is only necessary is to *calculate* it as accurately as required using the Riemann sum, and then to prove (2).

In a modern presentation, the value of A is often written as

$$\int_a^b y\,dx$$

and spoken as ‘the integral of y with respect to x’.
It is also useful to use other letters to allow other quantities to vary, for instance, the expression

$$A(x) = \int_a^x y\,dt \text{ where } y = f(t)$$

allows the area $A(x)$ from a to x to vary for $a \le x \le b$.

The area $A(x)$ for a continuous function $y = f(t)$ can be calculated by showing that the finite sum $\Sigma_a^x\, y\,dt$ can be made as close as desired to $A(x)$ using a partition with a suitably small mesh m. This is usually introduced in a first course on calculus as an informal version of the formal definition. It involves an implicit potentially infinite process as the mesh becomes arbitrarily small but not zero.

The meaning of $\int_a^x y\,dt$ has changed from the original idea of Leibniz, who used the ‘elongated s’ ∫ to signify an infinite sum of infinitesimal quantities. The idea of visualising an arbitrarily thin strip on a retinal display can be performed by stretching the graph of the function horizontally while maintaining the same vertical scale. Figure 11 shows a possible picture of this idea where the graph of $y = f(x)$ in the left box with a thin strip width dx is stretched horizontally to fill the right box. The *numerical* value of the area calculated by multiplying height y by width dx has a practical value $dA = y\,dx$ and its error is contained in the horizontal line of pixels, which can be made as small as required. This allows the learner to visualise a practical interpretation that can be extended theoretically to give the formal fundamental theorem of calculus: $y = dA/dx$ which may also be written as $A'(x) = f(x)$.

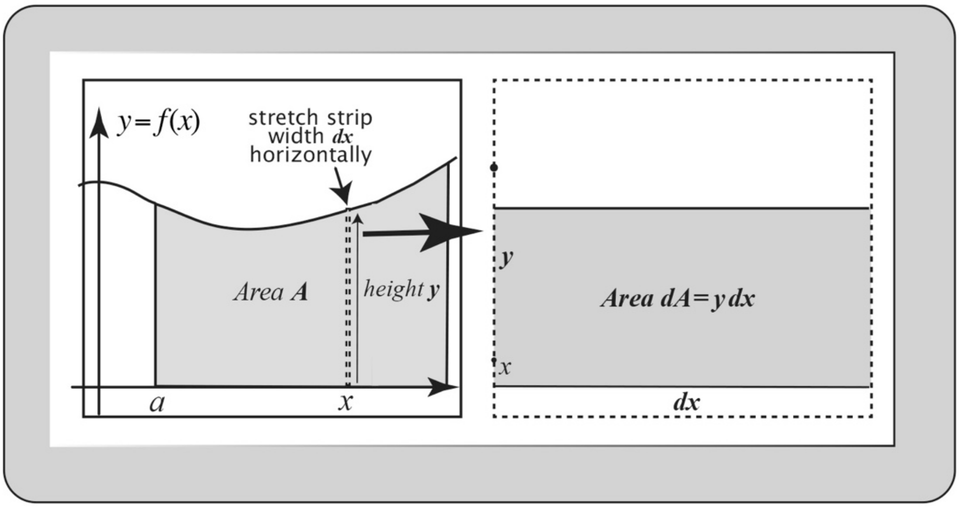


Figure 11: Visual representation of the fundamental theorem

6. Is the calculus about quantities with dimensions or is it about numbers?

The calculus is presented in at least two distinct forms which may operate in different ways with a problematic transition between them. One, which I will call ‘real calculus’, involves quantities that are real numbers, such as $y = ax^2$, another, which occurs in applications, I will call ‘dimensional calculus’ involving quantities with dimensions, such as expressing a distance s as a function of time t, say $s = ut$. In the first case, x and y are real variables and a is a real constant where the dependent variable y depends on the independent variable x. In the second case, t represents time, say in seconds, s is the dependent variable, representing distance, say in metres, u is a velocity in metres per second.

These two forms of calculus have distinct properties. In real calculus, successive derivatives are again real functions. In dimensional calculus, if distance is a function of time, its first derivative is a velocity, its second derivative is an acceleration, but what is the third derivative? Some say the sudden change of an acceleration is a 'jerk'. For $s = \sin(t)$, the third derivative is $-\cos(t)$ and this certainly does not *feel* like a jerk. And what is the meaning of the fourth derivative?

A solution of this impasse is to deal mainly with real calculus and to use dimensional calculus in specific applications with a clear contextual meaning, such as the relationship between time, distance, velocity, acceleration and force in Newtonian dynamics. Using real calculus is particularly valuable in dealing with successive differentiation and integration where the derivatives and integrals of real functions are again real functions.

Even so, *there is a fundamental impediment in elementary calculus that affects almost all of us*. When real calculus is represented visually as graphs, the variables are drawn as *lengths*. This is not much of a problem in differentiation as the derivative is a length divided by a length, which is a dimensionless number, and this has a simple interpretation in real calculus using real numbers. But it is a subtle impediment in integration, where the integral $A(x) = \int_a^x f(t)\,dt$ is represented as an *area*. If the scales on the axes are changed, then the visual area changes while the numerical area remains the same. This is particularly problematic when different scales are used as in the transformation of the picture to stretch the horizontal scale while maintaining the vertical scale, as in the case of 'pulling the graph of a continuous function flat'. It is therefore little wonder that the notion of 'pulling flat' has not been integrated into contemporary courses.

'Local straightness' on the other hand has been incorporated in some courses, often in the symbolic form of 'local linearity', as the numerically 'best linear approximation' to a curve, as in Harvard Calculus (Hughes-Hallet, Gleason, 1994). Local linearity involves sophisticated theoretical symbolism. Local straightness offers a more fundamental embodied meaning to differentiation, not only as an intuitive beginning, but also in more sophisticated situations throughout the long-term development of practical, theoretical and formal mathematics.

For instance, it is possible to give a simple example of a function that is everywhere continuous but nowhere differentiable and so its integral is differentiable once everywhere (with a continuous derivative) but not differentiable twice anywhere. Integrating a function n times gives a function that is continuously differentiable n times but not n+1 times. This offers an embodied meaning relating continuity and differentiability at a more sophisticated level than is usually encountered in formal analysis.

6.1 Practical continuum, theoretical closeness, and formal completeness

There are changes in meaning between practical, theoretical and formal versions of mathematics relating to the difference between what we perceive with our practical human senses, what we imagine theoretically in our mind, and how we move on to a set-theoretic formal approach. A *practical* drawing arises from the dynamic movement of a pen or pencil and the final static curve is a *continuum* that, as our eye moves along it, we see it as dynamically continuous. Yet we know from our experience with factorization of whole numbers that √2 is distinct from any rational number. In *theoretical* mathematics, real numbers can be conceived as separate infinite decimals. Later, *axiomatic formal mathematics* uses the *completeness* axiom to formulate a set-theoretic interpretation of the real line.

Practical, theoretical and axiomatic formal mathematics have their own characteristics. Practical arithmetic involves making numerical approximations that are 'good-enough' for a particular practical purpose. For example, a good-enough approximation for π might be 3.142 to four significant figures or 3.14159 to 5 decimal places. Solving the simple harmonic motion of a pendulum through an angle θ radians might be performed for small values of θ by replacing $\sin(\theta)$ by θ.

In calculus, the *practical slope function* of $y = f(x)$ may be calculated for variable x and a small constant value c by calculating $(f(x+c) - f(x))/c$ and (for differentiable functions) the limit object is the *theoretical slope function* $f'(x) = dy/dx$. The *practical integral function* for a variable x on an interval $[a,b]$ for $y = f(x)$ may be calculated as the finite sum $\Sigma_a^x f(t)\,dt$ for a partition of $[a,b]$ and as the mesh size gets suitably small, the limit object is the *theoretical integral function* $I(x) = \int_0^x f(t)\,dt$ and the fundamental theorem says $I'(x) = f(x)$.

The visual link between the practical approximation and the limit object in calculus can be represented using a retinal interface as in figures 8 and 11, which builds on the natural working of the human eye and brain. This allows the integral to be imagined as a sum of arbitrarily thin strips, neglecting higher order detail in a manner favoured by many engineers modelling physical problems.

7. Making sense reading a mathematical proof

An essential aspect of formal mathematics – both in a theoretical form relating to natural contexts in applications and in axiomatic formal pure mathematics – lies in giving meaning to mathematical proof. Many learners see this as a need to learn the proof by heart to reproduce it in examinations. However, it is possible to give a meaningful long-term interpretation of mathematical proof related to the manner in which the human eye and brain interpret text, symbolism and visual illustration. A written proof is laid out on the page so that the eye can look at various parts of it at will, seeing its structure as a whole, looking back at earlier details, focusing carefully on subtle key points.

Using eye-tracking techniques, Inglis & Alcock (2012) confirmed that undergraduates devoted more of their attention to parts of proofs involving algebraic manipulation and less to logical statements than expert mathematicians. Hodds et al. (2014) developed materials to encourage 'self-explanation' by reading a proof line by line, to identify the main ideas, get into the habit of explaining to themselves why the definitions are phrased as they are and how each line of a proof follows from previous lines. They were counselled not to simply paraphrase the lines of the proof by saying the same thing in different words, but to focus on making connections to grasp the main argument and explain how the given assumptions and definitions in previous lines led to the current line and contribute to the following lines. Students who had worked through these materials before reading a proof scored 30% higher than a control group on making sense of a proof several weeks later.

Language plays an essential role in formulating increasingly sophisticated ideas through naming concepts so that they can be called to mind as mental objects, describing and defining their properties and relationships. However, as contexts change, language that makes sense in one context may fail to make sense in another (such as the definition of a tangent in Euclidean geometry and in the calculus). This can act as an impediment to transition for the learner while the expert with a more experienced knowledge structure may see the change as an enlightenment.

Experts from different communities of practice have differing views on the role played by language in mathematics. Einstein, as a theoretical physicist, used thought experiments to produce his theory of relativity and claimed that 'words and language, whether written or spoken, do not seem to play any part in my thought processes' (quoted from Hadamard, 1945, p. 142). The linguist Lakoff sees language and the use of metaphor as the foundation of how mathematics grows in human thinking (Lakoff & Johnson, 1990; Lakoff & Núñez, 2000). From a neurophysiological viewpoint, Amalric & Dehaene (2016, 2019) using fMRI scanning showed that specialists in algebra, analysis, topology and geometry used non-linguistic areas of the brain to respond to the truth or falsehood of mathematical statements and concluded that 'Overall, these results support the existence of a distinct, non-linguistic cortical network for mathematical knowledge in the human brain.' Kahnemann (2011), a psychologist and economist, who was awarded the Nobel Prize in Economics for his book *Thinking fast, thinking slow*, formulated a distinction between immediate responses in a few seconds (as in Amalric and Dehaene's use of fMRI) and the long-term deep thinking required for formal proof.

In the centre of the brain is a diverse network of structures called the limbic system[3], which is involved in a range of aspects, including laying down and retrieving long-term memories, and emotional reactions to incoming sensory information (Figure 12). These have a profound effect on the nature of mathematical thinking.

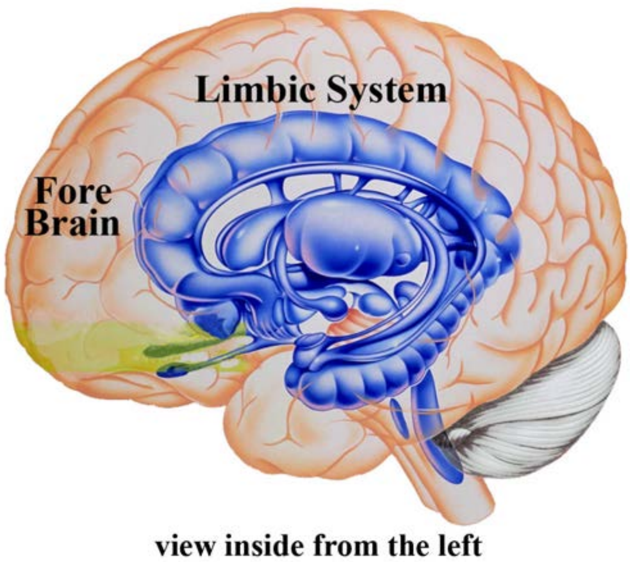


Figure 12: The limbic system

Incoming sensory data passes quickly to the limbic system, which can take immediate 'fight or flight' decisions before the forebrain is able to think through conscious decisions. It also suffuses the whole of the brain with chemical neurotransmitters that enhance or suppress connections. Positive aspects enhance motivation and set the brain on alert to solve problems while negative aspects depress thinking processes along with anxiety and fear. In the latter case, it is not just that the individual is unwilling to tackle a problem, the lack of connections makes it very difficult to think about the problem at all.

Creating an original mathematical proof is the final stage of presenting ideas in a coherent formal framework and is preceded by a varied range of activities, thinking about possibilities, formulating hypotheses which may be appropriate or erroneous, making and reformulating definitions and deductions (Byers, 2011).

The differing interpretations available confirm the need for what I will term *a meaningful long-term framework for mathematical thinking* that encompasses the many different aspects that are essential in a complex society. This chapter supports the case for a long-term development incorporating the three-world framework of embodiment, symbolism and formalism through practical, theoretical and axiomatic formal mathematics as in figure 2. It also includes other social and personal aspects related to the nature of mathematics and how it is conceived and interpreted by us as human beings.

[3] See: https://en.wikipedia.org/wiki/Limbic_system

8. Discussion

The meaningful long-term framework formulated here is a response to the various approaches that have been followed in the last half century. Within a single lifetime, the changes in technological tools have radically changed how we think about mathematics. In the United States, the development of nuclear weapons in the 1940s and the Russian launch of Sputnik in 1957 led to massive federal funding for huge 'New Math' projects to enhance American international influence (Woodward, 2004, Dossey et al. 2016). Meanwhile, in colleges and universities, computers have changed the whole approach to mathematics, particularly in calculus (Bressoud, Mesa, Rasmussen, 2015).

The New Math movement included the introduction of set-theory to explain mathematical concepts in more precise language and other approaches, such as 'guided discovery learning' where learners are encouraged to develop their own ideas through practical problem solving. These were promoted by a range of organisations, but also had a variety of opponents, from mathematicians who decried the lowering of technical facility (e.g. Kline, 1973: *Why Johnny can't add*) to teachers and parents who could not make sense of the new ideas. In 1972, federal funding was withdrawn. Throughout the US, different communities followed their own agendas with each state setting its own educational policies, often devolving decisions to local school boards.

Meanwhile, in other countries in the world, communities were seeking their own solutions. Over the following decades many different projects arose, some initially following a 'new math' approach, others, such as 'realistic mathematics' in the Netherlands, encouraged young children to construct their own mathematical ideas in real problem-solving situations. Some projects had a measure of success, but there were also problematic aspects. For example, in the Netherlands, it was found that students going to university lacked the fundamental skills previously required in university mathematics. Difficulties arose in dealing with whole number arithmetic, fractions and algebra (Gravemeijer, et al., 2016), which could be explained in terms of problematic transitions between practical problem solving and operational symbolism.

The array of possibilities is further complicated by the amazing evolution of technology from simple numeric calculators in the early seventies through an annual cycle of new developments that radically affect not only the ways that we think about mathematics, but also how we live our whole lives.

It is impossible for any individual to grasp the full picture. An expert may have an insightful view of a wide range of aspects but, as we saw in §7, experts from differing communities may have radically different interpretations. Instead of attempting to deal with the complexity of competing approaches, I chose to seek ways that could be used by anyone to improve the possibilities for learners to make meaningful sense of mathematics relevant to their own lives in a complex society.

This involved focusing on the long-term development of sophistication in mathematics and the underlying changes in meaning that can give insight to some and impediments to others. I wanted to encourage learners to develop a personal sense of confidence by realising that it is not their inadequacy that causes them to have difficulty in understanding more sophisticated ideas, it also relates to the changing nature of mathematics itself. I sought to help learners and teachers by finding general principles that remained consistent over several changes in context which could act as a secure basis to allow them to reflect on problematic aspects that needed to be resolved to take advantage of more sophisticated ideas. This led to the formulation of the Simon Principle to encourage teachers to be aware of supportive aspects that enable learners to build confidence over successive changes in context and to explicitly address problematic aspects to build a meaningful long-term development.

As an example, the principle of articulation can be used at every level to give meaning to mathematical expressions and the use of brackets. This is appropriate for young children in their early learning of mathematics, for older students who have difficulties with algebra and for teachers and experts to have a strategy to help learners develop meaningful mathematical thinking. It can be coupled with the conservation of number and the general principles of arithmetic to deal with the meaningful manipulation of expressions in arithmetic and algebra which may cause increasing difficulty over time if learned only by rote.

The beauty of the principle of articulation is that it extends the notion of flexibility of expressions representing either process or concept, which Eddie Gray and I formulated as the notion of 'procept'. This works in the interpretation of mathematical expressions throughout the whole of mathematics. Although mathematicians may not realise it explicitly, this flexibility is implicit in being able to look at an expression and 'know' how to break it down mentally into appropriate chunks to interpret its meaning.

Mathematicians generally use the notion of 'equivalence relation' to explain how different processes represent the same mathematical object. For instance, equivalent fractions represent a single real number and equivalent expressions represent the same function. In human terms this is made meaningful by the embodied representation of equivalent fractions as a single point and equivalent algebraic expressions having the same visual graph.

Embodiment also gives meaning to the natural idea of 'constants' and 'variables' as fixed and moving points. The process of 'tending to a limit' is then embodied in a practical sense by a variable point becoming visibly indistinguishable from a fixed point or a sequence of graphs becoming indistinguishable from the graph of a limit function.

Now we have retinal display on computers and smart phones, we can imagine zooming in on a graph to see that a differentiable function is 'locally straight'. It is then possible to look along the graph to see the changing slope and imagine the derivative as the changing slope of the graph itself, both visually and also translated into symbolism.

Meanwhile, in the USA, calculus has become a high school subject, still firmly based on a traditional sequence of four 'big ideas' (College Board, 2016):

- Limits and continuity,
- Derivatives,
- Integrals and the Fundamental Theorem of Calculus,
- Series.

These are all formulated in traditional language with no mention of the embodied meaning of 'local straightness', even though this can be *seen* and *sensed* by any learner using an interactive dynamic retina display. The language describes the derivative as 'the instantaneous rate of change' of the function, intimating how it is calculated as a limit, when the locally straight approach encourages the learner to look along the graph to *see* its changing slope to give it an embodied meaning.

The calculus reform project at Harvard (Hughes-Hallett, Gleason, et al., 1994) proposed that mathematics should be considered in three ways – graphical, numerical and algebraic – later extending to a 'rule of four' by adding 'verbal'. However, this begins with the more sophisticated symbolic idea of 'best linear approximation' rather than the natural idea of 'local straightness', where the derivative is simply the slope of the graph itself. Instead of building up from the learner's experience to the expert's insight, the curriculum takes the expert's formal knowledge and reformulates it in a manner claimed to be 'informal'.

When university mathematicians reflect on their teaching, they may develop a range of different mathematical approaches in their course design. I did it myself when I first taught analysis and was given freedom to interpret the syllabus. I designed an approach from the physical drawing of graphs to the formal definitions of analysis but avoided the compactness proof of the link between pointwise and uniform continuity by only defining uniform continuity. This had some success but technical difficulties arose, such as the fact that the simple function $f(x) = x^2$ is not uniformly continuous on the domain $\mathbb{R}$. My colleague Alan Weir decided to teach integration before differentiation but found that, whereas the derivative of x^n was straightforward to calculate for specific n, the Riemann sum was difficult for $n = 2$ and almost impossible for $n = 3$, so it was not possible to begin with simple computational examples. The book by Moss and Roberts (1968) bases its whole development on the notion of continuity of a function $f : D \to \mathbb{R}$ at a cluster point of D. For instance, it defines a sequence (s_n) to have the limit s if the function $f(x) = s_n$ for $x = 1/n$, $f(0) = s$ is continuous on the domain $D = \{x \mid x = 0 \text{ or } x = 1/n \text{ for } n \in \mathbb{N}\}$.

Many other approaches are possible. This may involve reorganising the mathematical content in pure mathematics by taking a different starting point, but seeking to simplify ideas in one aspect may lead to a difficulty arising somewhere else. It may also involve seeking to focus on appropriate aspects in a different subject area involving specific applications.

Different communities and individuals will take their own decisions as to how they teach mathematics and that is their privilege. The main issue here is how each learner can be encouraged to develop their own mathematical thinking in a way that is appropriate for them as individuals and also as part of a wider complex society.

This chapter offers a long-term framework for the meaningful development of mathematical thinking that takes into account the increasing sophistication of mathematical ideas and the cognitive and emotional growth of the individual. It also offers a contextual overview to encourage the comparison and cooperation of different communities of practice. It does not predict the future. It offers a framework for readers to challenge their own beliefs to make informed choices.

References

Amalric, M., Dehaene, S., (2016). Origins of the brain networks for advanced mathematics in expert mathematicians. *Proc. Natl. Acad. Sci. Unit. States Am*. 201603205. https://doi.org/10.1073/pnas.1603205113.

Amalric, M., Dehaene, S., (2019). A distinct cortical network for mathematical knowledge in the human brain. *NeuroImage* 189 (2019), 19–31.

Argand, J. R. (1806). *Essai sur une manière de représenter les quantités imaginaires dans les constructions géométriques*. Paris: Duminil-Lesueur. (2nd edition: Gauthier Villars, Paris, 1874.)

Bressoud, D., Mesa, V., Rasmussen C. (Eds). (2015). *Insights and Recommendations from the MAA National Study of College Calculus*. https://www.maa.org/sites/default/files/pdf/cspcc/InsightsandRecommendations.pdf

Byers, W. (2007). *How Mathematicians Think*. Princeton: Princeton University Press.

Cauchy, A. L. (1821). *Cours d'Analyse de l'École Royale Polytechnique*. Paris: Imprimérie Royale. See http://gallica.bnf.fr/ark:/12148/bpt6k90195m/f12

College Board (2016). *AP Calculus AB and AP Calculus BC Including the Curriculum Framework*. New York: College Board. https://apcentral.collegeboard.org/pdf/ap-calculus-ab-and-bc-course-and-exam-description.pdf

Desmos (2011). *Desmos Graphing Calculator*. https://www.desmos.com

Dossey, J. A., McCrone, S. S., Halvorsern, K. T. (2016). *Mathematics Education in the United States 2016*. ICME-13. NCTM. https://www.nctm.org/uploadedFiles/About/MathEdInUS2016.pdf

Filloy, E. & Rojano, T. (1989). Solving equations, the transition from arithmetic to algebra, *For the Learning of Mathematics*, 9(2), 19–25.

Gravemeijer, K., Bruin-Muurling G., Kraemer J, M. & van Stiphout, I. (2016). Shortcomings of Mathematics Education Reform in The Netherlands: A Paradigm Case? *Mathematical Thinking and Learning*, 18:1, 25-44, DOI: 10.1080/10986065.2016.1107821

Gray, E. M. & Tall, D. O. (1991). Duality, Ambiguity & Flexibility in Successful Mathematical Thinking, *Proceedings of the 15th Conference for the International Group for the Psychology of Mathematics Education*, 2, 72–79, Assisi, Italy.

Gray, E. M. & Tall, D. O. (1994). Duality, Ambiguity and Flexibility: A Proceptual View of Simple Arithmetic, *Journal for Research in Mathematics Education*, 26 2, 115-141.

Hodds, M., Alcock, L., & Inglis, M. (2014). Self-explanation training improves proof comprehension. *Journal for Research in Mathematics Education, 45*, 62-101.

Hughes-Hallett, D., Gleason, A. M. (1994*). Calculus.* New York: John Wiley and Sons.

Inglis, M., & Alcock, L. (2012). Expert and novice approaches to reading mathematical proofs. *Journal for Research in Mathematics Education, 43*, 358-390.

Hadamard, J. (1945). *The Psychology of Invention in the Mathematical Field,* Princeton University Press, (Dover edition, New York 1954).

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Macmillan.

Keisler, H. J. (1976). *Foundations of Infinitesimal Calculus*, Boston: Prindle, Weber & Schmidt.

Kline, M. (1973). *Why Johnny can't add: the Failure of the New Math*. New York: St. Martin's Press.

Kolb, H. (2007). Webvision: The Organization of the Retina and Visual System. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK11556/, 26 January 2019.

Lakoff , G. & Johnson, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.

Lakoff, G. & Núñez, R. (2000). *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being*. New York: Basic Books.

Mason, J., Burton, L. & Stacey, K. (1982). *Thinking Mathematically*, London: Addison Wesley.

Moss, R. & Roberts, G. (1968). *A Preliminary Course in Analysis*, London: Chapman & Hall.

Roberts, S. (2016). Mathematical Beauty: A Q&A with Fields Medalist Michael Atiyah. https://www.scientificamerican.com/article/mathematical-beauty-a-q-a-with-fields-medalist-michael-atiyah/

Stewart, I. N. & Tall, D. O. (2014). *Foundations of Mathematics*, Second Edition. Oxford: OUP. ISBN: 9780198531654

Stewart, I. N. & Tall, D. O. (2015). *Algebraic Number Theory and Fermat's Last Theorem*, Fourth edition. Boca Raton: CRC. ISBN 9781498738392

Stewart, I. N. & Tall, D. O. (2018). *Complex Analysis*, Second edition. Cambridge: CUP. ISBN: 978-1108436793.

Tall, D. O. (1980a). Intuitive infinitesimals in the calculus, *Abstracts of short communications, Fourth International Congress on Mathematical Education, Berkeley*. http://homepages.warwick.ac.uk/staff/David.Tall/pdfs/dot1980c-intuitive-infls.pdf

Tall, D. O. (1980b). Looking at graphs through infinitesimal microscopes, windows and telescopes, *Mathematical Gazette,* **64** 22– 49.

Tall, D. O. (1985). Understanding the calculus, *Mathematics Teaching*, 10, 49– 53.

Tall, D. O. (2013). *How Humans Learn to Think Mathematically*. New York: Cambridge University Press. ISBN: 9781139565202.

Tall, D. O. (2019). Complementing supportive and problematic aspects of mathematics to resolve transgressions in long-term sense making. *Fourth Interdisciplinary Scientific Conference on Mathematical Transgressions*, Krakow, March 2019.

Tall, D. O. & Katz M. (2014). A Cognitive Analysis of Cauchy's Conceptions of Function, Continuity, Limit, and Infinitesimal, with implications for teaching the calculus. *Educational Studies in Mathematics*, 86 (1), 97-124.

Tall, D. O., Tall, N. D. & Tall, S. J. (2017). Problem posing in the long-term conceptual development of a gifted child. In Martin Stein (Ed.): *A Life's Time for Mathematics Education and Problem Solving*. *On the Occasion of Andràs Ambrus*' *75th Birthday*, pp. 445-457. WTM-Verlag, Münster.

van Hiele, P. M. (1986). *Structure and Insight*. Orlando: Academic Press.

Woodward, J. (2004). Mathematics Education in the United States: Past to Present. *Journal of Learning Disabilities*, 37 1, 16–31.