Making Sense of Mathematical Thinking over the Long Term: The Framework of Three Worlds of Mathematics and New Developments

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

Abstract

The framework of 'three worlds of mathematics' was designed to reveal the growth of mathematical thinking in individuals over a lifetime to include all abilities and interests (Tall, 2013). This paper outlines the framework and introduces new developments that simplify and extend it. These involve aspects that anyone can observe, such as how we *say* mathematical expressions, how we *hear* someone else speak mathematically, how we *see* moving objects, how we *think* about mathematical symbols as processes or mental objects. These new developments have direct application to mathematical thinking at all levels, in different individuals, in differing cultural settings, and also in our understanding of the historical growth of the subject. They offer new ways of making sense of 'math wars' in which different approaches to mathematical ideas cause debates over which is preferable or even correct. The broader framework takes account of differing approaches by different communities of practice. The plan is to offer factual information so that readers can make their own judgement of how to proceed in their current situation, to make sense of mathematical thinking over the long term. This is formulated to take account of the fact that different communities may have valid approaches that are appropriate in their own context but are problematic in another.

1. Introduction

The last half century has seen immense changes in our understanding of mathematical thinking, not only through the development of mathematics education as a research topic but also through the invention of digital technology and a deeper understanding of the structure and function of the human brain. These changes have been so immense in such a relatively short time that they probably outstrip any other period of mathematical development in the whole evolution of our human species. It is an interesting time to be alive.

How Humans Learn to Think Mathematically (Tall, 2013) sought to construct an overall framework for long-term growth of mathematical thinking from child to adult. It also proved insightful in making sense of historical development where intellectual adults in history have a personal development from child to adult occurring in different societies with different accumulated resources. The framework is based on three distinct forms of thinking that have evolved in increasingly shorter periods of development: 'embodiment', evolving in many species over millions of years, 'symbolism', evolving in in a mathematical sense in *Homo Sapiens* over a period of around fifty thousand years, Greek ideas of 'formal' proof arising about two and a half thousand years ago, developing into set-theoretic 'axiomatic formal' proof in the last century.

Our children are faced with making sense of appropriate parts of mathematics within a lifetime. A major aim is to understand why some individuals are highly successful in gaining pleasure and power through thinking mathematically while others find mathematics a source of anxiety and confusion.

Part of the task is to consider the mathematical content and development of different topics in arithmetic, algebra, geometry, calculus and more advanced forms of mathematics. But it is equally important to study how we actually *think* about mathematics. This involves

Draft Paper for Tall, D. & Witzke, I. (2019). MINTUS: Beiträge zur mathematischen, naturwissenschaftlichen und technischen Bildung. Wiesbaden: Springer.

investigating what is known about the structure and operation of the human brain and to translate it into a form that is relevant to teachers, learners and others involved in the development and use of mathematics.

This paper first summarises the main ideas of the three-world framework in terms of the increasing sophistication of mathematics and how a human brain operates mathematically. It takes account of both the development of mathematical knowledge and the emotional reactions to changes in mathematical context. Some aspects of previous experience may continue to work and are *supportive* in a new context, but others may involve changes of meaning that are *problematic*. For example, arithmetic facts such as '2+2 makes 4' continue to be supportive, but more subtle implicit properties, such as 'you cannot have less than zero', or 'adding a number gives a bigger result' no longer hold when introducing negative numbers.

The new extensions to the framework consider simple observations of how we operate in everyday life – speaking, hearing, seeing, thinking, and communicating with others. These offer insight into how we can improve long-term understanding of mathematics by highlighting simple principles that remain supportive over several changes in context while dealing explicitly with problematic aspects that impede development. For instance, it is possible to begin to make sense of complicated mathematical expressions by realising that a simple expression such as 'two plus three times four' can have different answers depending on how it is spoken. (Try saying it in different ways by leaving short gaps between different words.)

By becoming aware of this ambiguity, it becomes more reasonable to seek meaningful ways to symbolise the difference in meaning rather than simply present 'rules' such as 'multiplication takes precedence over addition' to be learnt by rote. This leads to the Articulation Principle in which the spoken articulation of an expression gives it an unambiguous meaning. This proves to be valuable in giving genuine meaning to mathematical expressions, not only in simple arithmetic, but also throughout the whole mathematics curriculum.

Over the longer term, mathematics educators are aware that expressions such as '2 + 3' are initially interpreted as operations to be performed in time in a variety of possible ways, and then as a single mental object – the sum '2+3', which is '5'. This paper offers an explicit new way of interpreting expressions and sub-expressions as operations or objects that fits with the mathematical meaning, as opposed to reading words in the standard textual sequence.

In the wider scheme of making sense of mathematics, different communities will have very different needs to satisfy very different objectives. For instance, most people use relatively simple mathematics in their everyday lives, but various professions require very different kinds of mathematics for different tasks. Some may involve practical mathematics in commerce, some may require more theoretical mathematics to model a situation and predict its outcome. A small proportion of the population may go on to research more advanced areas of pure mathematics and logic.

A major aim of this paper is to present a general framework that takes an overall view of the differing approaches of different communities. Its purpose is to enable each of us to see our own viewpoint as part of a broader journey to make long-term sense of mathematical thinking appropriate for differing needs.

The journey begins in the next section by considering aspects of the operation of the human brain that support the long-term growth of mathematical thinking. While it is evident that language is a foundation for the success of our species, we will find that there are other aspects of human perception and action that are mathematically even more essential. Based on these, we will move on to consider observable aspects of human activity that contribute significantly to success and failure. This offers the potential for focusing on supportive aspects that enhance confidence and insight over several changes in context, while dealing with

problematic aspects that inhibit mathematical thinking for differing individuals in their own society. It will lead to new ways of simplifying increasingly sophisticated mathematical ideas that can be integrated into current approaches to teaching and learning.

2. Mathematical operations in the brain

The brain has two essentially symmetric halves performing complementary roles that are connected together to operate as a whole. Facility with language takes place in the left brain for all but a few who tend to be left-handed. Wernicke's area, making sense of spoken input from both ears, is in the back part of the left brain just behind the left ear. Broca's area, responsible for spoken output, is further forward on the left side. The right-hand side deals with more global non-verbal thinking.

The front part of the brain has an overview executive function, making more conscious decisions. In the centre of the brain is a complicated array of structures, collectively referred to as the limbic system, which performs diverse tasks, such as laying down and fetching long-term memories and responding emotionally to pleasure and danger.

Figure 1 shows a view from above the brain and a view of the inside of the brain, revealing one side of the limbic system, which is in two parts symmetrically placed on either side of the brain that are connected together.

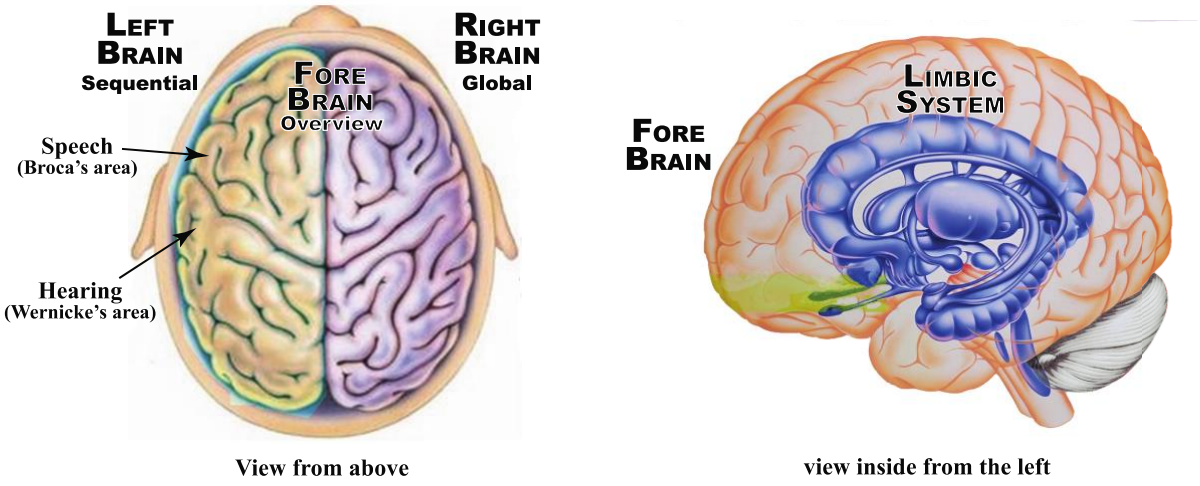


Figure 1: The human brain

As it stands, the complexity of the brain needs substantial work to translate its operation into a form suitable for teachers and learners to link to mathematical thinking in their everyday activities. An alternative approach is to work in the opposite direction to see how the long-term development of mathematics links to the operation of the brain.

Mathematical thinking arises in the young child with an intuitive sense of space and number and develops into more structured forms of geometry, arithmetic, algebra, calculus, and, for a few experts, into more formal mathematics, logic and mathematical proof. All of these involve thinking about *objects* (mental or physical), *operations* on objects and their *properties*. Some of these activities (such as geometry) focus more on *objects and their properties*. Others (such as arithmetic and algebra) focus on *operations and their properties*. More sophisticated formal mathematics, used mainly by pure mathematicians, focuses on structures based solely on *properties* given by axioms and definitions from which all other properties are deduced by formal proof.

2.1 Three worlds of mathematics

The framework for long-term mathematical thinking based on three interrelated strands of developing sophistication is called 'the three worlds of mathematics' (Tall, 2013). One focuses on objects and their properties, initially physical, then constructed mentally, termed *conceptual embodiment*. Another focuses on operations and their properties, called *operational symbolism*. Both develop in sophistication from *practical mathematics* based on the *coherence* of observed properties to *theoretical mathematics* where one property may be deduced as a *consequence* of another. A third strand, that develops from theoretical mathematics into the formal mathematics of the twentieth century based on *properties*, defined using set theory or logic, is called *axiomatic formalism*.

These three names can be shortened to 'embodied', 'symbolic' and 'formal' when their meaning is clear. However, each of them is has very different meanings in various theories and it is essential to be aware of their particular meaning in the three-world framework.

For instance, an 'embodied' approach for younger children may use physical materials to 'embody' mathematical concepts (as in multi-base blocks and logic blocks of Dienes (1960), Cuisenaire Rods (Cuisenaire, 1952), the Geoboard of Gattegno (1971)). More generally it may refer to how we 'embody' abstract concepts in our bodily perceptions and actions which involve mental imagery and gesture (Lakoff & Núñez, 2000). Conceptual embodiment refers to the long-term development of the properties of objects as their conception becomes more sophisticated, from making sense of the relationships between physical objects, to more abstract relationships between mental objects.

The term 'symbol' is generally used to denote 'a word or mark or anything that represents or signifies an idea, object, or relationship.' Bruner's (1966) classification of communication consists of three different modes: 'enactive' (based on gestures) 'iconic' (based on pictorial imagery) and 'symbolic'. He used the term 'symbolic' to apply not only to natural languages but also to specialist areas such as arithmetic or logic, which form the basis for the operational symbolic and axiomatic formal worlds respectively. The term 'operational symbolism' specifically refers to symbolic expressions that represent operations, such as addition of two numbers 3+2, with the understanding that the same symbol can also stand for a mental object, namely 'the sum of 3 and 2', which is '5'.

The term 'formal' is used by pure mathematicians to refer to structures specified by verbal axioms and definitions which includes the axiomatic approach of Euclidean geometry. In the three-world framework, Euclidean geometry is classified as 'theoretical mathematics' along with mathematics in natural science and other applications which involve modelling real world problems and deducing consequences. This is because it is inspired by naturally occurring objects (in this case, geometric figures) which are then formulated verbally in terms of axioms and common notions from which all other properties are deduced by Euclidean proof.

There is an essential difference between the long-term development of embodiment and that of symbolism. Whereas embodiment focuses mainly on objects and classifies their properties verbally in increasingly sophisticated ways, symbolism goes through many individual stages, encountering new ways of operating with new forms of number, first through calculation, then through manipulation of increasingly sophisticated symbolism. The development of symbolism is therefore more intricate than that of embodiment.

Before the end of the nineteenth century, the study of mathematics and science based on naturally occurring phenomena was described as 'natural philosophy'. It is useful to distinguish between *theoretical* mathematics, based on naturally occurring phenomena, and *axiomatic formal* mathematics, based only on properties defined using set theory and logic. Axiomatic formal mathematics can be applied to *any* context where the axioms and definitions hold. It is

conceptually more powerful than theoretical mathematics as it applies not only to known contexts, but also to any as yet unknown future context that satisfies the axioms and definitions. In this sense it is 'future-proofed' in that it applies to future evolution of ideas in axiomatic formal mathematics, always with the possibility that we will develop even more sophisticated ways of thinking in the future that we have not yet considered.

I will later show that the future is already here in the sense that it is possible to prove formal *structure theorems* that take us on to more sophisticated forms of embodiment and symbolism. This can still be encompassed in the current framework with three forms of mathematics (embodiment, symbolism, formalism) in three levels of sophistication (practical, theoretical, axiomatic formal) (Figure 2).

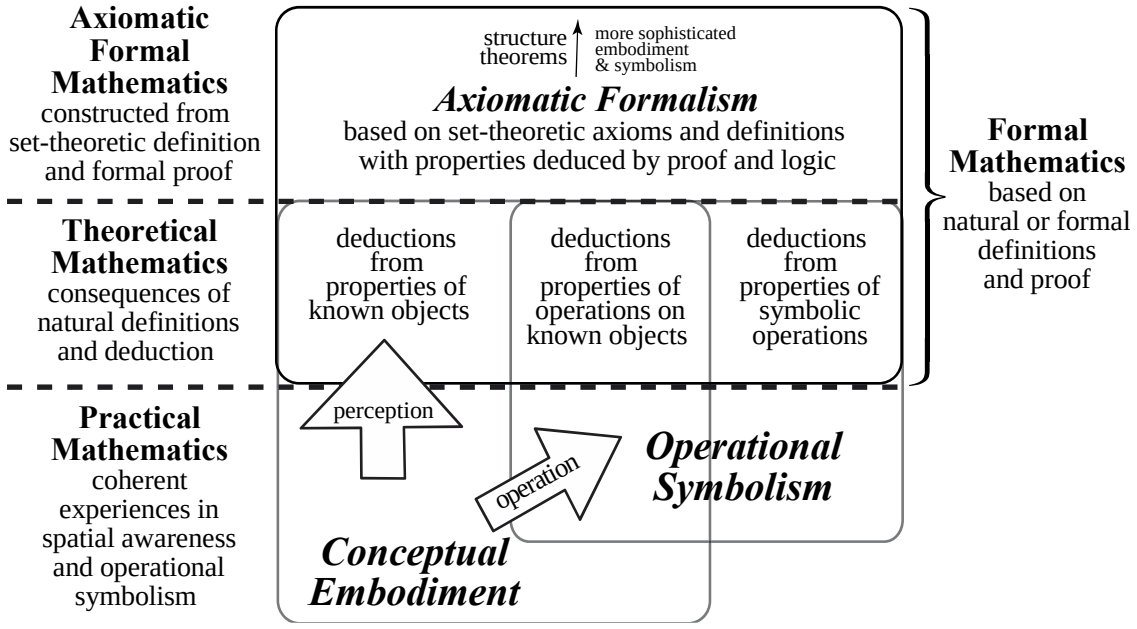


Figure 2: the long-term development of mathematical thinking

These three different worlds develop in succession over time, both in the life of an individual and in historical evolution. In the Piagetian theory of individual development, the young child begins with a 'sensori-motor' stage, passes through a 'pre-operational' stage to 'concrete operational', then to a 'formal operational' stage which develops in adolescence as abstract thought independent of concrete referents. Piaget's theory is foundational in focusing on the long-term changes of thinking in the individual, but it needs more specific clarification to apply to the long-term development of mathematics. The three-world framework for mathematics follows a corresponding path from embodiment through practical mathematics for everyday use and, as appropriate, to more theoretical and formal mathematics.

In historical evolution, the early development of embodiment existed in our ancestors, and in many other species, hundreds of thousands of years ago. Operational symbolism evolved in *Homo Sapiens* in the last fifty thousand years or so, proliferating in various communities in Egypt, Babylon, India, China around five thousand years ago, becoming increasingly theoretical in Greek mathematics with the first flowering of mathematical proof two and a half thousand years ago. Axiomatic formal mathematics has been around for little more than a century. Now new possibilities are emerging in our digital age, enabling *Homo Sapiens* to use new digital tools to enhance an embodied (enactive) interface, dynamic visualisation, symbolic computation and the emergence of new forms of artificial intelligence.

2.2 Studying brain structure and operation

The picture in figure 2 is by no means complete. It omits emotional affective aspects and offers no detail about brain activity, although further analysis of the growth of mathematical sophistication will include the study of supportive and problematic changes in context.

Modern techniques include non-invasive study of surface brain activity by attaching electrodes to the head, and magnetic resonance imagery (MRI) to map the internal structure. More subtle techniques using functional MRI (fMRI) can track the activity of the brain over time, but this only measures the flow of blood moving to more active parts of the brain over a period of seconds and lacks the resolution to record the subtlety of mathematical thinking that changes in milliseconds.

In the paper *Left Brain, Right Brain: Facts and Fantasies*, Corbalis (2014) exposes some of the suppositions about brain activity that are widely believed, yet either lack, or are contradicted by, empirical evidence. For instance, even though language may function mainly on one side while the other deals with non-verbal intuition, the actual operation of the brain is far more complex as the two sides cooperate together. Furthermore, there is evidence that, although language plays a vital role through verbalising properties of arithmetic and geometry, mathematical thinking also involves activities that may not link to language at all, as suggested by a recent study that declares:

> Our work addresses the long-standing issue of the relationship between mathematics and language. By scanning professional mathematicians, we show that high-level mathematical reasoning rests on a set of brain areas that do not overlap with the classical left-hemisphere regions involved in language processing or verbal semantics. Instead, all domains of mathematics we tested (algebra, analysis, geometry, and topology) recruit a bilateral network, of prefrontal, parietal, and inferior temporal regions, which is also activated when mathematicians or nonmathematicians recognize and manipulate numbers mentally. Our results suggest that high level mathematical thinking makes minimal use of language areas and instead recruits circuits initially involved in space and number. This result may explain why knowledge of number and space, during early childhood, predicts mathematical achievement. (Almeric & Dehaene, 2016)

The three-world framework observes that language plays different roles in each world. Conceptual embodiment relies on 'categorization' of properties, using language to formulate successive levels of embodied thought, with a focus on thought experiments to imagine situations spatially. Successful operational symbolism uses symbols to 'encapsulate' processes as flexible mental objects at successive levels of sophistication in ways that may not be available to learners who only learn by rote. Axiomatic formal mathematics can be inspired by spatial and/or symbolic thinking to formulate set-theoretic definitions and make conjectures to prove properties by formal proof.

The overall theory of three worlds of mathematics has two over-arching concepts:

> *set-before*: a genetically endowed facility 'set before' birth in our genes,

and

> *met-before*: an aspect of the current mental state as it is affected by experiences that were 'met before' by the individual in previous contexts.

(The term 'met-before' arose to consider the new context from the individual's personal viewpoint, as compared to the related term 'metaphor' which interprets the situation from a more sophisticated top-down philosophical viewpoint.)

Language, which plays a major role in formulating the overall theory, is specified as a 'set-before'. It is used extensively throughout the theory to give explicit explanations of our

conscious mental ideas while many links within our brain relate to visual and operational activities that are performed automatically and sub-consciously.

It is helpful, wherever possible, to become explicitly aware of these underlying processes and the longer-term development of more sophisticated processes. For example, the human brain is very good at recognizing the same thing from different viewpoints, such as being able to recognize a human face from any direction. Less obvious is the manner in which different mathematical operations give rise to the same mathematical concept.

2.3 From counting processes in time to the concept of number as a mental object

An important principle that takes time to realise is the fact that the number of objects in a given (finite) collection is independent of the way it is laid out and how it is counted (Figure 3).

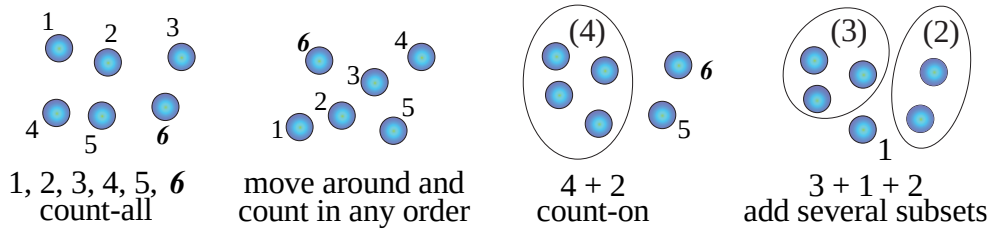


Figure 3: The number of objects in a collection is independent of order and layout

This underlies the general idea that the sum of a list of whole numbers is independent of order and method of calculation. The same idea works for the product of two whole numbers exemplified by the idea that six objects can be placed in two rows of three or three columns of two (Figure 4).

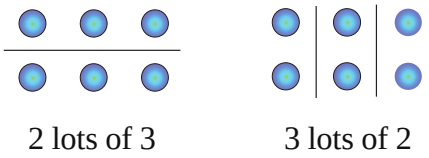


Figure 4: The product of two numbers is independent of order of calculation

These are both instances of Piaget's notion of conservation of number:

> **The Principle of Conservation of Number:** The number of objects in a collection is independent of the layout and of the manner of counting.

This leads to more general principles of arithmetic:

> **The General Principle of Addition for Numbers**: A finite sequence of additions of numbers is independent of the order of calculation.

> **The General Principle of Multiplication for Numbers**: A finite sequence of multiplications of numbers is independent of the order of calculation.

These principles belong initially in *practical* mathematics, for example, adding a column of numbers is independent of the order of operations and continues to generalise to addition of signed numbers, decimal representations, real numbers and even complex numbers. They operate as supportive principles throughout elementary mathematics.

The theoretical sequence, in which addition and multiplication are defined as binary operations that satisfy certain rules from which more general properties can be deduced, is far more sophisticated.

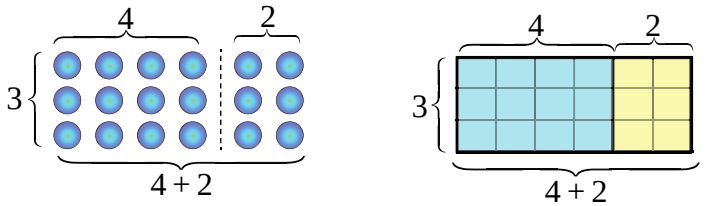
Brackets are not required in practical mathematics until addition and multiplication are used in the same expression. Figure 5 shows two representations of 3 rows with 4+2 objects in

each. The first has distinct objects and the numbers are used to *count* the objects: the second has an area with vertical side 3 and horizontal side 4+2 and numbers are used to *measure* the area. Using brackets to enclose the sum 4+2, both pictures represent

$3 \times (4 + 2) = 3\times4 + 3\times2.$

The same layout works whatever numbers are used, giving a generic view of the property that later generalises to the algebraic distributive property:

$a\times(b + c) = a\times b + a\times c.$



$3\times(4+2) = 3\times4 + 3\times2$ as whole numbers

$3\times(4+2) = 3\times4 + 3\times2$ as areas

Figure 5: The distributive property

This illustrates the **Distributive Law**, which easily generalises to having several numbers inside the brackets, such as

$3 \times (4 + 2 + 5) = 3\times4 + 3\times2 + 3\times5$

with the additional flexibility that the result of the calculation is independent of the order of terms inside the brackets.

It is also possible to formulate a more general principle for multiple products of sums inside brackets. However, this increases the complication and it is sensible in the initial stages to deal mainly with the simple case of a single number times a bracketed list of numbers, which is unchanged if the order of the list is changed.

The literature is full of detail showing that many students have difficulty passing from arithmetic to algebra.

We now consider how these problematic aspects may be related to the natural way in which we operate as human beings.

3. Simple observations that can be noticed by any reader

There are ways in which we can see the mental subtleties of mathematical thinking with our own eyes and ears. All we need to do is to pay attention to what is happening automatically as we read and speak mathematically as compared with other aspects of everyday life.

3.1 How humans read text

If you read this paragraph several times and notice what happens to your eyes, you will sense that they move in a sequence of jumps alighting temporarily on small parts of the text as the brain builds the meaning of the text by putting the pieces of information together.

Please read any paragraph on this page, several times if necessary, so that you become aware of what is happening. 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.

The retina at the back of the eye has millions of cells called rods and cones that react to light. Rods are more numerous and are highly sensitive for night vision. Cones are sensitive to colour and detail in daylight and predominate in a circular area around 5.5 mm in diameter (the macula) with an even more sensitive central area consisting mainly of cones around 1.5 mm in

diameter (the fovea) (Figure 6). The edge of the macula overlaps the blind spot where the eye sends signals to the brain along the optic nerve.

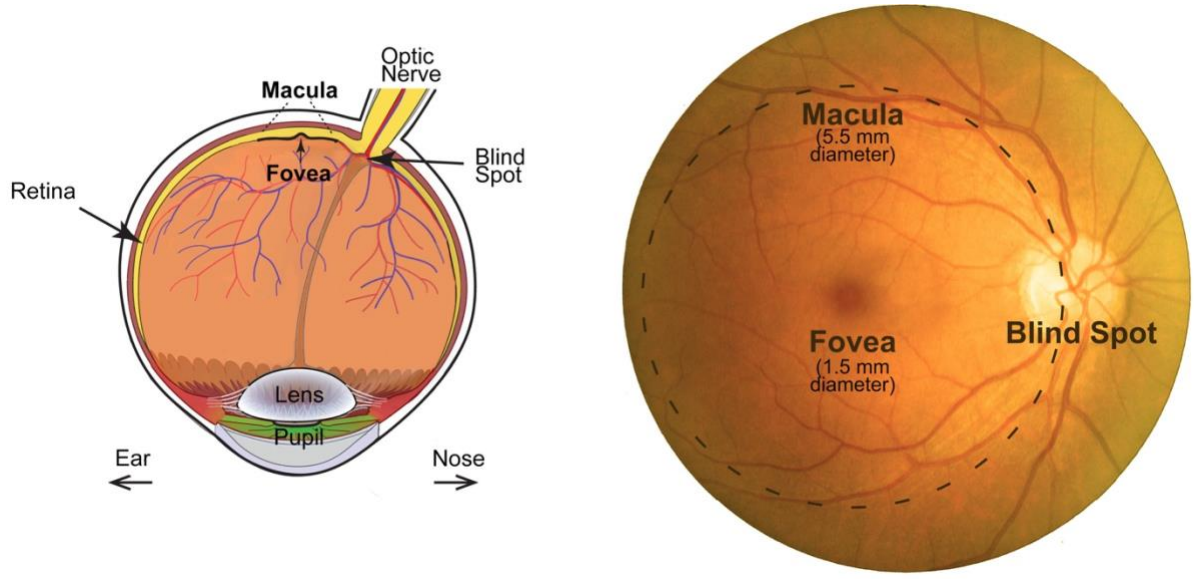


Figure 6: The central part of the retina seen through the pupil of the eye

The fovea is estimated to have around 200,000 cones (Kolb, 2007). The precise number is not important, but the order of size is. A modern smart phone with a so-called ‘retina screen’ may have a million or more pixels which is considerably larger than the number of cones in the fovea. Depending on how far you hold the phone from your eye, you may be able to focus on a different number of characters. When I hold my iPhone a comfortable distance from my eye to look at a list of song titles in iTunes, I can read the title ‘Wonderful’ in a single glance but it takes two or three saccades or even more to read ‘My one and only’ or ‘Nice work if you can get it’. You should look at text on a phone to get a sense of the phenomenon for yourself.

The eye focuses only on a few syllables of text at a time, and the brain puts the separate pieces together to build up the meaning. With languages that are written from left to right, reading text is performed sequentially in the same direction, though there may be small variations that can be accommodated over short stretches. (For instance, in German, the numtber 123 is read as ‘ein hundert, drei und zwanzig’, which is ‘1 hundred, 3 and twenty’, requiring the digits to be spoken in the sequence 1, 3, 2.) In this case, the number can be read in a single chunk and the brain is able to put the chunks together as the eye jumps along a line.

If the same technique is used to read an expression such as

$$2 + 3 \times 4$$

then this is likely to be read in sequence as ‘two plus three times four’. This natural order interprets ‘two plus three’ as ‘five’ and then ‘five plus four’ is ‘nine’. However, in arithmetic we are taught the convention that ‘multiplication takes precedence over addition’, so we must first calculate ‘three times four’ to get ‘twelve’, then ‘two plus twelve’ gives ‘fourteen’.

For many this is a bewildering experience, as it violates the natural sequence for interpreting language. It can give the impression that mathematics involves arbitrary ‘rules’ that need to be remembered even if they have no meaning. If such rules are remembered using language, neurophysiological observations suggest that they may occur in language areas of the brain without meaningful connections to areas related to space and number. If this happens, long-term learning in mathematics may involve learning ‘rules without reason’ so that mathematical ideas are not well-connected and become more complicated and error prone.

But need it necessarily be this way? When we speak language, we use other techniques such as tone of voice and articulation to give extra meaning. Why not seek to use such techniques to make sense of arithmetic?

3.2 How we speak mathematical expressions: the Articulation Principle

When we speak an expression such as '2 + 3 × 4', we can leave small gaps between words to give different meanings. An ellipsis symbol '…' consisting of three dots is often used to mark where words are omitted in a quotation, usually enclosed in square brackets as […]. Here I use the ellipsis symbol on its own to represent a slight break between characters when the expression is spoken. If we leave a slight gap in '2 + 3 × 4' after 3, written as '2 + 3 … × 4' then this is read as '2 + 3', which is '5', then '× 4' which gives '5 × 4' which is '20'. Written symbolically, the different articulations for

2 + 3 × 4

can be spoken as

2 + 3 … × 4, which is 5 + 4, giving 9

or as

2 +… 3 × 4, which is 2 + 12, giving 14.

Once the ambiguity for the expression 2 + 3 × 4 is recognised, it leads to:

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

This is not a definition in a mathematical sense. However, it is fundamental to giving meaning to mathematical thinking throughout the long-term development of the subject. For instance, the expression '$-x^2$' is usually read as 'minus x squared.' Does it mean 'minus x … squared' or 'minus … x squared'? According to the rule 'the product of two minuses is a plus', for $x = -2$, the first is +4, but the second is –4. If such rules are learnt verbally by rote without linking to deeper mathematical meaning, this can only lead to increasingly complicated error-prone long-term learning. The mathematics education literature is replete with horror stories of student misconceptions.

The articulation principle leads to a more natural way of expressing meaning through the use of brackets (or 'parentheses' in American English) to indicate what parts of an expression need to be given precedence. For instance,

2 + 3 … × 4 may be written as (2 + 3) × 4,

and

2 +… 3 × 4 as 2 + (3 × 4).

Using the articulation principle and the general principles for addition and multiplication, this offers a meaningful starting point for precise interpretation of mathematical expressions.

Combining addition and subtraction involves using brackets to clarify meaning of expressions such as 5 – 2 + 1. Here the principle of articulation shows that 5 – 2 … + 1 is very different from 5 – … 2 + 1 and these can be distinguished as (5 – 2) + 1, which is 4 and 5 – (2 + 1), which is 2.

In the absence of brackets, it is natural to perform operations in sequence left to right, so that the expression 5 – 2 + 1 gives 4. When adding a collection of whole numbers, the order of addition does not matter. This also happens when addition and subtraction are mixed. But there is one problematic aspect. If the terms are re-ordered as 1 – 2 + 5, then the first operation is not possible working with whole numbers counting objects because you can't

take two objects away if you only have one. In this case, the operations need to be performed in an order where each step gives a whole number result.

However, over the longer term, when contexts arise that include negative numbers (such as a bank account or temperatures above or below zero) then an extended general principle arises:

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

This principle makes sense as new number systems arise, including signed numbers, real numbers and complex numbers. It also holds later for constant and variable quantities in algebra and symbolic calculus.

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) and for real and complex numbers, where division by a non-zero number is always possible. For fractions, rational, real and complex numbers there is a longer-term principle:

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

Here the term 'quantity' will again apply to constant and variable quantities in algebra and calculus. In other more advanced contexts, such as linear algebra or group theory, there are major changes in which multiplication is no longer independent of order of operation. This problematic situation will be more easily addressed if learners have been encouraged over the years to deal explicitly with new ideas that do not fit with previous experience.

In any context where both principles are satisfied, their combination with the use of brackets and the distributive law will be called

> **The General Principles of Arithmetic** for brackets, exponents and operations +, –, ×, ÷.

These principles using the operations of arithmetic are satisfied in arithmetic and algebra and in all number contexts from whole numbers, integers, rational numbers, to real numbers and complex numbers. In particular, they make coherent sense in practical mathematics, satisfactory for everyday use.

3.3 Interpreting an expression as operation or object

To make sense of more sophisticated symbolism in arithmetic, algebra and calculus, it is helpful to be able to see how an expression is built up hierarchically from sub-expressions. This can be achieved in an explicit manner by paying attention to how the component parts operate successively as operation or object.

An expression, such as 4 + 2, is first encountered as an operation of counting that can be performed in several different ways. Later the same expression can be conceived as a mental object, the sum 4 + 2, which is 6.

This idea that an expression can be conceived as an operation that takes place *in time*, or as a mental object that can be manipulated as an entity in the mind, is crucial to coping with increasingly sophisticated use of symbolism throughout long-term mathematical development.

The literature introduces a variety of terminology and subtle differences in meaning to discuss the mental compression of a process into an object. Here I will use the term 'operation' interchangeably with 'process', and the word 'object' interchangeably with 'concept'. Gray & Tall (1994) referred to an expression that can be used as a process or concept as a *procept*. The transition of turning an operation into an object is called 'encapsulation' or 'reification'. Another, simpler word is 'objectify'.

To objectify an idea means to give it a name, then to talk about it as if it were an entity as a whole. Language already has a construct to turn a process into an object, namely the notion of 'gerund' where a participle such as 'walking' is a process when used as part of a verb, as in 'I am walking', but then becomes a noun in a statement, as in 'walking is good for my health'.

The transition from process to object occurs throughout mathematics. For instance, the process of sharing, represented by fractions, can have many different processes giving the same object. Dividing something into six equal shares and selecting three (written as 3/6) is a different process from dividing into four and selecting two (2/4) but they both give the same quantity (½) and, when marked on a number line, they are the same point. The fractions 3/6 and 2/4 are said to be 'equivalent' but, as a rational number, they are one and the same. As operations they are different, but embodied as a point on the number line, they give a single point.

In algebra, a symbol such as $x^2 - 1$ may be conceived both a process ('square the value of x and take away 1') and also as an object that can be operated upon, such as being factorised to give another expression $(x + 1)(x - 1)$. In the calculus, a symbol such as $\int f(x)\,dx$ is both an instruction to carry out the operation of finding the integral of the function $f(x)$ and also the value of the integral which is an entity that can itself be manipulated.

3.4 Representing an expression as operation or object

As mathematics becomes more sophisticated, the shift from operation to object is not always made explicit. In learning to be aware of the two meanings, it is possible to make the distinction by placing boxes around sub-expressions that are thought of as objects (Tall, 2019a). A symbol such as 4 + 2 can be written as:

$\boxed{4}+\boxed{2}$ as the *process* of adding objects 4 and 2,

$\boxed{4\ +\ 2}$ as the mental *object*, 4 + 2.

Initially it may be helpful to draw the boxes explicitly, but once the distinction is made, it may be imagined in the mind's eye to take account of the hierarchical structure of expressions.

To interpret more extended expressions involving operations with different orders of precedence, the technique is to look along the expression from left to right to find the occurrences of the operation with the highest preference and put them inside a box. For example, with the convention that 'multiplication has precedence over addition' the operation of highest precedence in the expression '2 + 3 × 4' is '3 × 4'. Placing this into a box, as an object, reveals the expression as a sum of objects:

$\boxed{2}+\boxed{3\times 4}$.

By the general principles of arithmetic, the objects $\boxed{2}$ and $\boxed{3\times 4}$ can be combined in any order, as can the numbers in the product $\boxed{3\times 4}$. Of course, the numbers 3 and 4 here are also objects and could be placed inside boxes, but visually it is clearer not to overcomplicate the notation.

If there is more than one instance of the highest order of operation in succession, then these should be placed in a single box. For instance, in the expression '3 + 2 × 4 × 4 – 1' the terms '2 × 4 × 4' are associated together to give

$\boxed{3}+\boxed{2\times 4\times 4}-\boxed{1}$

By the general principles of arithmetic, the boxes can be placed in any order as can the operations inside the product box $\boxed{2\times 4\times 4}$.

This principle extends naturally to more general expressions where there is an agreed order of precedence, such as the order given in the USA curriculum as

Parentheses – Exponents – Multiplication/Division – Addition/Subtraction

which is memorised using the mnemonic PEMDAS (Please Excuse My Dear Aunt Sally). Parentheses are given the highest precedence, then Exponents, then Multiplication and Division together at the same level, followed by Addition and Subtraction together at the lowest level. In the UK the mnemonic BIDMAS is used for the same purpose with levels representing

Brackets – Index – Division/Multiplication – Addition/Subtraction.

Neurophysiological evidence suggests that, if these mnemonics are learnt by rote, they may connect to language areas of the brain but not to areas that deal with number (Maruama et al. 2012). Personally, I wonder how 'Dear Aunt Sally' links to mathematics … There is enormous evidence in the literature to show that multiple difficulties arise in many students as they encounter more complicated expressions.

These new principles have the potential to introduce meaningful connections to the conventions of operational symbolism. The articulation principle reveals the ambiguity of speaking and hearing mathematical expressions, offering a meaningful reason to require the use of brackets and to interpret notation more precisely. The duality of symbolism as operation and object offers insight into the hierarchical meanings of expressions nested one within another.

3.4 Compression of knowledge over the longer term

Over time, as the brain makes new connections, more sophisticated thinking becomes possible as operations that occur in time are symbolised and conceived as mental objects that can then be manipulated at a more sophisticate level.

The successive compression of ideas into more compact forms can be shortened by further conventions to reduce the number of symbols required. When single letter variables are involved, we can omit the multiplication sign between numbers and variables and between variables themselves. By the general principle of multiplication, we can write a product of several terms with the number first and the variables in alphabetical order. So, $b \times 2 \times a$ can be written as $2ab$. If a power is involved, as in x^2, with the power written as a superscript as x^2, then the power is of higher order than a product, so $2 \times x$^2 $\times a$ can be written as $2ax^2$. With the lower order operations +, × remaining explicit, this gives a compact notation for a quadratic expression such as $ax^2 + bx + c$ where the term ax^2 clearly means $a \times x^2$ because the power has a higher precedence than the product.

Later developments represent expressions in two-dimensional layouts rather than simply on a straight line. For instance, division of rational expressions is written spatially with one expression written above another with a horizontal line in between. Once again, a rational expression can be seen as an object or an operation (Figure 7):

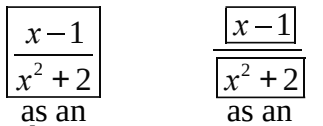


Figure 7: A rational function as an object or operation

The same technique works for any expression as a two-dimensional template written by hand or using a digital layout system such as TeX or MathType, whether it is a unary operation (such as the square root) or a more general expression, such as the solution of a quadratic equation (Tall 2019a, 2019b). The process of reading an expression as a hierarchy can be interpreted, first by seeing the whole expression as an object, then looking at it as an operation in which any expressions in brackets are boxed as individual objects, then burrowing down recursively in each object as an operation eventually builds the hierarchical meaning of the whole expression.

This is not an easy task for most learners. It is virtually impossible for the many who simply learn the rules by rote, but it is a process that seems to be performed implicitly by those who make sense of the hierarchical structure.

The ‘box method’ is not proposed as another rote-learnt technique to help students get ‘the right answers’. What is more important is that the learner senses the different strengths of binding operations to make sense of more complicated expressions. Over time, with increasing familiarity, it may enable the structure to be conceptualised in the mind’s eye to manipulate the symbols subconsciously while focusing consciously on more sophisticated ideas.

3.5 How the eye follows a moving object

Hold a finger in front of your eye and move it sideways, keeping your gaze on the finger as it moves. Do this now and sense what is happening.

Whether you keep your head still and move your eyes, or you move your head to follow your finger, you should find that the eye remains smoothly in focus while the background is out of focus. This is using the same mechanisms in the eye and brain that are employed in reading, but rather than jumping along text in saccades, there is a single saccade to jump to focus on the finger and then the moving finger is followed in a smoother fashion.

This relates to the fact that the fovea in the eye which takes in the highest detail has around 200,000 cones which gives a linear diameter with around 500 cones in a line. Each of these takes a fraction of a second to register a signal. The time for this to happen can again be evidenced by personal experience. Standard movies on a computer are currently set to renew the picture at 25 frames a second in the UK and 30 frames a second in the USA though faster speeds are now becoming widely available. Yet only a decade or so ago, movies on a computer were set at a much slower rate and a speed of around 15 frames per second is as low as can be used to give a sense of continuous movement. Any lower and the viewer becomes aware of the separate pictures.

An object moving along in a line will successively alight a thin strip of cones in the retina as it moves along, continually refreshing in short periods of time, to detect movement. What we actually see is not a mathematical real line with infinite decimals, but a practical line which may be described as a *continuum*.

The Oxford English Dictionary defines a continuum as ‘A continuous sequence in which adjacent elements are not perceptibly different from each other, but the extremes are quite distinct’, while the Cambridge Dictionary says, 'something that changes in character gradually or in very slight stages without any clear dividing points’.

When we look at a point moving along a line, our eyes see a continuum as the cones in our retina recognising the point come into play successively. If we watch a football match on a retina screen, even though the actual ball may be moving smoothly, the image is moving in imperceptible small steps. It is natural to imagine a point which moves on a line as a *variable* and a point that stays in the same position as a *constant* (Figure 8).

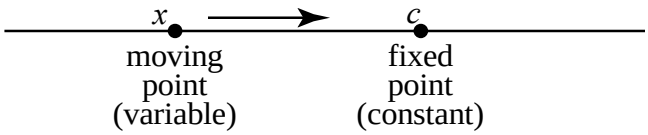


Figure 8: constant and variable points on a line

In a practical embodied sense, we can see a variable point x moving to a constant c and, according to what we *see*, the variable x will become indistinguishable from c. We can therefore imagine that as $x \to c$ the variable tends to the fixed value c. Meanwhile, in the operational

symbolic world, if we consider an expression such as $(x^2 - 4)/(x - 2)$ then, for $x \neq 2$, we can factorise and cancel to get

$(x^2 - 4)/(x - 2) = x + 2$ (for $x \neq 2$).

This brings us to the problem that has dogged us for three and a half centuries: we can let x get 'as close as we wish' to 2 and the answer is $x + 2$, but we cannot put $x = 2$, because then the expression is not defined.

The solution to this dilemma is simple. It is a matter of whether the focus is on the process or on the object. If we think about the *process* of x getting arbitrarily close to 2 without actually getting there, then this is never-ending, but if we focus on the *object* that the expression gets close to, then this is the constant 4.

3.6 Practical, theoretical and formal levels of conceptualising a limit

The three-world framework identifies three levels in the calculation of a limit, whether it is the limit of a continuous function, of an infinite sequence or series, the derivative, integral, the fundamental theorem of calculus, the solution of a differential equation, functions of several variables, partial derivatives, or vector calculus. These three levels are: *practical*, *theoretical* and *formal*.

The *practical limit* is the result of an approximation calculated numerically, symbolically or visually. It is an *object* that arises as a close approximation. Visually the embodied limit object can usually be *seen*. For example, at a later stage of the calculus, consider the Taylor series for $\sin(x)$:

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots + \frac{x^{2n-1}}{(2n-1)!} + \cdots$$

Drawing successive practical approximations for $n = 1, 2, \ldots, 6$, using Mathematica, we can *see* the graph of successive approximations stabilise visually on the graph of $\sin(x)$ (Figure 9: Kidron & Tall, 2015).

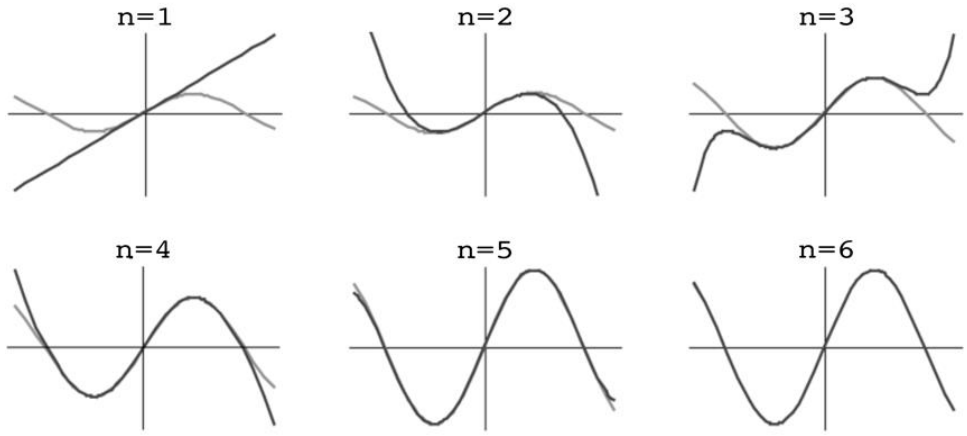


Figure 9: A sequence of practical graphs stabilises on the embodied limit

This is not a formal proof, it is not intended to be. It is a practical visual embodiment in which the sequence of practical approximations stabilises on the limit object. It gives human meaning that can develop into more sophisticated interpretations.

The *theoretical limit* is the object that arises from the (infinite) process of stabilisation. It is the object that the process of approximation gets 'as close as is required.

The *formal limit* in all cases is re-defined in terms of quantified epsilon-delta or epsilon-*N* definitions and proof. Handling the quantifiers is far more sophisticated and is usually postponed to a formal course in analysis. (See Tall, 2013 or Stewart & Tall, 2014).

3.6 Embodying limits, differentiation, continuity, integration and the calculus

Now we have visual interactive displays, it becomes possible to manipulate visual representations of the limiting process as it stabilises on its limit object. This offers insightful new ways of embodying the notions of continuity, differentiation, integration and so on. In particular, it allows us to build a coherent long-term sequence of meanings, starting from practical experience of drawing by hand and moving to the theoretical ideas that can be approached as close as is desired, then (if one desires) to move on to the formal definitions given in mathematical analysis.

Once we grasp that the curve that we draw with a pencil on paper is a continuum, we can begin with the idea that a continuous curve is given by the practical action of drawing a curve dynamically with a pencil without the pencil leaving the paper. In the same way, a number line is just a practical continuum with a chosen unit to represent the numerical length 1 where the line is marked with integer points at equal intervals to the left and right of the origin, with rational numbers added by dividing the units into equal size parts and the realisation that there are even more points, such as $\sqrt{2}$, $\sqrt{17}$, π, e, etc which are not precisely rational numbers.

The real number line is a theoretical construct in which we imagine any position on the real line can be represented precisely as a rational or irrational number. This will later be defined more formally in terms of Cauchy sequences or Dedekind cuts. In the initial stages of drawing graphs, numbers are theoretical 'infinite decimals' that can be represented practically as accurately as is desired by calculating a suitable number of places.

By drawing a horizontal x-axis and a vertical y-axis, we can describe the plane as a two-dimensional continuum with theoretical points identified precisely as ordered pairs (x, y) of real numbers. Now a real function $y = f(x)$ can be drawn practically as a physical curve which we can imagine as a theoretical graph made of up of theoretical points $(x, f(x))$.

The study of calculus involves the two operations of differentiation and integration which are complementary. However, we need to be aware of how we interpret the picture, particularly if we change the scale on the two axes while maintaining the same numerical values. If the scale is changed, the visual slope of the tangent will change, but the symbolic value of the derivative $f'(x)$ and of its practical slope function $(f(x+h)–f(x))/h$ for a small value of h remain the same.

The same happens for the area under the curve. The numerical value of the practical area found by adding up strips width h, height $f(x)$ and the corresponding integral remain the same but the picture looks different. When we calculate the derivative and integral using visual pictures and symbolic calculations, we need to interpret the calculations as remaining the same, even if the scales on the two axes are changed in the picture.

As we study the calculus, we can consider differentiation and integration beginning with either one first. Historically calculating areas and volumes came first because it was of practical use while the rate of change was not a focus of attention because time could not be measured accurately. In modern teaching it is more sensible to study differentiation first because the practical derivative only involves calculating the practical slope function $(f(x+h)–f(x))/h$ for a small value of h, whereas practical integration involves adding the areas of many strips height $f(x)$, width h, which is intrinsically more complicated.

3.6.1 Differentiation: a differentiable function is 'locally straight'

In Leibniz's original formulation of differentiation (1684), he defined what we now term the derivative at a point (x,y) as the quotient dy/dx where dx and dy are the components of the tangent. Its first definition was therefore as the quotient of two lengths. The devil is in the detail: what is meant precisely by 'the tangent' and how can it be calculated algebraically. The word 'tangent' comes from the Latin 'tangere' to touch. In Euclidean geometry, for a circle, the

definition is simple: it is the line through a point on the circle that is at right angles to the radius. It 'touches the circle at just one point and does not cross it.' However, if this definition is used for a more general curve, it is highly problematic. There are many studies that show the complications students encounter with the notion of tangent.

This relates to the long-term development of conceptual embodiment which uses increasingly sophisticated levels of language to categorize human perception and action. In the context of circle geometry, the phrase 'a straight line that touches the circle but does not cross it' clearly identifies a tangent. When the word 'circle' is replaced by 'curve', the description is no longer adequate. To operate with more general curves as graphs in the calculus requires a new mental image to make sense of a tangent in the new context. This is precisely the problem of shifting the context to a new level where language may mean something different, as observed by van Hiele (1986) .

Before the advent of high-resolution dynamic graphics, the usual way a tangent was represented visually was as a static picture in a book (figure 10).

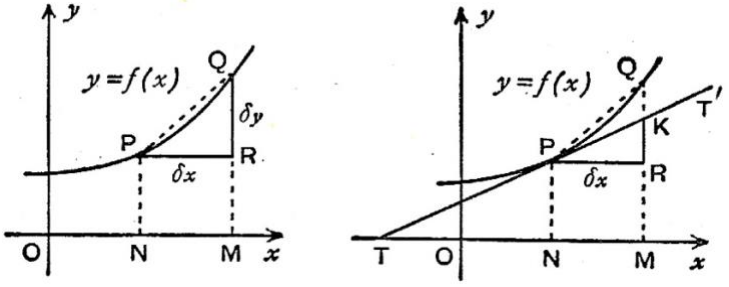


Figure 10: A static picture introducing the tangent (Durell & Robson, 1933)

This method was introduced in England to encourage the use of the continental Leibniz notation alongside the more cumbersome notation of Newton in a textbook by Robert Woodhouse (1803). From here it spread through the English-speaking world including the USA. Woodhouse used the notation δx for the change PR in x and δy for the change RQ in y, to give the slope of the chord PQ as the quotient $\delta y/\delta x$. But he disapproved of Leibniz's interpretation of dy/dx as a quotient of infinitesimals and insisted that it represents the limit of $\delta y/\delta x$ as δx tends to zero. This interpretation passed on from generation to generation. Now we are again aware of Leibniz's original definition, we can see it as the quotient of the components of the tangent vector. The problem is how to interpret the situation and calculate the derivative in a meaningful way.

Using a graphical display, we can zoom in on the curve at P and, if the graph is sufficiently smooth to have a tangent, we will find that a small portion of the magnified graph looks less and less curved as is it magnified. In practical terms, under high magnification, a small part of the graph is indistinguishable from a straight line: it is 'locally straight'.

Magnification of a graph can be performed in many current graphic software programs. Unlike the picture in figure 10, where a physical magnification will also magnify the thickness of the curve, a digitally magnified graph so that, unlike the magnification of a physical picture such as figure 10, the magnified graph can be drawn with the same thickness as the original. Figure 11 shows a picture representing a smart phone with a standard resolution in the square on the left and a small square centred on a point P with coordinates (x, y) has been selected and its contents magnified to fill the square on the right. As the magnification is increased, if the graph has a derivative at P, the magnified portion will become less curved until it looks 'locally straight'. What is important here is not the particular software, but the imaginative idea that, as the point x is moved to the left or right, then the slope of the graph changes and it is possible to look along the graph to *see* its changing slope.

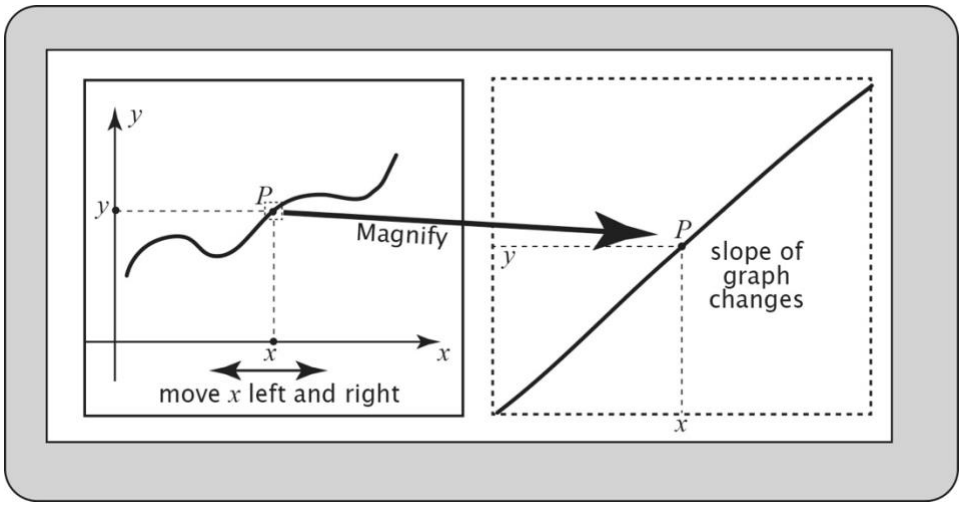


Figure 11: Under high magnification the graph of a differentiable function is locally straight

Software can be designed to plot the numerical value of $(f(x+h)–f(x))/h$ for small fixed h to give the *practical slope function*, which, for a differentiable function, for sufficiently small h will stabilise on the *theoretical slope function* – the derivative $f'(x)$.

Figure 12 shows the practical slope functions for sin(x) and cos(x), visually revealing the practical slope functions stabilising on cos(x) and –sin(x) respectively. Now we can *see* why the derivative of cos(x) is *minus* sin(x): it is the graph of sin(x) *upside-down*.

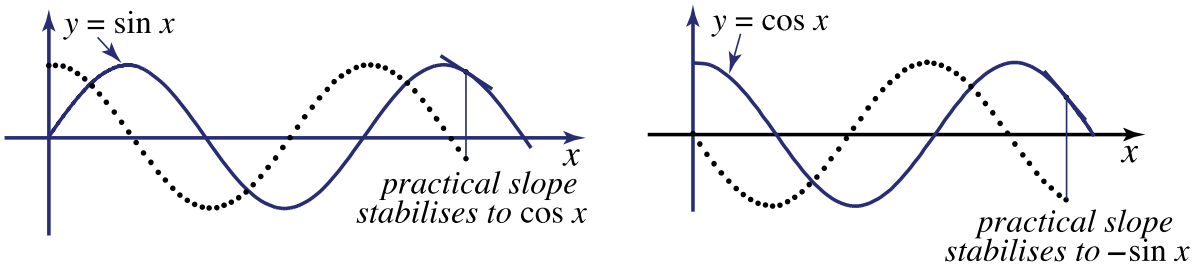


Figure 12: Seeing the practical slope functions of sin(x) and cos(x)

This approach can be used to visualise the derivatives for all the standard functions in the calculus in a practical and theoretical approach to the calculus. It also provides the foundations for increasingly sophisticated ideas in formal mathematical analysis and logical non-standard analysis (Tall, 2013, chapter 11, Stewart & Tall, 2014, 2018).

3.6.2 Integration: a continuous function pulls 'locally flat' when stretched horizontally

The theory of integration to calculate the area under a graph $y = f(x)$ from $x = a$ to $x = b$ involves taking thin strips width dx (which may vary) and height $f(x)$ for successive values of x in the strip and add them all together. The sum may be written as $\Sigma_a^b f(x)\,dx$ or, more compactly, as $\Sigma\, y\, dx$ when the context is clear. Note that the symbol dx now simply refers to the actual (variable) width of the strip.

Calculating such a sum can now be performed efficiently using technology. Adding them to give an algebraic formula is far more difficult. For example, if $f(x) = x_n$ for a whole number n, then the derivative is easy to calculate as nx_{n-1}, but the sum $\Sigma_a^b f(x)\,dx$ of strips involves the formula for the sum of nth powers $\Sigma\, r_n$. Even the case $n = 2$ is difficult for beginners and successively higher values of n become too complicated, even for experts.

An important aspect is the meaning of 'continuity'. Informally, a function $y = f(x)$ is said to be continuous if 'the graph can be drawn without taking the pencil off the paper.' When we

draw a continuous function, some functions, such as $y = x^3$ from $x = -5$ to 5, grow so large (in this case from $y = -125$ to $y = 125$) which would be too tall to draw on the page of a book. Such graphs are often drawn with different x and y scales. What is rarely done is to stretch the picture horizontally to stretch a thin x-interval to fill the width of a picture while maintaining the same vertical scale. This proves to be the representation that links the informal version of continuity to the formal definition.

Figure 13 shows a graph drawn with a continuous movement on the left in a fixed rectangle and on the right the graph is stretched horizontally while maintaining the vertical scale. Imagine the graph being drawn in a fixed window on a high resolution display.

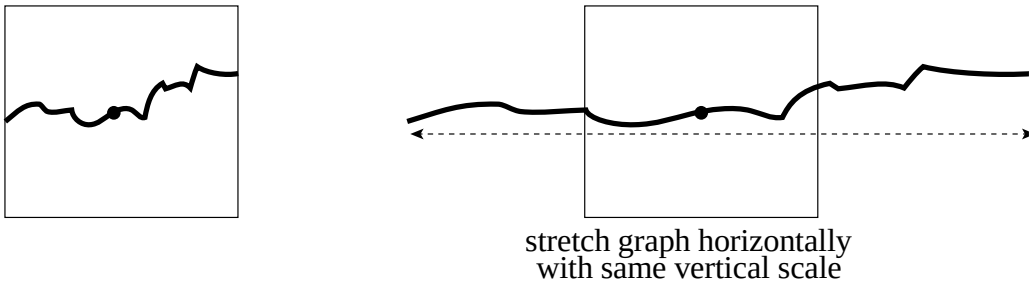


Figure 13: Pulling a continuous graph flat within a fixed window

The graph is said to 'pull flat' if it is stretched out to look like a horizontal line of pixels. This links directly to the formal notion of continuity. Suppose that the point x_0 is in the middle horizontally and the line of pixels is height $f(x_0) \pm \varepsilon$, then for the graph to 'pull flat' to lie in the line of pixels, we need to be able to find $\delta > 0$ such that if x lies in the interval between $x_0 \pm \delta$, then $f(x)$ lies between $f(x_0) \pm \varepsilon$ (Figure 14). This links to the formal definition of continuity in the form that a function is continuous at x_0 if, for given any $\varepsilon > 0$, there can be found a $\delta > 0$ such that $|f(x) - f(x_0)| < \varepsilon$.

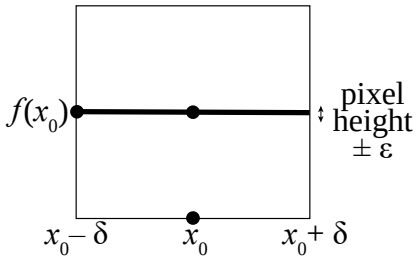


Figure 14: A continuous graph pulled flat

To represent the dynamic process of stretching a graph horizontally, it is possible to program two boxes side by side to allow the user to select a thin vertical strip, height y, width dx above a point x on the horizontal axis in the left box and then picture it stretched horizontally in the box on the right, as in figure 15.

The exact area under the graph $y = f(x)$ from a to b is denoted by A and the exact change in area from x to $x + dx$ as dA. Stretching the graph horizontally changes the visual appearance, but the numerical calculation of the rectangular area y times dx remains the same. If the pixels covered in drawing the graph represent a height h, the numerical difference between dA and $y\,dx$ is less than $h\,dx$. Adding together the rectangular strips, the practical area calculation $\sum_a^b y\,dx$ differs from the theoretical area by less than $\sum_a^b h\,dx = h(b - a)$. If this can be carried out for any value of $h > 0$, however small, then the process of calculating the practical area $\sum_a^b y\,dx$ can be made as close as is desired to the limit object, the theoretical area, $\int_a^b y\,dx$.

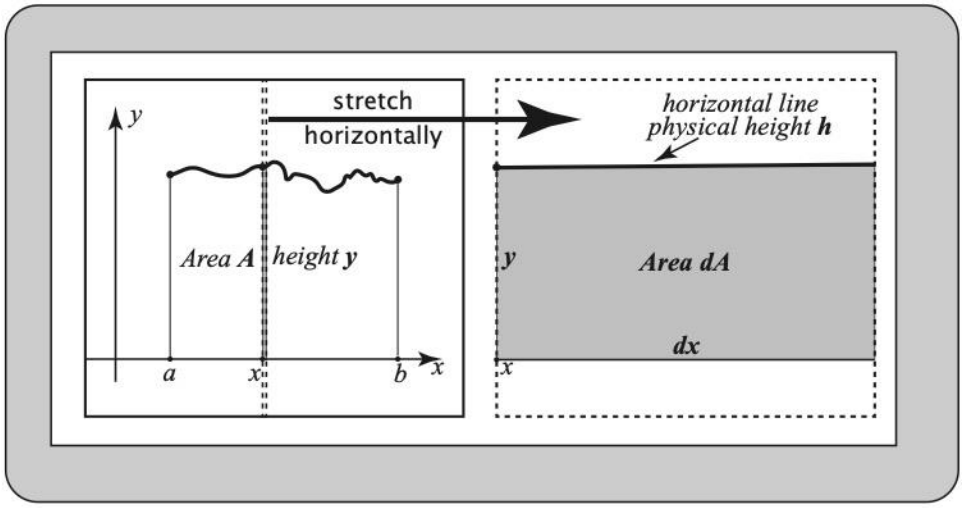


Figure 15: When stretched horizontally, the graph of a continuous function 'pulls flat'

3.6.3 The Fundamental Theorem

We are now in a position to visualise the Fundamental Theorem of Calculus. In the picture of the horizontally stretched strip, the difference between the actual area *dA* and the calculated value of *y dx* lies in the horizontal line width *dx*, height *h*. The error between *dA* and *y dx* is therefore less than *h dx*. The total error between the precise area and $\Sigma\, y\, dx$ is therefore less than $(b - a)\times h$. As we draw more accurate pictures in our mind's eye, we can imagine *h* being as small as desired, so the difference between the finite pratical sum $\Sigma\, y\, dx$ and the value of the area *A* can be made as small as desired. Denoting the theoretical limit value *A* as $\int y\, dx$, it is possible to think of the integral as the precise area where *dx* as an arbitrarily small variable.

Problem solved!

Or is it?

4. Competing views of different communities

Now we reach the great impasse. Can we *really* think of *dx* in the integral as an infinitesimal? Since the real numbers have been formulated as a 'complete ordered field' which can be proved to not contain any infinitesimals, the visual number line has widely been seen as 'complete' with its rational and irrational numbers. Many pure mathematicians researching mathematical analysis denied the existence of infinitesimals on the number line while applied mathematicians usually think of them pragmatically as 'arbitrarily small quantities.' So, two apparently conflicting theoretical approaches continue side by side.

The reason for this can be explained in terms of beliefs that arise in different communities of practice, each of which works coherently in its own community but is unacceptable to the other. This happens, for example, in religion where one community regards the beliefs of another to be unacceptable and, if an individual moves from one religion to another, the first community may consider it to be a *transgression* while the second regard it as an *enlightenment* Tall, 2019a).

This idea of shifting contexts also applies in other situations, including mathematics. It occurs in van Hiele's theory of levels where an expert may seek to offer enlightenment to a student but they may use language in different ways so that each one fails to understand the thinking of the other.

It also happens when an individual is faced with a shift to a new context where the transition is problematic and the individual is not able to make a meaningful transition. Over the longer term the individual may seek to learn the new material by rote to pass an exam but

does not build connected mental structures that are compressed into flexible concepts appropriate for longer-term evolution of ideas.

As new contexts are encountered, some concepts (such as the theory of prime numbers for whole numbers) may not be relevant in another context (such as the arithmetic of real numbers): or may even be inappropriate (as in algebraic number theory where factorisation may not be unique).

To address the relationship between different communities and different contexts requires a higher-level *multi-contextual overview* (Tall 2019a, 2019b). In a religious context this may involve a dialogue between different faiths to identify aspects that they have in common and others in which they differ. A resolution may be found by a multi-faith collaboration in which different religions share those aspects that they have in common and agree to differ on conflicting aspects that that each community holds sacred.

In the historical evolution of mathematics, conflicts may arise as new contexts are encountered and different communities hold their own views for their own purposes. The notion of infinitesimal is a classic case. In the early twentieth century, infinitesimals were banned from conventional standard analysis because they did not fit into the system of real numbers and their use was deprecated, yet they work in simple and meaningful ways in many applications.

This has been a recurring mantra over the centuries. Negative numbers cannot exist, 'because you cannot have less than zero', complex numbers cannot exist 'because the product of two non-zero (real) numbers must be positive, so $\sqrt{-1}$ must be 'imaginary'. In his critique of the calculus, Bishop Berkeley (1734) denied the existence of infinitesimals, saying it is 'impossible to understand them in any sense whatsoever.'

Axiomatic formal mathematics takes us into a new contextual level where it is possible to imagine the number line has more points on it than just real numbers. The practical real line that we see with our human eyes is a continuum which needs to be imagined theoretically to think of it consisting of an infinite number of rational and irrational points. A further leap of imagination is required to be able to conceptualise infinitesimal quantities in analysis. This was proposed using higher level logic by Abraham Robinson (1966). Although some pure mathematicians hailed this as a magnificent insight, many others in classical analysis considered it as a transgression and retained their long-established beliefs..

When I developed a course on the development of mathematical thinking for pure mathematics undergraduates in the early seventies, I translated Robinson's logical approach into algebraic set theory, beginning with simple examples of ordered fields that contained the real numbers as a subfield and proved a theorem that was so simple I didn't have the courage to submit it to any mathematical journal. Frankly, it could be set as an undergraduate exercise for anyone who grasped the axiomatic formal approach.

The theorem applies to any ordered field K which contains the real numbers as an ordered subfield. If K extends the real numbers, it must contain at least one element k that is not real and is either 'finite', in the sense that $a < k < b$ for some real numbers a, b, or positive infinite (satisfying $k > c$ for every real number c), or negative infinite (satisfying $k < c$ for every real c). The completeness of the real numbers can then be used to prove

> Any finite element x in an ordered extension field K of the real numbers is either a real number or of the form $x = c + \varepsilon$ where c is real and ε is infinitesimal.

This is a *structure theorem* which proves new forms of conceptual embodiment and operational symbolism. The conceptual embodiment can be seen in the form of a map $m(x) = (x–c)/\varepsilon$ which magnifies infinitesimal detail near c and draws a practical picture for finite $m(x)$. For those values of x for which $m(x)$ is finite, the *optical microscope* μ is defined to have real values by

defining $\mu(x)$ to be the unique real number that differs from $m(x)$ by an infinitesimal. (See details in Tall, 2013, chapter 13.) This works quite naturally in higher dimensions and, using an optical microscope to look at infinitesimal detail around a point $(x, f(x))$ on the graph of a differentiable function, the image seen in the optical microscope is a full line of slope $f'(x)$ (Figure 14).

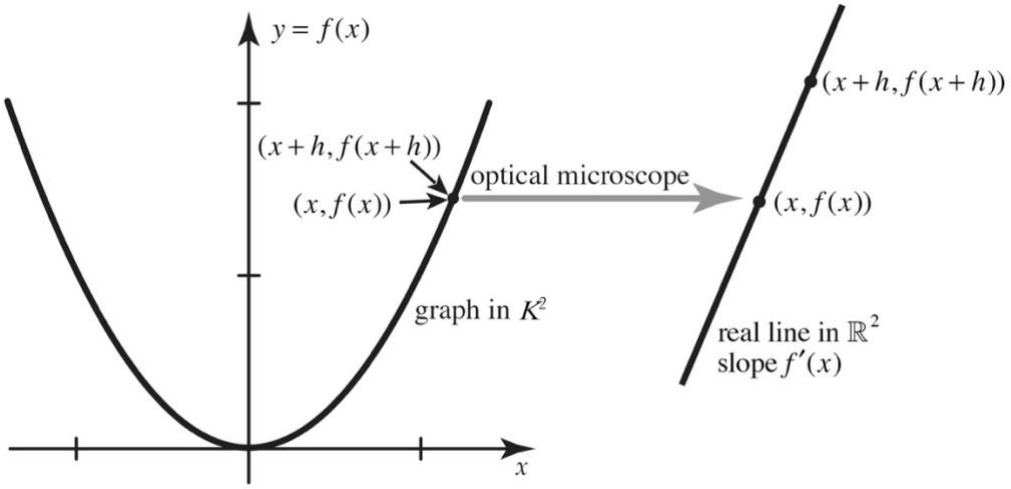


Figure 16: An optical microscope pointing at an infinitesimal part of a locally straight graph

The same technique works with different scales on the x and y axes to visualise the notion of local flatness of continuous functions and applies more generally to multiple dimensions, differential equations, partial derivatives, complex functions and other aspects of mathematical analysis (Tall, 2013, Stewart & Tall, 2014, 2018).

In this way, the theory of three worlds of mathematics does not end with axiomatic formal mathematics at the highest level: structure theorems rise up from the formalism to give even more sophisticated forms of conceptual embodiment and operational symbolism. The three worlds of mathematics spiral upwards together as embodiment and symbolism interact with each other to inspire axiomatic formal theory which in turn leads to more sophisticated embodiment and symbolism through proving structure theorems.

Since only a tiny percentage of the population will spend their lives studying axiomatic formal mathematics, the consequence of this insight is not to encourage the population to fly higher in the stratosphere. Such higher levels of mathematics may be of benefit to society as a whole (such as the use of large prime numbers to encrypt information on the internet). The message here is that different communities may have genuine reasons to think mathematically in different ways that are appropriate both for individuals within that community and for society as a whole.

5. Discussion

The framework of three worlds of mathematics offers a multi-contextual overview of the long-term evolution of mathematical thinking which builds from the past and develops into the future. Different readers may interpret it in different ways. It is the privilege of an individual to form a personal opinion. However, we also have responsibilities to others. The question for the reader to consider is whether the framework encourages *you* to reflect on the attitudes of others and the nature of your students' sense-making in a way that is supportive for their (and your) long-term thinking.

The framework offers immediate ways to re-think how we as individuals make sense of mathematics and how we can help others to progress in their mathematical thinking. The Principle of Articulation can be introduced at *any* level to open up discussion about the meaning

of mathematical expressions. It can be used with young children or with adults with learning difficulties. It can be helpful to teachers to encourage long-term strategies to grasp the meanings of operations in arithmetic and algebra and it can help experts and curriculum designers to plan for long-term success. By using the Articulation Principle and the long-term principles for the operations of arithmetic, the underlying principles of algebra arise naturally and offer a supportive basis to build confidence throughout the whole curriculum.

Many of the problems raised in the literature can be linked to the interpretation of symbolic expressions. For example, children first encounter simple operations such as $2+3 = 5$ where there is a process on the left giving a result on the right. In algebra an equation such as $2x+1 = 7$ with a process on the left and a number on the right can be 'undone' by reversing the process. Meanwhile an equation with expressions on both sides is better understood as having an object on either side expressed in different ways and is solved by 'doing the same thing to both sides.'

Over the long-term, there are successive transitions as different processes give the same object: counting a set in any way gives the same number, equivalent fractions become a single rational number, algebraic equivalences become the same function, equivalent Cauchy sequences give the same real number. While this is usually interpreted symbolically using the concept of equivalence, it is more meaningfully sensed as different symbolic representations of the same object. Visually, equivalent fractions represent the same point on the number line, algebraically equivalent expressions and trigonometric identities give the same graph, and, more generally, infinite limiting processes stabilise visibly on their limit object.

This may be approached by building from natural embodiment to symbolism by encountering several different experiences from which generalities grow (as proposed insightfully by Dienes, 1960). But a more powerful alternative is to identify foundational supportive principles so that learners can use them to build confidence to address the problematic aspects that impede the transition to new contexts.

A serious problem is a fragmentation of the whole system into dealing with learning by developing expertise in separate parts of the whole: pre-school, early learning, kindergarten, primary, secondary, high school, college, adult learning, university, post-graduate, special needs, gifted and talented, and so on. All of these are essential, but they need to be seen as part of a greater whole, so that different communities of practice are aware of a bigger picture. What happens currently is that learning is broken into stages, with tests to decide who passes on from one stage to another.

Personally, I always enjoyed examinations because they incentivised me to reflect on what I had learnt and helped me put everything in perspective. But this is not an experience shared by many.

Instead there may be a desire to pass the examination by rote learning, especially when there are problematic aspects involving a change in meaning. Over the longer term, cumulative changes that occur without making meaningful connections are likely to make mathematics more complicated. Long-term success may be enhanced by meaningful connections that compress complex operations into mental objects that can be manipulated in simpler ways in more sophisticated situations.

5.1 Evolving practice from theory

The book *How Humans Learn to Think Mathematically* (Tall, 2013) builds theory from practice through analysing the development of mathematical thinking of individuals from pre-school beginnings to post-graduate research. Since that publication, the theory has expanded to incorporate a range of insights reported in this paper. Of particular value are experiences that

everyone can observe for themselves and simple principles that can be introduced explicitly to learners to enhance long-term growth. They open up a new stage in the relationship between theory and practice. Instead of using research to develop theory from practice, *we may reverse the direction and use theory to develop practice, based on meaningful principles*.

For example, the **Articulation Principle** enables us to give precise meanings to symbolic expressions. For instance, $2 \times \ldots 3 + 4$ may be written as $2 \times (3 + 4)$ where $3 + 4$ is to be calculated first give 2×7. More generally, a sub-expression inside a bracket may be seen as a single entity to be calculated first and given the highest order of precedence.

The **General Principle for Addition and Subtraction** underpins the rule that performing these operations on a list of quantities in different ways does not change the result, so Addition and Subtraction have the same order of precedence.

Likewise, the **General Principle for Multiplication and Division** underpins the rule that Addition and Subtraction have the same order of precedence.

It is now a matter of seeing some operations bound together more strongly than others. Brackets (Parentheses) have the highest order of precedence. Exponents such as x^3 take precedence over multiplication, so that $2x^3$ is seen as $2(x^3)$ rather than $(2x)^3$. Over time the learner becomes acquainted with simpler expressions involving a small number of terms, such as $a + 2b$ where the implicit multiplication is bound together as $(2 \times b)$ and the terms a and $(2 \times b)$ can be re-ordered as $(2 \times b) + a$. More generally, the interpretation of expressions can be made meaningful by encouraging the learner (and the teacher) to imagine the strength of binding of operations between quantities and to deal with stronger bound sub-expressions first. Meanwhile any subsequence of operations having the same order of precedence can be moved around and performed in any order.

These general principles now give a precise *meaning* to the order of operations in the mnemonic

P > E > M = D > A = S.

As shown in §3.4, this meaning extends to more sophisticated symbolic expressions throughout mathematics, giving meaning to the long-term development of symbolism.

The enhanced framework also emphasises the fundamental role of embodiment to give human meaning to the powerful use of symbolic operations in increasingly sophisticated theory. In addition, it acknowledges the underlying emotional linkages between supportive and problematic aspects that cause not only pleasure or fear but also our very ability to make, or fail to make, connections in mathematical thinking. It enhances making mathematics meaningful over the longer term.

5.2 Evolving mathematical thinking in the present and future

The enhanced framework observes that different communities interpret mathematics in different ways that may be appropriate in their own context but may not apply in others. It formulates a broader overview to encourage individuals in different communities to communicate with each other, to respect viewpoints that may be appropriate for others but may differ from their own. This is not a simple task, as each community will have ways of working that they share amongst themselves but which may not be shared by others.

In writing up the developing theory as it evolved, I found that I was expressing ideas in different ways for different audiences. These included elementary school teachers (Tall, 2017), a conference on ‘mathematical transgressions’ for philosophers of mathematics and educators (2019a), cognitive science (2019b), the ‘Aha!’ experience (2020a), and the relationship between university mathematics and mathematics education (2020b). Reviews of these papers were encouraging, but there is a need for reflective research into the effectiveness of the

practical approach to the theory in different contexts. This is no longer a simple matter as different communities may view the situation in different ways and come to different conclusions. What is of paramount importance is that the teaching and learning in a particular community is fit for the purpose of that community.

A major ingredient is the Articulation Principle that can be used to give meaning to operational symbolism at any level from young children encountering simple arithmetic to adults who may have deeply problematic difficulties with arithmetic and algebra. A further aspect is the realisation that embodiment, in terms of gesture, dynamic visualisation and mental thought experiment, plays an important role in the long-term development of sophistication, even at the highest formal level where structure theorems pave the way to support human thought processes to imagine new levels of thinking. This includes the way in which we imagine constant and variable quantities that allow us to *see* arbitrarily small quantities as infinitesimals that can be magnified to see them in a finite magnification. It extends to the visualisation of concepts such as continuity, differentiation, integration, passing through levels of practical, theoretical and formal mathematics.

This has serious consequences that can affect the whole curriculum, such as calculus in the United States where the College Board (2016) specifies a curriculum that can be tested without mentioning visual ideas such as local straightness of differentiable functions and local flatness of continuous functions. Yet the *MAA National Study of College Calculus* (Bressoud et al., 2015) reports serious difficulties with the calculus that may be meaningfully resolved by taking account of their dynamic visual representations coupled with the flexible interpretations of the symbolism.

In my most recently completed publication at the time of writing (Tall, 2020b) I proposed an overall principle for long-term meaningful learning which I named in honour of my 11-year old grandson who explained the idea of the Articulation Principle to me (Tall, Tall & Tall, 2017):

> **The Simon Principle**: The teacher should be 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.

This requires a total rethink of the entire curriculum so that fundamental principles are fully integrated into the experience of teachers and learners. It also suggests that mathematicians and curriculum designers should reflect on how to make mathematics meaningful over the long-term in a manner appropriate both for the developing individual and the needs of society. This is likely to involve mathematicians, mathematics educators and others involved in the use of mathematics to realise that their belief systems need to be radically overhauled to make sense of the present and to prepare for the future.

References

Amalric, M. & Dehaene, S. (2016), Origins of the brain networks for advanced mathematics in expert mathematicians, *Proceedings of the National Academy of Science of the US*A, 113 (18), 4909-4917, https://doi.org/10.e1073/pnas.1603205113

Berkeley, G. (1734), *The Analys*t (ed. D. R. Wilkins, 2002). Retrieved from www.maths.tcd.ie/pub/HistMath/People/Berkeley/Analyst/Analyst.pdf, January 29, 2019.

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

Bruner, J. S. (1966). *Towards a Theory of Instruction*, Cambridge, Mass: Harvard University Press.

Chomsky, N. (2006). *Language and Mind*. Cambridge: Cambridge University Press.

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

Corballis, M. C. (2014). Left Brain, Right Brain: Facts and Fantasies. *PLoS Biol* 12(1): e1001767. https://doi.org/10.1371/journal.pbio.1001767

Cuisenaire, G. (1952). *Les Nombres en Couleurs*. https://en.wikipedia.org/wiki/Cuisenaire_rods.

Dienes, Z. P. (1960) *Building up Mathematics*. London: Hutchinson.

Durell, C.V. & Robson, A. (1933). *Elementary Calculus I*. London: G. Bell & Sons.

Gattegno, C. (1971). *Geoboard Geometry*. New York: Educational Solutions Worldwide.

Gutiérrez, A., Jaime, A., Fortuny, J. (1991). An alternative paradigm to evaluate the acquisition of the Van Hiele levels, *Journal for Research in Mathematics Education*, 22 (3), 237–251.

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

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.

Kidron, I. & Tall, D. O. (2015). The roles of embodiment and symbolism in the potential and actual infinity of the limit process. *Educational Studies in Mathematics* 88: 183. doi:10.1007/s10649-014-9567-x.

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

Leibniz, G. W (1684). Nova methodus pro maximis et minimis, *Acta Eruditorum, October 1684*, 467-473.

Maruyama, M., Pallier, C,. Jobert, A., Sigman, M., Dehaene, S. (2012). The cortical representation of simple mathematical expressions. *Neuroimage* 61(4):1444–1460.

Robinson, A. (1966). *Non-Standard Analysis*. Amsterdam: North Holland.

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

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

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

Tall, D. O. (2017) Making sense of elementary arithmetic and algebra for long-term success. Draft chapter for Japanese Elementary School Teachers.

Tall, D. O. (2019a). 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. (2019b). From Biological Brain to Mathematical Mind: The Long-term Evolution of Mathematical Thinking. To appear in Marcel Danesi (ed.), *Interdisciplinary Approaches to Mathematical Thinking*. Springer

Tall, D. O. (2020a). Building Long-term Meaning in Mathematical Thinking: Aha! and Uh-Huh! To appear in Bronislaw Czarnocha and William Baker (Eds): *Creativity of Aha! Moment and Mathematics Education*. Sense Publications.

Tall, D. O. (2020b) Long-term principles for meaningful teaching and learning of mathematics. To appear in Sepideh Stewart (ed.): *Mathematicians' Reflections on Teaching: A Symbiosis with Mathematics Education Theories*. Springer.

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.

van Hiele, P. M. (2002). Similarities and Differences between the Theory of Learning and Teaching of Skemp and the Van Hiele Levels of Thinking. In Tall, D. O. & Thomas, M. O. J. (Eds), *Intelligence, Learning and Understanding – A Tribute to Richard Skemp* (pp. 27–47). Post Pressed, Flaxton, Australia.

Woodhouse, R. (1803). *The Principles of Analytical Calculation*. Cambridge: Cambridge University Press.