A Sensible approach to the Calculus

David Tall
University of Warwick
Coventry CV4 7AL
United Kingdom
<david.tall@warwick.ac.uk>

In recent years, reform calculus has used the computer to show dynamic visual graphics and to offer previously unimaginable power of numeric and symbolic computation. Yet the available technology has far greater potential to allow students (and mathematicians) to make sense of the ideas.

***A sensible approach to the calculus** builds on the evidence of our human senses and uses these insights as a meaningful basis for various later developments, from practical calculus for applications to theoretical developments in mathematical analysis and even to a logical approach in using infinitesimals. Its major advantage is that it need not be based initially on concepts known to cause student difficulty, but allows fundamental ideas of the calculus to develop naturally from sensible origins, in such a way as to make sense in its own right for general purposes, support the intuitions necessary for applications, provide a meaning for the limit concept to be used later in standard analysis and further, to provide a sensible basis for infinitesimal concepts in non-standard analysis.*

Introduction

This presentation reveals my current thinking on the nature of calculus, based on the ways that we humans naturally think about the ideas. In particular it considers how we develop through our perceptions, operations and use of language to formulate increasingly sophisticated ideas. I suggest that this involves three distinct forms of mathematical thought, one growing from our natural perceptions, one from the actions that we perform and translate into symbolic computation and manipulation, and one in which we formulate logical definitions and develop the structures of formal proof. (Tall, 2004, 2008),

This is part of an ongoing development that I began in thinking about the calculus over 35 years ago (Tall, 1975) and, while some ideas are long established, other significant advances are presented here for the first time. These involve an analysis of how our ideas depend on our previous experience within a global theory of cognitive development from early childhood to research mathematics. This affects not only students who are learning analysis, but also we 'experts' who view mathematics from our own viewpoint which we may share with our particular expert community. We need to clarify precisely what we would desire students to learn and the development that is possible for students in our current technological age.

Culturally, the calculus is the product of thousands of years of evolution that have shaped its current form. This includes the early methods of the Greeks to compare areas and volumes, and their development through the 'prime and ultimate ratios' of

Newton and the infinitesimals of Leibniz, on to the formal epsilon-delta definitions and proofs of mathematical analysis. Various constructs have changed meaning over time, for example *dy*/*dx* originally meant a quotient of lengths to Leibniz, but now it is usually re-interpreted as a limit that makes the meaning more subtle. Here I will return to the idea of *dy*/*dx* as a quotient of the components of the tangent vector.

The limit concept has proved to be an excellent foundation for mathematical analysis at the highest level. However, we now know from our own experience and many research studies that it is a source of cognitive difficulties for students. My quest is for a 'sensible approach' to the calculus that begins in naturally perceived phenomena and flowers into a knowledge structure of great power in applications and also provides a meaningful foundation for more subtle mathematical developments at a later stage.

As a schoolboy I absolutely adored the beautiful book on *Elementary Calculus* by Durell and Robson (1934), working assiduously through its visionary presentation, doing every problem over a three-month period and finishing it triumphantly at 3pm on Christmas Day 1956: it was probably the best Christmas present I ever had. In contrast, as an undergraduate at Oxford in 1960, I struggled with Mathematical Analysis, initially finding it almost impenetrable, though on reviewing my notes it began to make sense and I scored the highest mark on the analysis paper of all Oxford mathematics students in my year. If I had found it so difficult, what had happened to everyone else? The contrast between the two experiences was dramatic. How could calculus give me so much joy as a boy in school when analysis was so problematic for me as the highest scoring student in a highly prestigious university?

Subsequently, as a mathematician I appreciated the power of the limit concept and the precise formal theorems that could be deduced from it. Later, as a mathematics educator, I lamented the loss of the natural beauty of the ideas of calculus that had given me so much personal joy.

My objective is not to produce a watered-down version of mathematical analysis 'made simple' for students who struggle. As a mathematician I seek to develop fully functional mathematical thinking, including precise mathematical definitions and proof. As an educator, I consider it essential to present the ideas in a sequence that makes sense to students, including those who study the subject for its use in applications without any desire to follow it into more advance pure mathematical studies.

This does not mean looking at mathematics from the viewpoint of an expert (which the learner is as yet unlikely to share) and 'talking down' the ideas in an 'intuitive' way. (This happens all to often in calculus books where mathematicians profess to offer an 'intuitive' viewpoint of continuity and limits.) My quest is to seek a 'built up' viewpoint, carefully designed to reach the subtleties of mathematics from the viewpoint of the learner. To do this requires more than mathematics alone and more than the viewpoint of the learner, it requires a complementary blending of both.

The reform of calculus teaching has been considered around the world for many years now. However, after reform projects have attempted a range of different

approaches using technology, what has occurred is largely a retention of traditional calculus ideas now supported by dynamic graphics for illustration and symbolic manipulation for computation. In this presentation I consider the theory and practice of a sensible approach to the calculus which builds on the natural viewpoint of the student and offers a conceptual foundation for more sophisticated development.

Where do we begin?

The first question is to ask where we begin in the quest to blend together mathematics and human development to build a theory of calculus that fits together naturally for the human learner. Mathematicians already have a sophisticated view of the limit concept and its use as a foundation for modern mathematical analysis. The consequence is that the limit concept is often introduced to beginners in terms of intuitive ideas of 'as near as we please' or 'for sufficiently large n'. Meanwhile, other previous experiences, such as the notion of a tangent in Euclidean geometry, give the intuitive idea that a tangent 'touches the curve at one point and does not cross it' in a way that is problematic in the calculus.

To 'make sense' of the concepts of the calculus, including the notion of continuity, limit, tangent, derivative, and so on, we need to consider how we, as individuals, think about these ideas. The first thing to do is for the reader to reflect for a moment and write down what she or he thinks these calculus concepts actually *mean*. Not just their definitions, but how we might describe the meaning of the ideas and their relationships in a way which makes sense *to us*, as individuals, and how these ideas might make sense to a student.

When I present these ideas in a workshop, I invite members of the audience to talk to one another for a time, to write down what they mean to them as individuals and how the ideas are related:

> Function, continous function, limit, tangent, derivative.

It is important at this stage that the reader makes explicit what she or he thinks of these concepts. Please take a few minutes to sketch out your ideas and then you will be able to compare what you have written down with the sensible approach to calculus presented here.

Human perception

In the book *A Mind So Rare*, Merlin Donald (2001) analyses the nature of human consciousness and proposes that consciousness occurs at three levels that he suggests are:

1. *selective binding* to give a thinkable concept (around 1/40th of a second),
2. *short-term awareness* monitoring change (two to three seconds),
3. *extended awareness* over long periods of time using language, symbols, pictures etc to build coherent knowledge structures.

The first two of these relate particularly to the fundamental ideas of the notion of

change and rate of change in the calculus. The operations in the brain take a specific small time to put together and we are not capable of perceiving changes that occur in arbitrarily short periods or arbitrarily small quantities, although we can use extended awareness to imagine them. The perceptual idea of continuity involves the short-term awareness monitoring change. This is where we 'sense' change. It is only through the third level use of extended awareness that we can build up a coherent mathematical knowledge structure for a more formal concept of continuity.

Preliminaries

When students begin to study the calculus, their success depends on their previous experience and current knowledge. This should include the conception of a function defined on a specific domain and giving a specific output $y = f(x)$ for a specific input x in the domain (see, for example, Tall, McGowen & DeMarois, 2000a, 2000b). This is essential to formulate the numerical approximation to the slope as $(f(x+h) - f(x))/h$ and to be able to manipulate such expressions to understand the derivation of the general rules for the calculus. It is also assumed that the students can interpret the graphs of functions such as simple polynomials, rational functions, trigonometric functions and the relationship between powers $a = b^c$ and logarithms $c = \log_b(a)$, as appropriate.

Perceptual Continuity

Figure 1: A perceptually continous graph

The perceptual notion of continuity is based on the idea of drawing a curve with a pencil in a stroke of the hand without taking the pencil off the paper. We can see it at level 1 as a whole gestalt which we recognise as being in one piece without any gaps, in a single pencil stroke. At level 2 we can imagine our finger tracing along it over a short period of time. We sense this in several ways: visually as we look along the curve, enactively as we draw the curve, or trace along it with a finger, and mentally as we imagine the graph being drawn in our mind's eye.

The question is: how do we formulate this in a way that transforms these 'natural' experiences into the formal definition? The answer lies in stretching it horizontally on a computer screen.

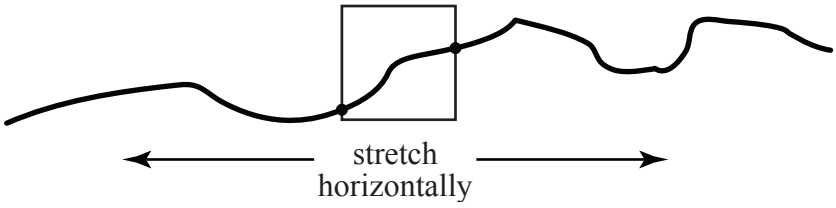


Figure 2: stretching horizontally

The graph will stretch off screen, but the viewer will see only the displayed part.

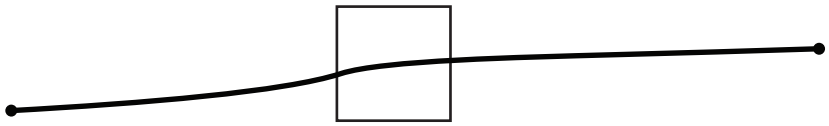


Figure 3: stretched

Then continue stretching the curve, looking only at the part on the screen.

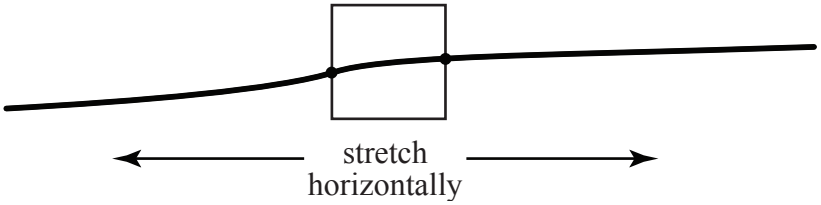


Figure 4: stretch again

Until the visible part on the screen is pulled flat:

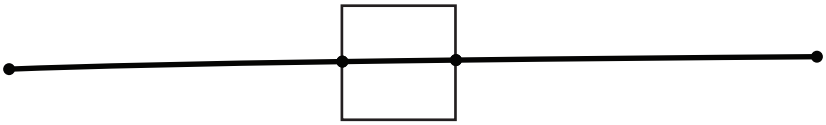


Figure 5: the graph 'pulls flat'

We now say that a graph is 'naturally continuous' if, maintaining the same vertical scale and increasing the horizontal scale, the visible part of the graph in a fixed window eventually pulls flat.

This natural process has a formal counterpart. Imagine a picture of a graph on a high resolution screen, and suppose the middle point $(x_0, f(x_0))$ on the graph lies in a pixel of height $\pm\varepsilon$ in the picture, then to 'pull the graph flat' (that is so that it lies in the horizontal line of pixels) it is necessary to find a value $\delta > 0$ such that whenever x lies between $x_0 - \delta$ to $x_0 + \delta$ then $f(x)$ lies in the horizontal line of pixels between $f(x_0) - \varepsilon$ and $f(x_0) + \varepsilon$. This is precisely the formal epsilon-delta definition in the form: 'Given a pixel height $\pm\varepsilon$, a $\delta > 0$ can be found so that, given x_0, if x lies between $x - \delta$ and $x + \delta$, then $f(x)$ lies in the range $f(x_0) \pm \varepsilon$.'

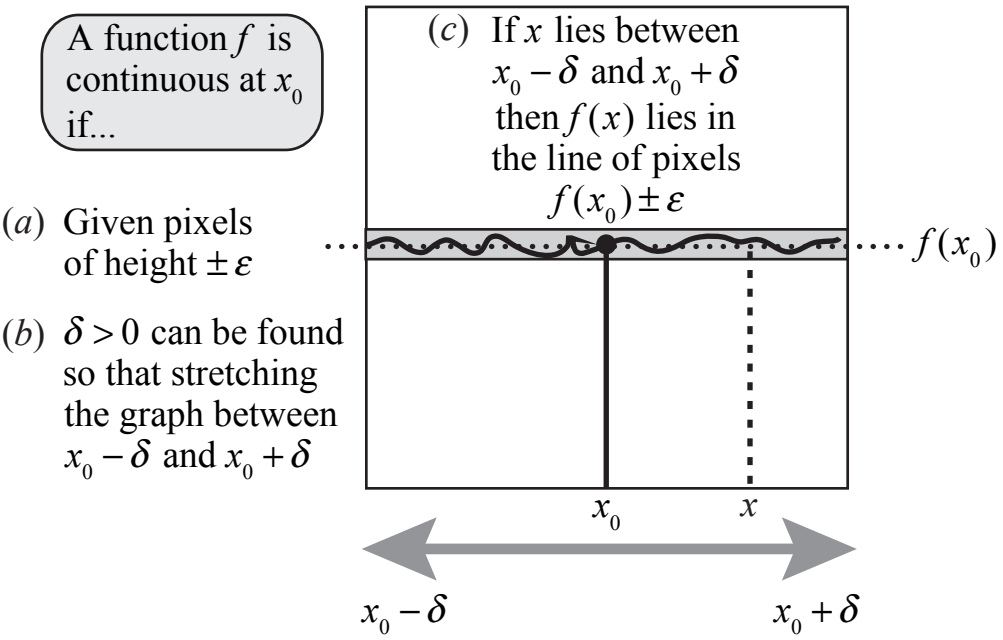


Figure 6: A natural interpretation giving the formal definition

My experience is that serious mathematicians are concerned about the validity of this kind of 'natural' approach. Surely the argument is an intuitive picture that does not give the full force of formal continuity. Does it work for more general cases, such as a function defined only on the rationals? What about weird functions such as

$$f(x) = \begin{cases} 0 \text{ for } x \text{ irrational,} \\ 1/n \text{ if } x = m/n \text{ is rational in lowest terms} \end{cases}$$

This is continuous at all irrational points and discontinuous at every rational.

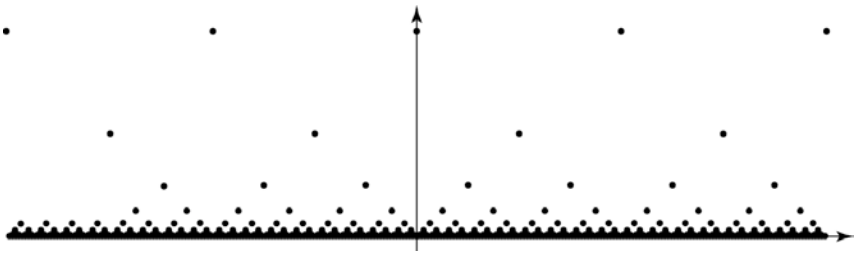


Figure 7: A very discontinuous function that is continuous at every irrational point

Despite the unusual picture, the graph satisfies the definition. It 'pulls flat' for any window centred on an irrational (since for given $\varepsilon > 0$ one can find an interval excluding any rational m/n for which $n > 1/\varepsilon$). However, it does not 'pull flat' for any rational.

Continuous functions defined only on rationals can also have 'gaps', such as

$$f(x) = \begin{cases} 1 \text{ if } x^2 > 2, \\ 0 \text{ otherwise.} \end{cases}$$

This clearly has a disconnected jump either side of $x = \sqrt{2}$, but the graph is not

defined at this point. Everywhere that the function is defined, the graph will pull flat.

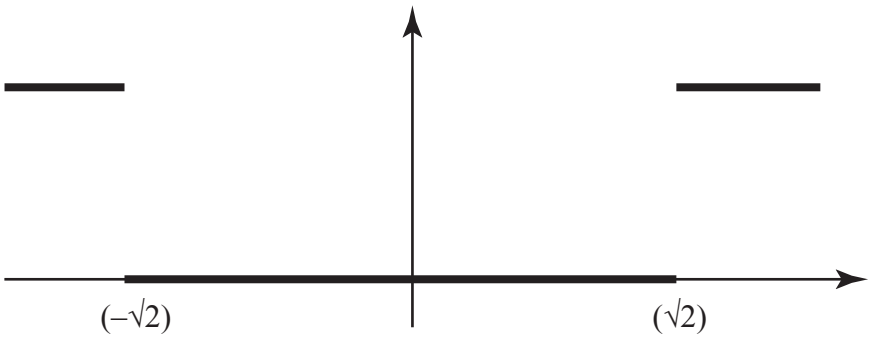


Figure 8; A function continuous on the rationals that has gaps in the graph

The question arises as to whether considerations such as these are relevant for the average student taking elementary calculus. My response is *absolutely not*. They are ideas relating to the nature of mathematical analysis with precise set theoretic ideas of concepts such as formal limit, completeness, connectedness and so on.

For a student starting the calculus, it is natural to draw graphs as curves on paper with a pencil or on a screen with a pixel where points have a finite size. In this case a graph is drawn from some value of $x = a$ and moves smoothly to an endpoint $x = b$. If the graph is drawn over a closed interval $[a,b]$, the physical drawing does not consist only of the abstract points $(x, f(x))$, it *covers* the points with a pencil line of finite thickness.

For a given pencil, choose a value of $\varepsilon > 0$ sufficiently small so that when a point is marked at a point (x, y) it also covers a small square width $x \pm \varepsilon$, height $y \pm \varepsilon$. A formally continuous graph can then be drawn physically as follows. For the given value of ε, find a value δ, such that for any for t in the interval centre x, width δ, the value of $f(t)$ lies in a vertical range with centre $f(x)$ and total height ε. If δ is larger than ε, then replace its value by ε, and then the rectangle of width 2δ, height 2ε will be covered by the mark made by the pencil point. Draw successive rectangles at steps 2δ apart, each with its middle point centered on the graph. Place the pencil point over successive rectangles and drag it along the curve to draw a naturally continous graph. By using a finer pencil and a corresponding value of ε, this can be done for any size pencil, however small it may be.

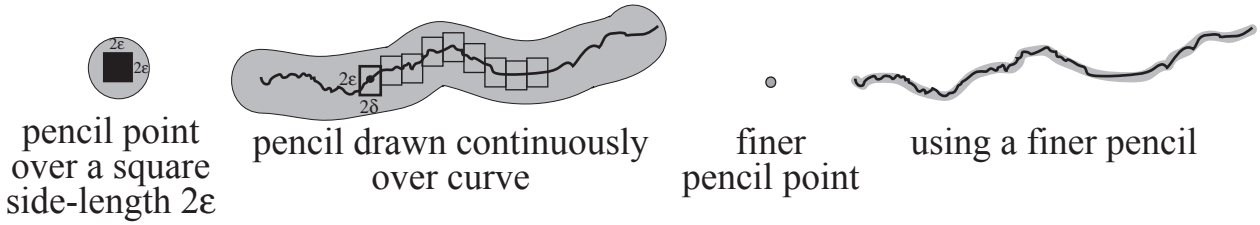


Figure 9: Drawing a continuous graph

These relationships between natural continuity and the formal definition are not part of elementary calculus. Their purpose is to convince the teacher that a natural approach to the calculus can use natural continuity as a basis for a first course in the knowledge that it can provide a sound cognitive basis for later formal analysis.

Perceptual limits

The idea of a sequence of points 'getting close' to a limit point or a sequence of graphs 'getting close' to a limiting graph both create the possibility of cognitive obstacles which cause deeply-held beliefs that are considered difficult to remediate. In particular, the idea of a sequence of numbers tending to a limit often gives a view of a variable quantity that is 'arbitrarily small' so that the number 0.999… is conceived as 'just less than one' rather than precisely equal to one. The symbolic and visual aspects of convergence are here in conflict.

While it is evident that the no term of this particular sequence of decimal approximations is ever equal to the limit, visually if one plots the points physically on a line, then they are soon indistinguishable from the limit. This can be seen more generally in a dynamic picture where a sequence of points a_1, a_2, … tend to a limit a. As they are added successively to a picture, the marked points eventually become indistinguishable from the limit a. When one focuses on the latter points, by successively removing a_1, a_2, … what is left reveals that after a certain stage all the later approximations are indistinguishable to our human eyes from the limit a.

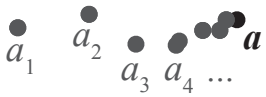


Figure 10: A sequence of points tending to a limit a

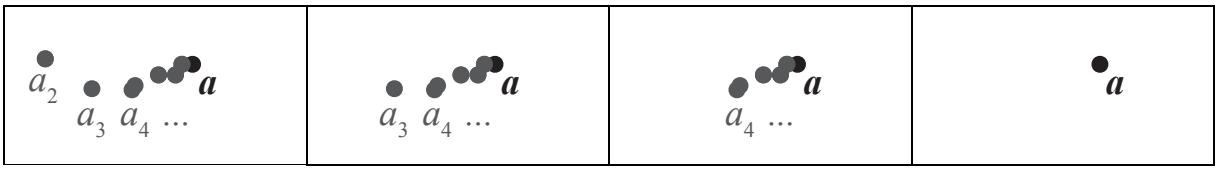


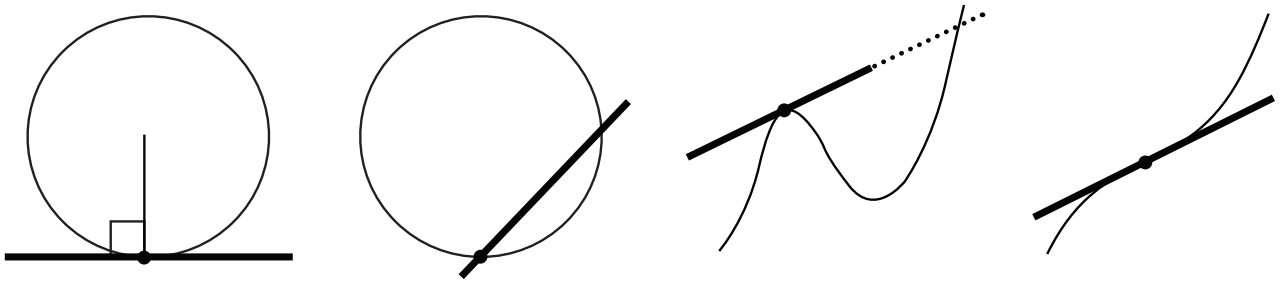
Figure 11: Removing the initial points until the terms are indistinguishable from a

What is important here is to see the limit and then to see the later terms of the sequence become indistinguishable from this limit.

Numerically, beginning calculus students have been operating in a 'good enough' world of arithmetic. The sequence 3.1, 3.14. 3.141, 3.1415, … tends to π in the sense that various approximations, such as 3.14, 3.1412, $^{22}/_{7}$ are 'good enough' to be indistingishable from π in a given practical context.

Perceptual tangents

Our previous experience of tangents in geometry give us specific insights that colour our notion of tangent in calculus. In geometry the tangent to a circle is at right-angles at the end of a radius. It appears to touch the curve precisely once (or esoterically in two 'coincident' points), lies outside the circle, and does not cross it. So what precisely is a tangent in the calculus? Does it touch a curve at a single point and not cross it?



tangent not a tangent Is this a tangent? Or this?

Figure 12: What is a tangent?

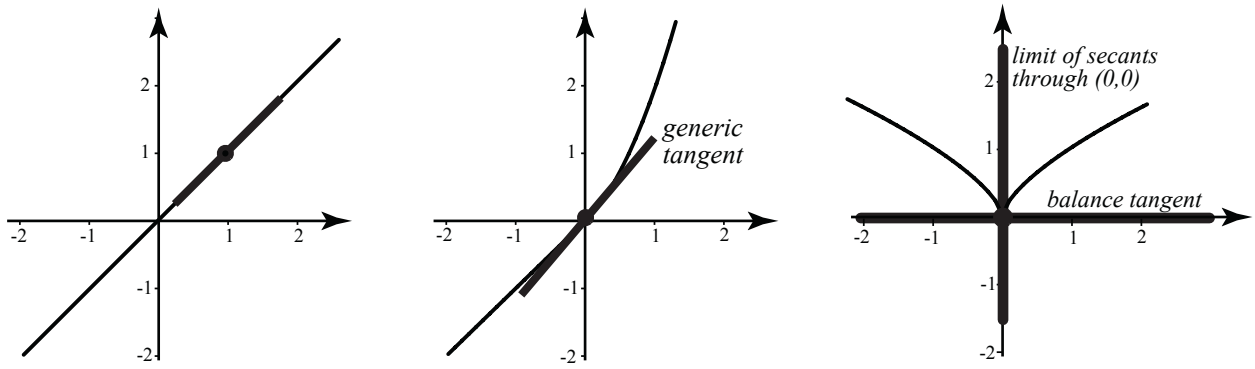
Standard practice is to assume that a beginning calculus student 'knows what a tangent is' and uses this idea to speak of the derivative as the slope of the tangent at a point. Such an approach is fraught with subtle difficulties if a tangent 'touches, but does not cross' the curve at a point.

If we consider various possibilities, we will find this idea causes conceptual difficulties. For instance, the tangent to the graph $y = x$ at a point does not 'touch the graph at a single point'. It is identical with the original graph.

The notion of 'touching at a single point' causes difficulty with a function such as

$$y = \begin{cases} x & (x \le 0) \\ x + x^2 & (x \ge 0) \end{cases}.$$

In this case, the tangent at the origin coincides with the graph to the left, but students asked to draw the tangent at the origin, often draw a 'generic' tangent that touches the graph at only one point (figure 13) by turning it at a slight angle rather than drawing the actual tangent which coincides with the graph to the left of the origin (Vinner, 1982).



$y = x$ $\quad y = \begin{cases} x & (x \le 0) \\ x + x^2 & (x \ge 0) \end{cases}$ $\quad y = x^{2/3}$

Figure 13: How do students 'see' the tangent to a curve at a point on a curve?

Even more problematic is the tangent to $y = x^{2/3}$ at the origin. If one draws a secant through $(0,0)$ and $(x, x^{2/3})$, then as x tends to zero, the limit of the secants is a vertical line. This is often claimed to be the tangent. However, students starting

calculus see other possibilities, perhaps as a horizontal 'balance' tangent that touches and does not cross the curve, or perhaps suggesting many tangents which are lines that go through the origin without cutting through the graph itself.

If one magnifies these pictures at the points concerned, then a significant clarification occurs. The first two graphs magnify until they 'look straight', whereas the third graph is not a straight line segment. When it is sufficiently magnified, it is a half line segment pointing down and going back up (Figure 14).

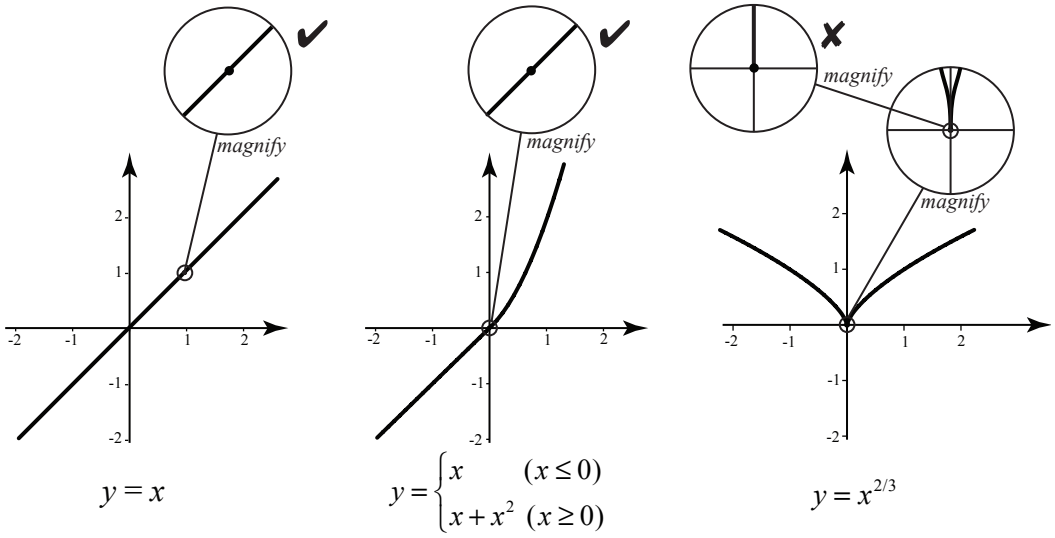


Figure 14: Are these graphs locally straight at the origin?

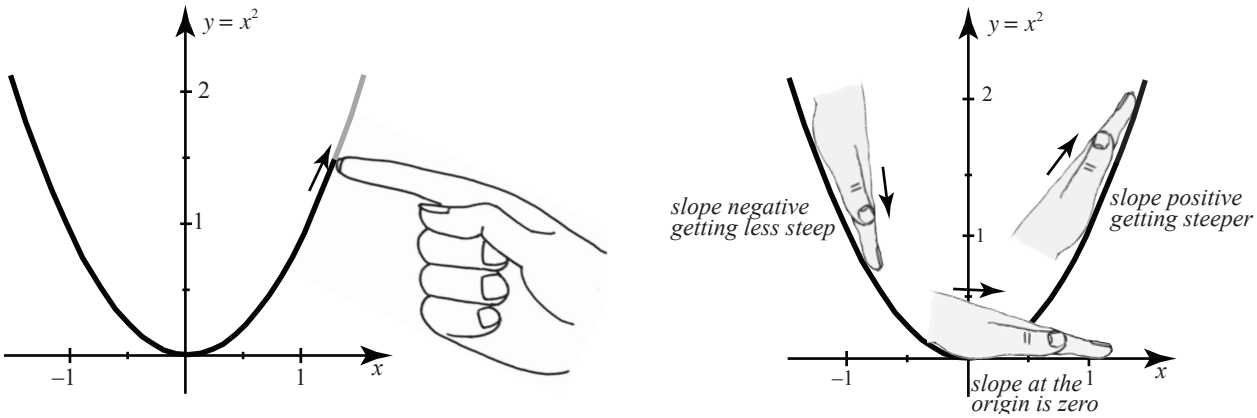
Now, using level two dynamic perception, we can *see* the highly magnified graph looking like a straight line which looks the same as the tangent because the difference between them is covered by a line of pixels, or if drawn with a pencil, by the thickness of a pencil-stroke.

The case of the function $y = x^{3/2}$ requires further clarification. Some choose to declare it to have a vertical tangent at the origin because the limit of secants through the origin is vertical. However, from the left the slope is negative infinite and from the right it is positive infinite, so it has different left and right derivatives.

For consistency, it is appropriate to visualize a graph under high magnification and to make the convention that a graph is only considered to have a unique tangent if it magnifies to look straight when one zooms in on a sufficiently small portion of the curve. This enables us to imagine the natural dynamic continuity of a point moving along a curve so that were it to continue in the direction it has at any time, then it would move in the direction of the tangent. For instance, if a weight at the end of a piece of string is being swung round in a circle and the string is cut, then it will continue in a straight line along the tangent.

Local Straightness

A graph can be *seen* as an object and one may trace a finger along it to *sense* it as an object. Then it is possible to slide a straightened hand along the curve to sense its changing slope as a natural conception.



Tracing a graph to see and feel the graph as an object

Sliding a hand along the graph to sense the changing slope

Figure 15: Tracing a graph with a finger and sensing the change of slope

Looking at a tiny part of the graph it may be possible to *see* the slope of the curve as it changes gently along its length. A simple way to do this is to look at a small part of the curve and place ones fingers over the graph on either side of it to confine one's view to a small portion of the graph. Often this reveals a small portion to 'look straight' without looking too closely.

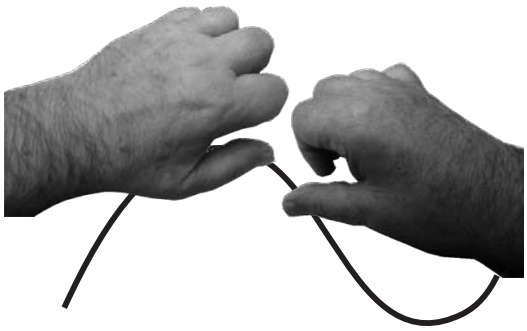
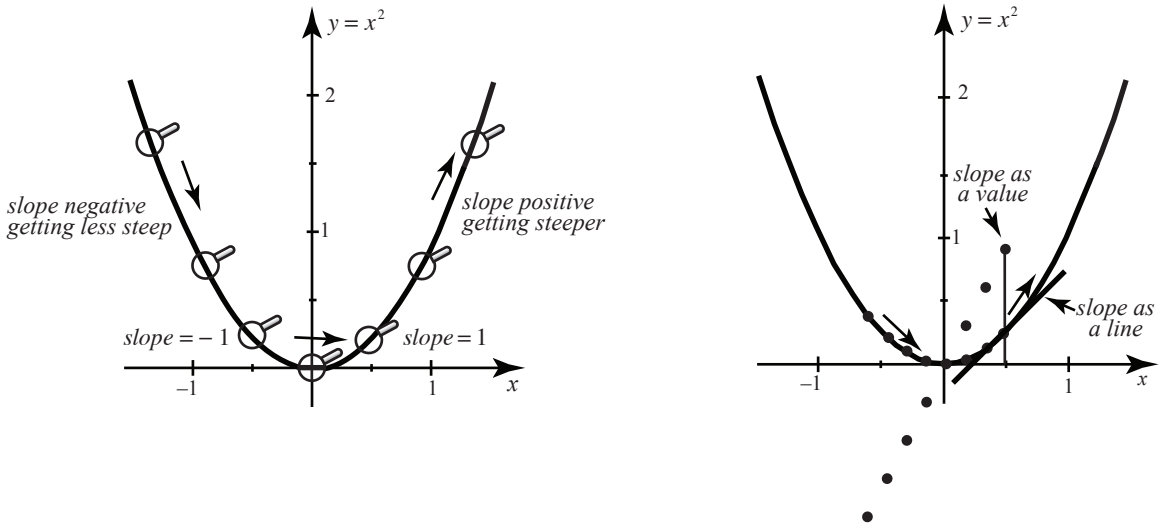


Figure 16: A small part of the curve looks fairly straight

A better method is to imagine using a small magnifying glass to magnify a small portion of the graph. This mental image allows one to 'see' the changing slope as the magnifying glass is moved along.



Moving a magnifying graph along the curve to see the changing slope

Plotting the changing value of the slope on a computer as a new graph

Figure 17: Sensing and seeing the changing slope

Using information technology it is possible to use software to plot the numerical value of the slope as a point. As this happens dynamically, one can see the graph of the slope function for x^2 to stabilize on the graph of $2x$.

The general method, using the idea of local straightness is to begin with some (locally straight) graph $y = f(x)$ and draw the slope function which stabilizes to give a new graph representing the slope of the original. Let us denote the operation by D and denote the slope function as Df (where D stands for 'derivative', namely the slope function *derived* from the original). In this case, for $f(x) = x^2$ we can see that $Df(x) = x^2$.

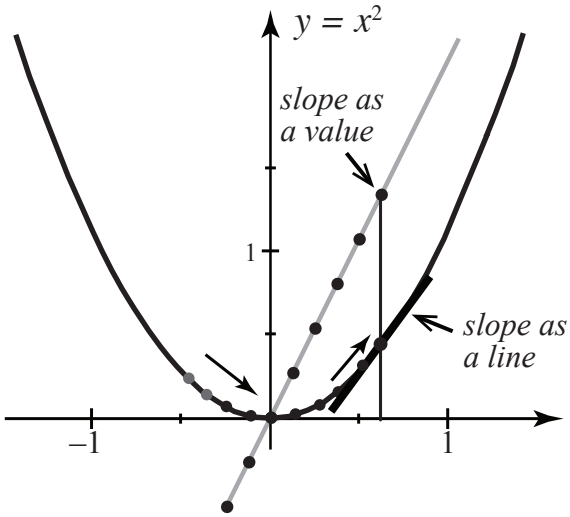
This conception of the derived function originates fundamentally at Donald's level two of short-term awareness. Our task is now to set it in a wider sense of global awareness as a theoretical concept specified symbolically and formally.

The derived function is also written as $Df(x) = f'(x)$. The operation can be repeated to give the derivative of the derivative as $D(Df(x)) = D^2 f(x) = f''(x)$.

It is important at this stage to be aware of the fundamental idea:

> The derivative function $f'(x)$ is the result of a *global* operation D that operates on the original function f to give the derivative function
>
> $$Df(x) = f'(x).$$

With this fundamental idea in mind, it is time to relate the dynamic visualisation to the corresponding symbolic operation, linking human embodiment to mathematical symbolism to give a meaningful symbolic formula for the operation D in a range of different cases.



Slope from x to $x+h$

$$= \frac{f(x+h)-f(x)}{h}$$

$$= \frac{(x^2+2xh+x^2)-x^2}{h}$$

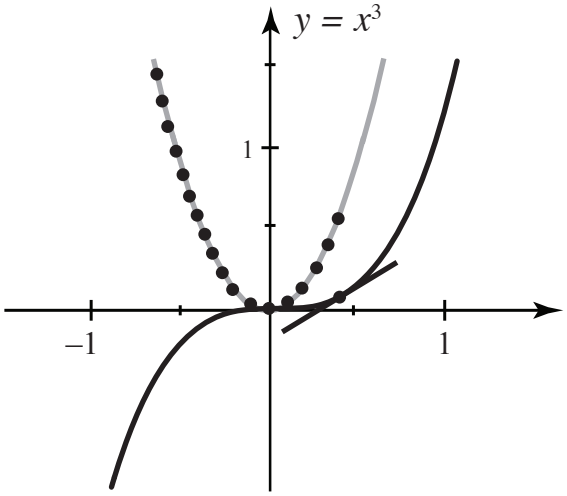
$$= 2x+h$$

For small h,
the slope stabilises to $\mathbf{2x}$.

Dynamic Visualisation **Symbolism**

Figure 18: Correspondence between visualisation and symbolisation

The same technique is possible for x^3, to give $D(x^3)=3x^2$.



Slope from x to $x+h$

$$= \frac{f(x+h)-f(x)}{h}$$

$$= \frac{(x^3+3x^2h+3xh^2+h^3)-x^3}{h}$$

$$= 3x^2+3xh+h^2$$

For small h,
the slope stabilises to $\mathbf{3x^2}$.

Dynamic Visualisation **Symbolism**

Figure 19: The derivative of x^3

The same technique and the binomial theorem for $(x+h)^n$ gives $D(x^n)=nx^{n-1}$. It is also possible to study the visualisations when n is fractional or negative and link the general idea of differentiation to what the student may know about powers.

Once the student has a link between the dynamic visualisation and the symbolism it becomes more appropriate to introduce the notation:

$$Df(x)=\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}$$

and begin to relate visualisation and symbolism for the standard functions.

Looking along the graphs sine and cosine (measured in radians, because this gives a natural way of relating the angle to the length of the circumference) reveals the

graph of $D(\sin x)$ stabilizes as $\cos x$ and $D(\cos x)$ stabilizes on the graph that is $\sin x$ upside down revealing $D(\cos x) = -\sin x$.

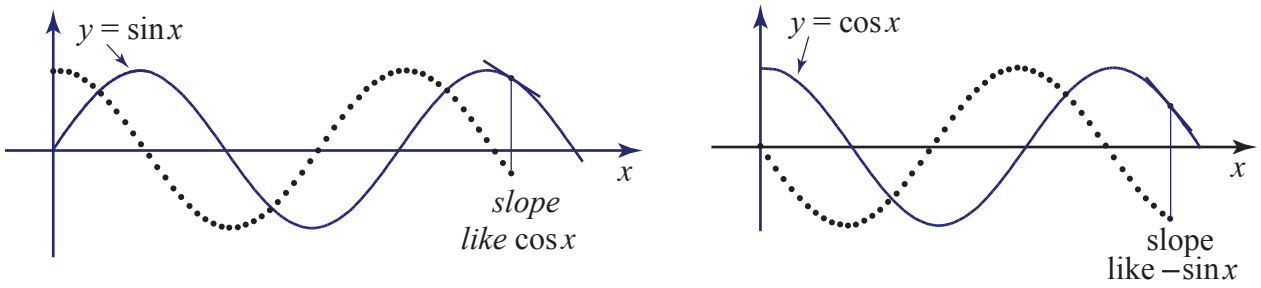


Figure 20: The slope of sinx is cosx and the slope of cosx is sinx upside down

At this point one can look at the symbolism in parallel to see how the symbolic computation works (which is usually difficult for students who may have only rote-learnt the formulae. However, now they can *see* the limit and realise that the minus sign in the derivative of cosine *x* is a natural property of the derived function.)

Moving on to the case of graphs of the form k^x where k is a constant, an investigation of the slopes of 2^x and 3^x reveals both have steadily increasing graphs, and each has steadily increasing slope functions. However, the graph of 2^x has a slope graph that is lower than the original, while 3^x has a slope graph that is higher.

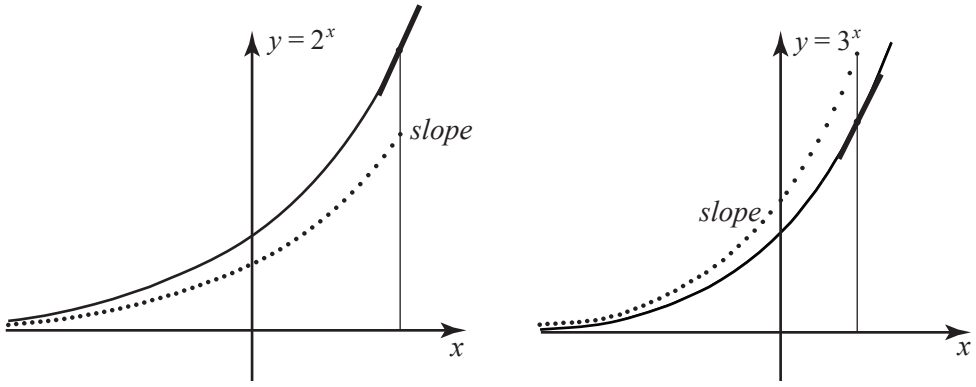


Figure 21; The slopes of exponential graphs

Our dynamically continuous perception can imagine k changing continuously from 2 to 3 suggesting that somewhere between 2 and 3 there should be a value e such that the slope of the graph of e^x and its slope function are the same.

By hoping that this function can be approximated by a (possibly lengthy) polynomial,

$$e^x = a_0 + a_1 x + a_2 x^2 + \ldots$$

then this must equal its derivative,

$$D(e^x) = a_1 + 2a_2 x + 3a_3 x^3 + \ldots$$

Putting $x = 0$, using $e^0 = 1$ gives $a_0 = 1$ and, comparing term by term gives $a_1 = a_0$, $2a_2 = a_1$, $3a_3 = a_2$, … to yield the values $a_1 = 1$, $a_2 = \frac{1}{2}$, $a_3 = \frac{1}{(2\times3)}$, …

So

$$e^x = 1 + x + x^2/2! + \ldots + x^n/n! + \ldots$$

Putting $x = 1$ gives

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \ldots + \frac{x^n}{n!} + \ldots$$

This is easily calculated without even using a calculator to give $e = 2.7182818285$ accurate to ten decimal places.

For the expert, this approach involves hidden problems, such as the idea that e^x is given by a polynomial of unspecified length. However, for the student coordinating good-enough arithmetic with dynamically changing graphs, it offers a natural extension of previous experience. In particular, by personally calculating *e*, the student experiences why the later terms become so small that they become irrelevant.

A sensible approach to the calculus

We are now in the position to consider 'a sensible approach' to the calculus, based on our human perceptions (Donald's level 1 and 2) which can then be analysed to produce mathematical theory (level 3). This involves a parallel development of conceptual embodiment (which involves the complementary use of human perception and action) and proceptual symbolism, which involves manipulation of symbols that arise from operations. (The term 'procept' denotes a symbol that can be used flexibly as either process or concept, and includes number, fraction, algebraic express, derivative, integral, and so on (Gray & Tall, 1994).)

A sensible approach is based on natural continuity (where a graph 'pulls flat') and local straightness (to give the concept of derivative). The limit concept is implicit in this development and can be made explicit at a later stage once the student can 'see' the derivative *Df* by looking along the originl graph *f* to imagine its changing slope as a new graph $Df(x)$ which may be written as $f'(x)$.

The Leibniz notation

The very first definition of the derivative of Leibniz (1684) did not mention limits or infinitesimals. Instead, he began with the notion of *tangent*, and defined the derivative to be *dy/dx* where *dx* and *dy* are the horizontal and vertical components of the tangent.

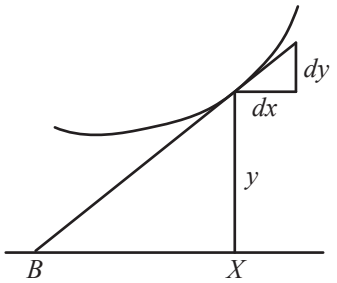


Figure 22: The original definition of Leibniz

He allowed *dx* to have any chosen value and then defined *dy* to be given by

$$dy = dx \times \frac{y}{BX}$$

where y is the abcissa and BX is the subtangent. This leads to the problem of calculating BX and to do this one needs to know the equation of the tangent. Leibniz's solution was to imagine the curve to be a polygon consisting of an infinite number of infinitesimal straight sides.

We can use the notion of dx and dy being the components of the tangent and, given the natural concept of 'local straightness', we now see that under high magnification, a tiny portion of the graph will look like a straight line. This means we can imagine not just the slope of the tangent, but the slope of the graph itself.

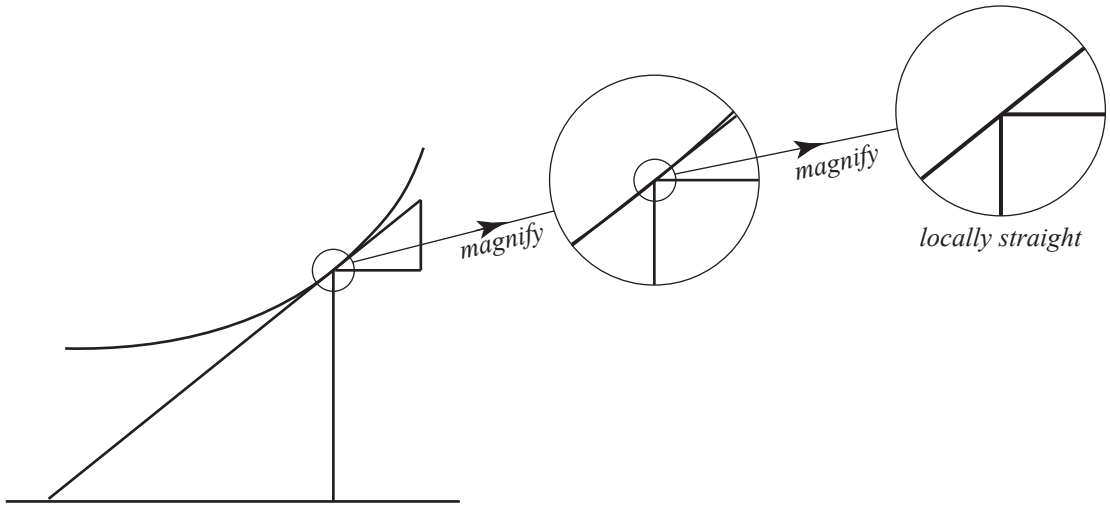


Figure 23: Magnifying the Leibniz triangle

This view enables us to see that

$$Df(x) = \frac{dy}{dx}$$

revealing the derivative function $Df(x)$ as the quotient of the components of the tangent. These may be called *differentials.* A differential is simply a component of the tangent vector. The vector (dx, dy) is the *direction* of the tangent vector.

The standard derivatives

Using this approach, all the derivatives of standard functions x^n, $\sin x$, $\cos x$, e^x, $\ln x$ can be *seen*. Only $\ln x$ remains. This can be done by noting the inverse relationship $y = \ln x$ and $x = e^y$ and interchanging the axes.

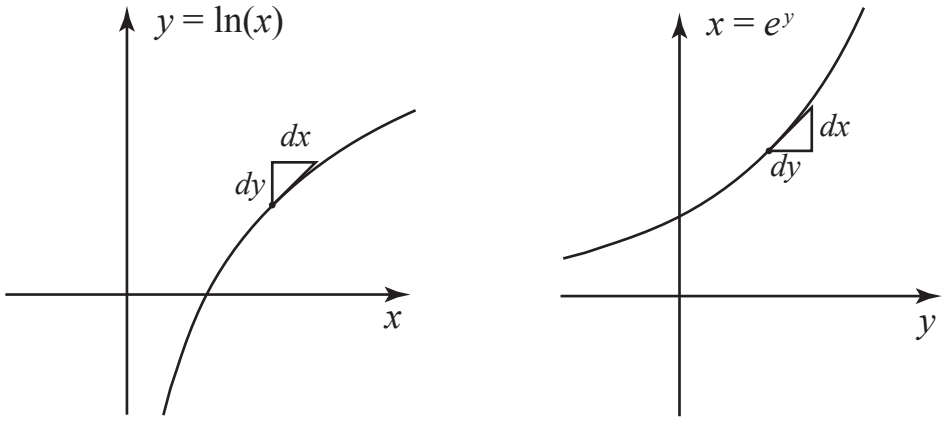


Figure 24: The derivative of the inverse function is found by interchanging the axes

Since for $x = e^y$ one knows

$$\frac{dx}{dy} = e^y = x$$

so

$$\frac{dy}{dx} = \frac{1}{x}.$$

This is possible because *dy/dx* is here considered as a *quotient of components*, not as a limit.

The rules of the calculus and the need for the limit concept

Having motivated the idea of the derivative as the changing slope of a given function, and visualised the derivatives of various standard functions such as $\sin x$ and e^x, it is time to develop general methods to calculate the derivatives of combinations of functions such as $e^x \sin x$ or $e^{\sin x}$.

This involves the derivation of the formulae for $D(f(x)+g(x))$, $D(f(x)g(x))$, $D(f(x)/g(x))$ and $D(f(g(x))$, all of which can be performed by the usual techniques of calculating the slope from *x* to $x+h$ and considering what happens as *h* gets small. This is particularly necessary for the product and quotient.

The chain rule can be calculated in the same way, but it also has an alternative visual meaning by representing $y = f(x)$, $z = g(y)$ as a graph in three-dimensional *x-y-z* space where the projections on to the three coordinate planes represent the graphical relationships between the variables in pairs. The components of the tangent vector to the curve are (dx, dy, dz) and, are *lengths*, so we can write the derivative of *z* with respect to *x* as

$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}.$$

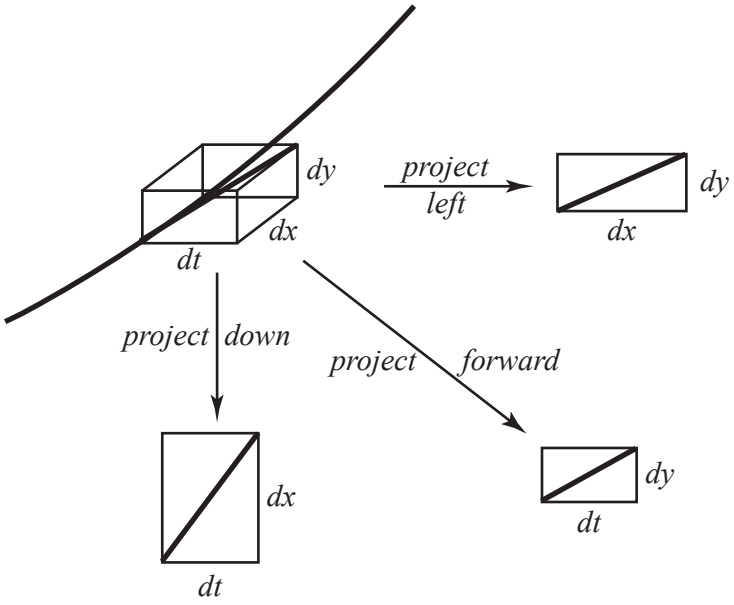


Figure 25: The components of the tangent vector in three dimensions as dx, dy, dz.

Writing the composite function as $h(x) = g(f(x))$, the corresponding function notation is

$$Dh(x) = Dg(y)Df(x)$$

which can also be written as

$$h'(x) = g'(f(x))f'(x)).$$

There is a technical point here. The formula using the Leibniz notation as a ratio of components is only applicable if the denominator is zero, so the argument cannot be used if $f'(x) = 0$, for then $dy = 0$. But in this case, $dz = g'(y)\,dy$ is zero, so the chain rule $h'(x) = g'(f(x))f'(x))$ .still holds because both sides are zero.

Parametric Functions

A similar picture can be drawn for a parametric function $x = x(t)$, $y = y(t)$ by visualising the curve in *t*-*x*-*y* space as *t* varies. The following picture shows $x = \cos t$, $y = \sin t$ as *t* increases from 0 to 10. The projections onto the three coordinate planes show *x* as a function of *t*, *y* as a function of *t* and the relationship between *x* and *y* as the point $(x(t), y(t))$ moves around a circle.

If one draws a tangent to the curve in three space, with components dt, dx, dy, then these are the sides of a cuboid and one may write

$$\frac{dy}{dx} = \frac{dy}{dt} \Big/ \frac{dx}{dt}\ .$$

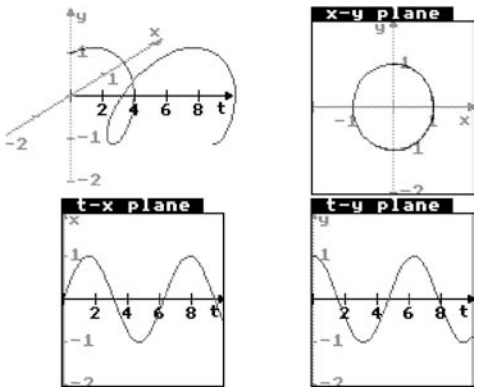


Figure 26: The parametric curve $x = \cos t$, $y = \sin t$ *in three-space with projections onto the three coordinate planes*

Because *dt*, *dx*, *dy* are all *lengths* in this interpretation, the equation is valid as quotients of differentials (Tall, 1992). Since *dt* can always be taken to be non-zero, the only technicality occurs when $dx = 0$ and the tangent is perpendicular to the *x*-axis.

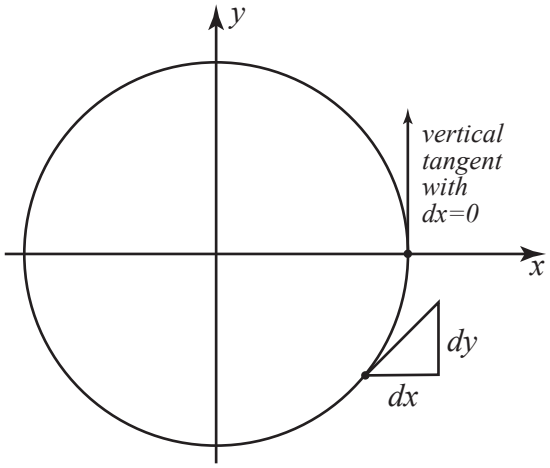


Figure 27: A vertical tangent to a parametric curve

Implicit functions

An implicit function $F(x, y) = 0$ involving a relationship between *x* and *y* occurs as a naturally drawn precisely when it is possible to imagine tracing a finger along it. This will give a parametrization $x = x(t)$, $y = y(t)$ as the point moves along in time *t*. In this case, the implicit function can be written as $F(x(t), y(t)) = 0$ and differentiated with respect to *t* to give $DF(x(t), y(t)) = 0$.

For instance, if

$$F(x, y) = x^2 + y^2 - 1,$$

then the differentiating both sides of the equation $F(x, y) = 0$ gives

$$2x\frac{dx}{dt} + 2y\frac{dy}{dt} = 0$$

and, because these are equations involving the differentials as *lengths*, we can

multiply through by *dt* to get the 'differential equation'

$$2x\,dx + 2y\,dy = 0.$$

When $dx \neq 0$, this can be rewritten to calculate the derivative of *y* with respect to *x* in a part of the graph where *y* is given as a function of *x* to get

$$x + y\frac{dy}{dx} = 0.$$

Differential Equations

Differential equations are just equations involving differentials (the components of the tangent vector). As such they specify the direction of the solution curve at a point. Because a differentiable function is *locally straight*, one can approximate the graph through a point by a short straight line in the direction of the tangent. In the picture, the differential equation is

$$\frac{dy}{dx} = \frac{x}{y}.$$

A solution curve has been built by joining pieces together. To get a better picture, the slope of the curve given by the differential equation is calculated at the midpoint of the line segment. (Technically this gives a second order approximation to the solution curve and a much more accurate picture.)

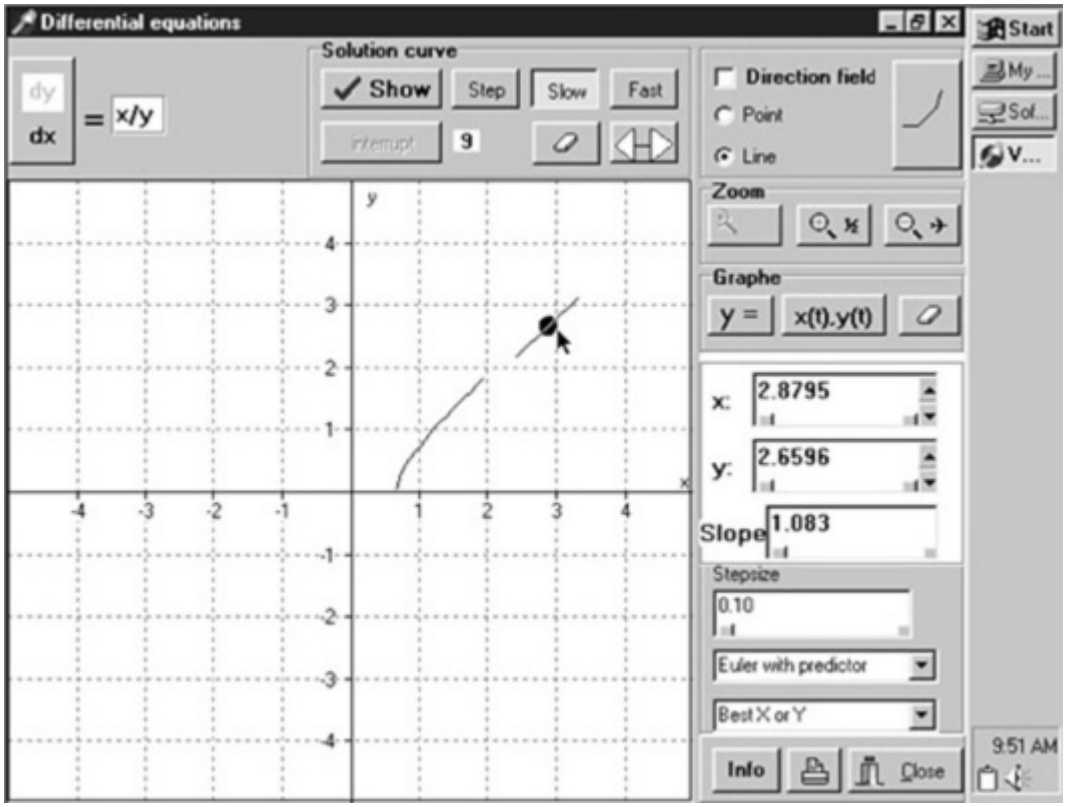


Figure 28: Building a solution of a differential equation by following the direction it specifies (Blokland & Giessen, 2000)

Partial Derivatives

For a function of two variables $z = f(x, y)$, the partial derivatives

$$\frac{\partial z}{\partial x} = \lim_{h \to 0} \frac{f(x+h, y) - f(x, y)}{h}, \quad \frac{\partial z}{\partial y} = \lim_{h \to 0} \frac{f(x, y+h) - f(x, y)}{h}$$

can be visualised in three dimensions as the slopes of the tangent plane in the vertical planes given by y = constant and x = constant, respectively. If increments dx and dy are given to x and y, and the resulting increment to the tangent plane is dz, then we can calculate

$$dz = \frac{\partial z}{\partial x} dx + \frac{\partial z}{\partial y} dy$$

and, clearly, cancellation of the dx and the ∂x are not possible as it would give something like $dz = \partial z + \partial z$ and this is clear nonsense.

I can reveal to you that this is because the notation is inadequate. What we should do is to look at the vertical planes that intersect the surface and its tangent plane. An increment dx in the vertical x-z plane will intersect the tangent plane in a tangent line whose vertical component may be denoted dz_x, and similarly an increment dy gives a corresponding vertical increment to the tangent plane written as dz_y. The equation now becomes

$$dz = \frac{dz_x}{dx} dx + \frac{dz_y}{dy} dy$$

where cancellation is now possible to give

$$dz = dz_x + dz_y$$

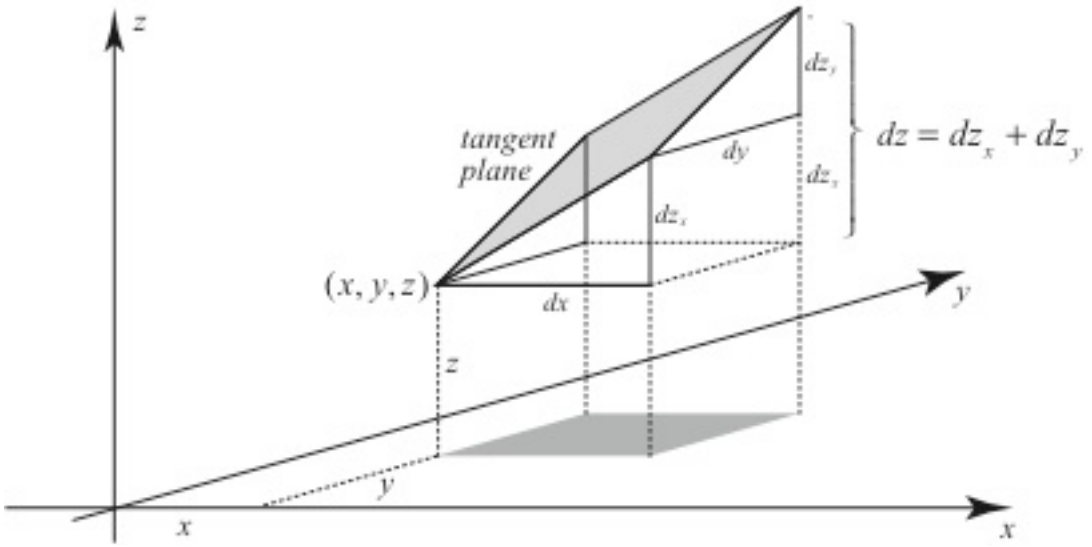


Figure 29; The tangent plane to a surface z=f(x,y) (Tall, 1992)

Integration

I handle integration by using a slightly different notation. A practical calculation of the area under a graph $y = f(x)$ from $x = a$ to $x = b$ takes a partition of intermediate points $x_0 = a$, x_1, x_2, …, $x_n = b$, and defines $dx_k = x_k - x_{k-1}$. Usually, the points are taken in order with $x_{k-1} < x_k$, but this is not necessary or desirable, particularly if one wishes to consider $b < a$. The 'mesh-size' of the partition is the largest value of $|dx_k|$. The Riemann sum from *a* to *b* is

$$\sum_a^b f(x)\,dx = \sum_{k=1}^n f(x_k)\,dx_k$$

As the mesh-size gets smaller, for a naturally continuous function *f* we can 'see' that this stabilizes to the area $A(a,b)$ under the graph, which is often written as

$$A(a,b) = \int_a^b f(x)\,dx.$$

Leibniz originally used the even more economical symbol $A = \int y\,dx$ for the area which he envisaged as the sum of strips height *y*, width *dx*. He considered that, when *x* is increased by a quantity *dx*, the area *A* is increased by a quantity *dA* which is the area of a very thin strip width *dx*, height *y*.

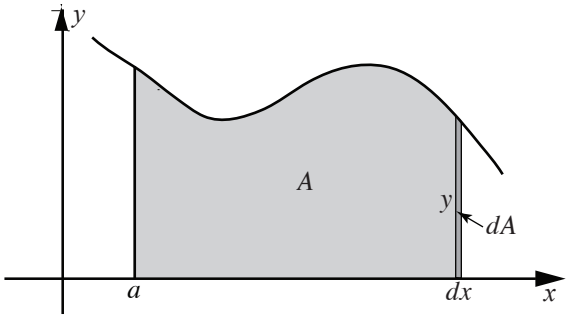


Figure 30: The increase dA in area equals $y\,dx$

The thin strip of area *dA* is not an exact rectangle as the top is part of the curved graph. However, for a continuous function, the final strip width *dx* can be taken so small that when the strip is stretched horizontally, then the graph will pull flat and the increase in area looks like a rectangle width *dx*, height *y*.

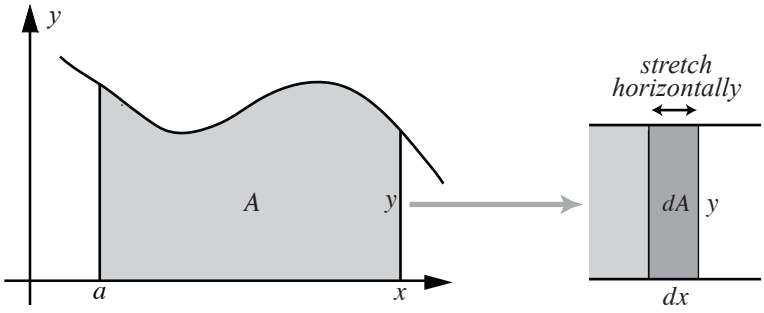


Figure 31; Stretching a thin strip that pulls flat

Using 'good-enough' arithmetic, the area is

$$dA = y\,dx.$$

Any error that occurs because the curve is not precisely horizontal is contained within the thickness of the pencil line used to draw the graph.

Dividing the equation through by dx, Leibniz obtained the relation:

$$\frac{dA}{dx} = y.$$

If the area is measured from a different point a', the area function will differ by a constant c which represents the area between a and a'. The two equations

$$A = \int y\,dx + c \quad \text{and} \quad \frac{dA}{dx} = y$$

express in the simplest terms that the operations of integration and differentiation are essentially inverses of each other, giving the *Fundamental Theorem of Calculus*. This is an amazing compression of knowledge, expressing the essential connection between change and growth in two brief equations!

Calculus and Analysis within a long-term framework of development

In recent years, I have been combining and distilling my personal experiences of researching the nature of mathematical thinking to produce a single framework for its long-term development, which proves to be valid to analyse not only the development of a single individual (Tall, 2004, 2008), but also the development of mathematical thinking in history (Katz & Tall, to appear).

The framework involves three distinct forms of mathematical development:

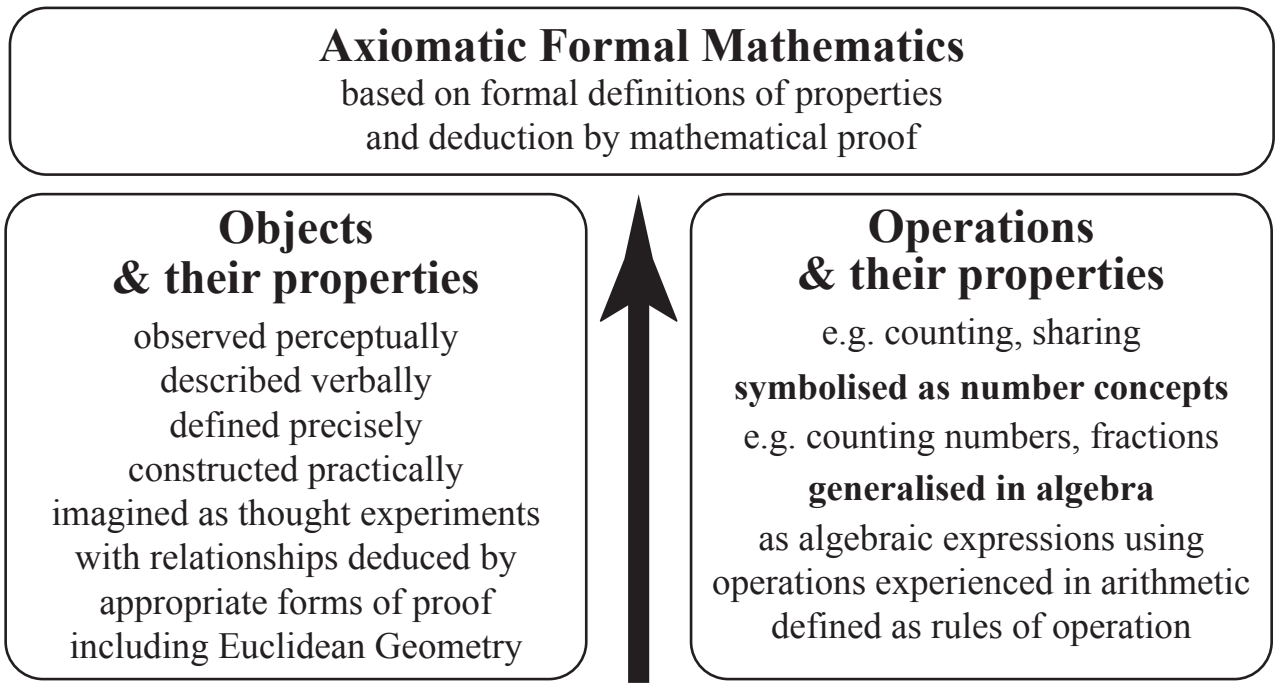


Figure 32: Three worlds of mathematics

Each has its own modes of operation and development to such an extent that I specify them as three distinct mental *worlds* of mathematics (Tall, 2004):

> (1) *Conceptual embodiment* builds on human perceptions and actions, developing mental images that are verbalized in increasingly sophisticated ways and become perfect mental entities in our imagination.
>
> (2) *Proceptual symbolism* grows out of physical actions into mathematical procedures that are symbolized and conceived dually as operations to perform and symbols that can themselves be operated on by calculation and manipulation (procepts).
>
> (3) *Axiomatic formalism* builds formal knowledge in axiomatic systems in a suitable foundational framework (such as formal set theory or formal logic) whose properties are deduced by mathematical proof.

Within this framework, graphs and the notion of slope inhabit the conceptual embodied world of objects and their properties in terms of natural continuity and local straightness. The symbolic concepts of function and the derivative lie in the world of proceptual symbolism. Elementary calculus develops as a blend of the two. Meanwhile, mathematical analysis inhabits the axiomatic formal world which involves a substantial change in meaning with formal definitions including the epsilon-delta definition of limit.

While the world of axiomatic formal mathematics is a working environment for the presentation of formal definitions and formal proof, it is not a suitable environment for elementary calculus which builds more naturally on embodiment and symbolism.

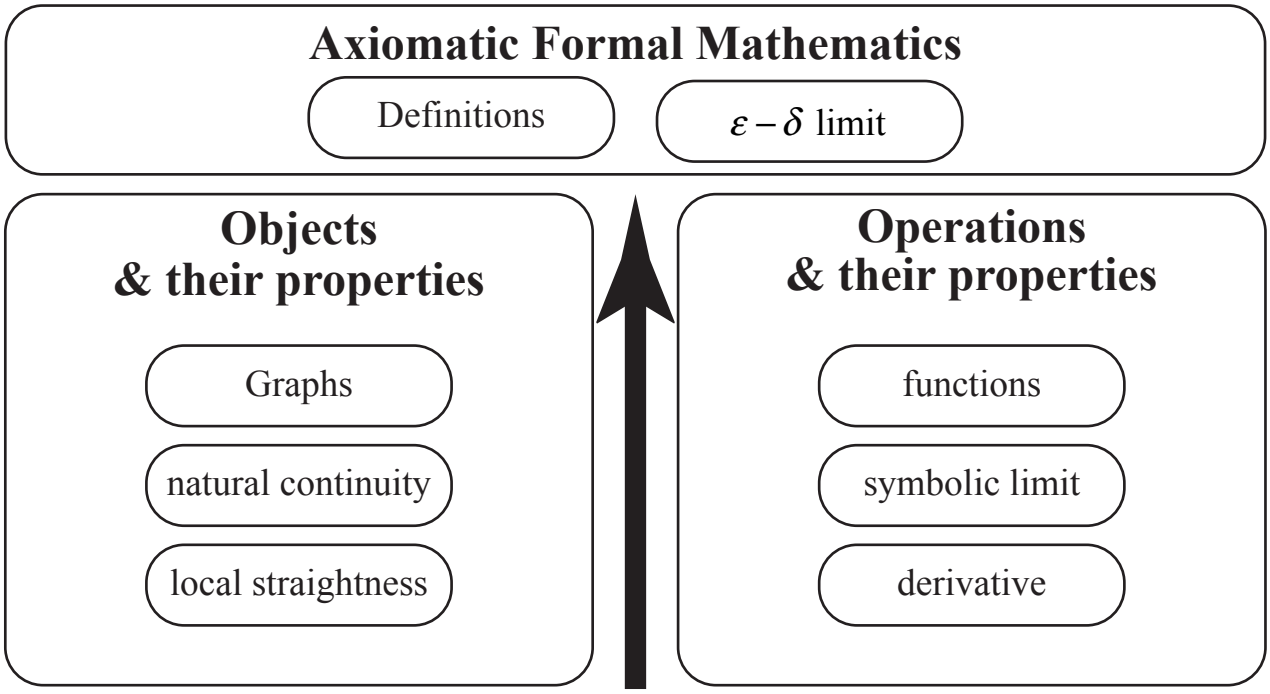


Figure 33: Calculus and Mathematical Analysis in the Three World framework

Limitations and extensions of other theories

At this point is is possible to reflect on other theories that each make enormous contributions to the development of mathematical thinking and yet are in need of extension or modification to provide a fuller insight into the conceptual growth of calculus and mathematical analysis.

The APOS theory of Dubinsky and his colleagues is based on Piaget's concept of reflective abstraction in which

> ... a physical or mental action is reconstructed and reorganized on a higher plane of thought and so comes to be understood by the knower. (Beth & Piaget 1966, p. 247).

This is developed into APOS theory (Asiala et al. 1996), where ACTIONS are routinized as PROCESSES, encapsulated as OBJECTS and embedded in a SCHEMA of knowledge. Dubinsky's approach uses the programming language ISETL and its first focus is on ACTIONS involving programming numerical algorithms that can be written in a functional manner that may then be used as mental entities (OBJECTS) in their own right. The graph is an afterthought, drawn only *after* the function is programmed symbolically. Cottrill et al. (1996) reported that students were able to conceptualise functions as processes, but only a few understood the formal definition of limit, and not one student in their study applied the formal definition spontaneously.

There are two distinct aspects here, one is the apparent difficulty of conceiving of a function as a process, but not as an object; the second is the difficulty of the formal definition of limit.

Conceptualizing a function as an object in APOS theory involves focusing on symbolism and programming the function notion to encapsulate a *process* as an as yet unconstructed symbolic object In sharp contrast, the sensible approach to the derivative advocated here begins with an embodied operation on a visible *object* (the graph of a function *f*) to see its changing slope and construct a new visible *object* (the stabilized slope function *Df*). I hypothesize that a theory formulating the encapsulation of a *process* into an as yet unknown *object* will have less sensible meaning than the operation on a visible *object* to give a new visible *object*.

The three-world framework focuses on the difficulty in making sense of the formal definition of a limit by placing it in a distinct world of operation, deducing theoretical constructs from multi-quantified definitions using formal proof. This is the world inhabited by mathematical analysis. Elementary calculus arises naturally in the parallel worlds of embodiment and symbolism using a visual and practical approach to translate concepts of change represented by graphs, rate of change (slope function) and cumulative growth (area) into symbolic manipulation of functions, derivatives and integrals.

A very different view of cognitive growth is formulated by Lakoff and Nunez (2001) in their book *Where Mathematics Comes From*, claiming that all thinking is embodied, and concepts develop from sensori-motor structures in the brain. This very compelling theory seems in harmony with a sensible approach. It builds on natural continuity and remarks on the clear distinction between elementary calculus

and formal analysis in terms of the differing metaphors that arise from natural dynamic continuity in calculus and formal static epsilon-delta arguments in analysis. This is consistent with the distinction between the elementary worlds of embodiment and symbolism in the calculus and formalism in mathematical analysis.

However, mathematicians see the nature of mathematical analysis in a variety of different ways. Some seek a natural approach building on their previous experience. Such a developmental path can be seen in the earlier description of transforming the concept of continuity into the formal epsilon-delta definition. Other mathematicians see formal mathematics as a completely new start, building theorems deductively from the definitions, and constructing a new formal structure based entirely on definition and proof.

Pinto & Tall (1999, 2001) identified a spectrum of performance in students studying analysis from those who *give meaning to* formal definitions from their concept imagery in what they termed a *natural* approach, to those who *extract meaning from* the definitions by learning how to handle multi-quantified statements and proving theorems by mathematical deduction using a *formal* approach. Such an analysis has been confirmed by Weber (2004) who added a further category of procedural learning in students, and other studies demonstrating different students successfully following either a natural or a formal route (Pinto & Tall, 2002, Alcock & Simpson, 2005).

This suggests at least two distinct routes to mathematical analysis, one prefaced by a natural transition from concepts such as natural continuity and local straightness to formal definitions, another by formal deductions within an axiomatic system. Whichever method is used, the eventual product is a knowledge structure where all the theorems are deduced from fundamental axioms and definitions. At one end of the spectrum is a knowledge structure linked to embodied images, at the other is a knowledge structure based on linguistic definitions and formal deduction.

Núñez, Edwards, & Matos (1999) speak of natural continuity, based on embodiment offering a grounding for mathematics education, consistent with the search for a natural route from calculus to analysis. They declare that mathematics is embodied, created by the human mind and that it cannot exist in any platonic sense outside the human mind. However, this is not to say that there is not a coherent structure of mathematics 'out there' to be discovered by human perception and reflection.

For example, working with a simple system which starts somewhere, makes a single step, then another, then another, produces an unending sequence of distinct entities that we call 1, 2, 3, … . This is a natural structure that necessarily *has* properties, which may be discovered to have a specific form. Two and two is always four and not five. Products of numbers give composite numbers and those that are not composite, called primes, are in this structure to be discovered. Systems based on such elemental operations are not simply created by the human who names them, they are universal and are discovered by human reason.

It is a surprise to me that the theoretical framework of Lakoff and his colleagues, with a significant basis in linguistics, prefers a natural development from sensory perceptions rather than a formal approach from carefully defined linguistic definitions. Formal mathematical thinking has an advantage over natural thinking. By selecting specific axioms, which may focus only on a specific aspect of a natural situation, such as the epsilon-delta definition of continuity, a whole new word of formal consequences follow which give new possible insights. Furthermore, formally deduced theorems from specific axioms may give a precise formal structure that can now be interpreted to give entirely new forms of embodiment, now based not only on sensory perception, but on logical deduction. The theorems proved operate not just in a specific situation, but in *any* context where the axioms hold.

For example, our sensory perception cannot *see* infinitesimal quantities that are abitrarily small. However, in the formal world, we can speak of an ordered field extension of the real numbers (as a complete ordered field). Such a field *must* contain infinitesimals, and any finite quantity is uniquely expressed as a real number plus infinitesimal. Furthermore, the field can now be represented as a visual number line where it is possible to distinguish visually between any two elements u,v by magnifying part of the picture containing the two quantities by a factor $1/(u-v)$ (Tall, 2002). Thus formal mathematics, built on linguistic definitions can give rise to new mental embodiments that can be pictured physically in a manner that has previously not been thought possible.

The formulation of three worlds of mathematics encompasses three different modes of thinking in mathematical development, from human embodiment of perception and action with a parallel operational development in symbolism and a more sophisticated form of definition and deduction in axiomatic formalism.

Elementary calculus belongs in the parallel worlds of embodiment and symbolism based on the perceptual ideas of natural continuity and local straightness. Mathematical analysis belongs in the more sophisticated world of axiomatic formalism where students learn to argue rigorously from formal quantified definitions (Tall & Mejia-Ramos, 2004).

Mathematical analysis can certainly be developed formally from axioms and definitions, but it can also be based on the blending of embodiment and symbolism that offers not only a sensible foundation for the calculus but also a natural basis for the full range of higher level developments. These include the modelling and solution of problems in applied mathematics, the formal concepts of mathematical analysis and also a sensible basis for the infinitesimal methods of non-standard analysis.

References

Alcock, L. & Simpson, A.P. (2005). Convergence of Sequences and Series 2: Interactions between Nonvisual Reasoning and the Learner's Beliefs about their own Role. *Educational Studies in Mathematics* 58 (1), 77-100.

Asiala, M., Brown, A., DeVries, D., Dubinsky, E., Mathews, D., Thomas, K. (1996), A framework for research and curriculum development in undergraduate mathematics education, *Research in Collegiate Mathematics Education II,* 1-32.

Beth, E. W., & Piaget J. (1966). *Mathematical Epistemology and Psychology* (W. Mays, trans.), Dordrecht: Reidel.

Blokland, P., & Giessen, C., (2000). *Graphic Calculus for Windows*. Available from http://www.vusoft2.nl.

Cottrill, J., Dubinsky, E., Nichols, D., Schwingendorf, K., Thomas, K., Vidakovic, D. (1996). Understanding the Limit Concept: Beginning with a Coordinated Process Scheme. *Journal of Mathematical Behavior*, 15 (2), 167-192.

Donald, M. (2001). *A mind so rare*. New York: Norton.

Durell, C. V. & Robson, A. (1934). *Elementary Calculus*. London: Bell.

Gray, E. M. & Tall, D. O. (1994). Duality, ambiguity and flexibility: A proceptual view of simple arithmetic, *Journal for Research in Mathematics Education*, 25 2, 115–141.

Lakoff, G. & Nunez, R. (2000). *Where Mathematics Comes From*. New York: Basic Books.

Leibniz, G. W. (October 1684). Nova methodus pro maximis et minimis, itemque tangentibus, qua nec fractas, nec irrationales quantitates moratur, & sinulare pro illis calculi genus, *Acta Eruditorum* 467-473.

Núñez, R. E, Edwards, L. D. & Matos, J. F. (1999). Embodied cognition as grounding for situatedness and context in mathematics education, *Educational Studies In Mathematics* 39, 45–65.

Pinto, M. M. F. & Tall, D. O. (1999). Student constructions of formal theory: giving and extracting meaning. In O. Zaslavsky (Ed.), *Proceedings of the 23rd Conference of PME, Haifa, Israel,* 4, 65–73.

Pinto, M. M. F. & Tall, D. O. (2001). Following students' development in a traditional university classroom. In Marja van den Heuvel-Panhuizen (Ed.) *Proceedings of the 25th Conference of the International Group for the Psychology of Mathematics Education* 4, 57-64. Utrecht, The Netherlands

Pinto, M. M. F. & Tall, D. O. (2002). Building formal mathematics on visual imagery: a theory and a case study. *For the Learning of Mathematics*. 22 (1), 2–10.

Tall, D. O. (1975). A long-term learning schema for calculus/analysis, *Mathematical Education for Teaching*, 2, 5 3–16.

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

Tall, D. O. (1992). Visualizing differentials in two and three dimensions, *Teaching Mathematics and its Applications*, 11 1, 1–7.

Tall, D. O. (2004). The three worlds of mathematics. *For the Learning of Mathematics*, 23 (3). 29–33.

Tall, D. O. (2008). The Transition to Formal Thinking in Mathematics. *Mathematics Education Research Journal,* 20 (2)*,* 5-24.

Tall, D. O. (2009). Dynamic mathematics and the blending of knowledge structures in the calculus. *ZDM – The International Journal on Mathematics Education*, 41 (4) 481–492.

Tall, D. O. & Mejia-Ramos, J. P. (2004). Reflecting on Post-Calculus-Reform: Opening Plenary for Topic Group 12: Calculus, International Congress of Mathematics Education, Copenhagen, Denmark. http://www.icme-organisers.dk/tsg12/papers/tall-mejia-tsg12.pdf

Tall, D. O., McGowen M. & DeMarois, P. (2000a). Using the Function Machine as a Cognitive Root for building a rich concept image of the Function Concept, *Proceedings of PME-NA*, 1, 247–254.

Tall, D. O., McGowen M. & DeMarois, P. (2000b). The Function Machine as a Cognitive Root for the Function Concept, *Proceedings of PME-NA*, 1, 255–261.

Vinner, S. (1982). Conflicts between definitions and intuitions: the case of the tangent, *Proceedings of PME 6*, Antwerp, 24–28.

Weber, K. (2004). Traditional instruction in advanced mathematics courses: a case study of one professor's lectures and proofs in an introductory real analysis course, *Journal of Mathematical Behavior*, 23 115–133.