The Anatomy of a Discovery in Mathematics Research

David Tall

Mathematics Education Research Centre,
University of Warwick,
COVENTRY CV4 7AL, UK.

What are the activities that constitute mathematical research? Various well-known mathematicians such as Poincaré, Hadamard and Polya have given descriptions, but these tend to be general reflections on the process considered after the event. The aim of this paper is to describe, within practical limits, the thought processes in a specific piece of research as they happened. It highlights the research activities of the author over ten days, relating these to the previous development of ideas over a period of years and the developments which followed. There were flashes of insight, the coming together of previous experiences, analogies both useful and false, and intuitions having the ring of truth which proved to be embarrassingly inaccurate.

After grappling with ideas which seemed complex at the time, the final product was a theory so inevitable that it seemed like a mathematical truth discovered. A year later it seemed naive, even trivial and, when presented to students, they found it straightforward, simple and obvious. But the tortuous route by which the author came to build up the theory is a story worth telling, if only because the way that it actually happened (as witnessed by notes taken at the time) was far less glamorous and logical than the memories that were subsequently recalled. In some instances memories a year later were quite different from the evidence as concretely represented by the notes. It seems that we remember the salient features of a past event and reconstruct the detail when required. In this way our recollections are far more rational than the actual processes. Even in the telling of the story it has been necessary to select material and so a certain amount of rationalisation has inevitably crept in. In doing this I have attempted to give an overall impression of the research activity and, within this programme, select certain themes that intertwine together as the work progresses. I have written of myself in the third person, as a separate observer might have done. This allows me to talk of the incorrect turns I took without (too much) embarrassment. The story has been written in such a way that events are usually reported without revealing subsequent occurrences that were unknown at the time. In this way the reader may participate in the hopes and fears as they happened without knowing of later reversals of fate. When this rule is broken it is signified by the relevant passage being placed in square brackets.

Published in *For the Learning of Mathematics*, 1, 2 (November, 1980), 25–34.

The spur

If *AB* is a line segment half the length of *CD*, are there the same number of points in *AB* as *CD*, or more, or less? (Figure 1.)

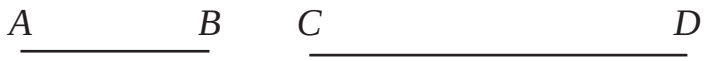


Figure 1

If *WXYZ* is a square whose side-length is the same as *AB*, does *WXYZ* (and its interior) have as many points as *AB*, or more, or less? (Figure 2.)

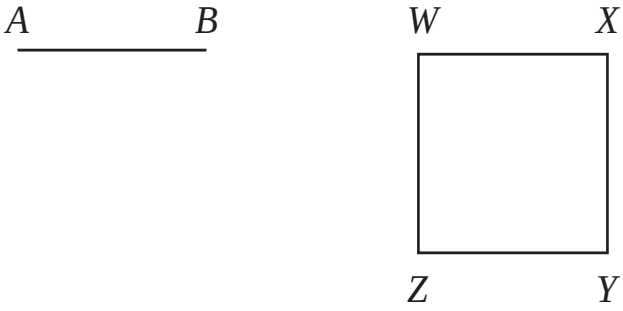


Figure 2

These questions were typical of a number in a questionnaire [1] to investigate a child's concept of infinity posed by Professor Fischbein of Tel-Aviv University. A natural intuitive response by a child might be that there are twice as many points in CD as AB and many more points in the square than on the line. These answers are not in accord with the accepted theory of cardinal numbers. They prompted the author to embark on a discussion with Professor Fischbein which led to a formal theory vindicating these intuitive responses, [12], [14].

An earlier ambition recalled

A project of the author which had lain fallow for several years provided the second ingredient for the ensuing research work—his interest in infinitesimals. He was aware of Abraham Robinson's pioneering work on non-standard analysis [7] and had briefly studied a text on infinitesimal calculus by Keisler [5]. But he failed to understand either book completely and this gave rise to an ambition to develop a simpler theory of calculus using infinitesimal techniques. (The research here described gave rise to such a theory, subsequently published, [12], [13].)

In brief, Robinson's theory allows one to consider an extended number line which includes infinitesimal quantities which are smaller in size than any positive real number. The reciprocals of such quantities are infinite in the sense that they are larger in size than any real

number. Robinson's extended number system, the hyperreal numbers, has three kinds of element:

(i) the set I of infinitesimals,

(ii) the set F of finite numbers, of the form $x+\varepsilon$ where r is a real number and ε is an infinitesimal,

(iii) the set L of infinite elements.

If one includes zero as an infinitesimal then algebraically I is an ideal in the ring F.

For the purpose of reading this paper, all the reader needs to know is that for any finite hyperreal number $\_=x+\delta$, the real number x is called the *standard part* of a and is denoted by st a. In intuitive terms taking the standard part is essentially the process of ignoring the infinitesimal part of a finite hyperreal. In algebraic terms, however, it has a very precise meaning. It may be shown that the map $\mathrm{st}:F\to\boldsymbol{R}$ is a ring homomorphism with kernel I.

In Robinson's calculus one may therefore define the derivative $f'(x)$ by computing $(f(x+\varepsilon)-f(x))/\varepsilon$ (for a non-zero infinitesimal ε) and then taking the standard part. For instance, if $f(x) = x^2$, then

$$\frac{f(x+\varepsilon)-f(x)}{\varepsilon}=\frac{(x+\varepsilon)^2-x^2}{\varepsilon}=2x+\varepsilon,$$

so $f'(x)=\mathrm{st}(2x+\varepsilon)=2x.$

The historical process of computing with infinitesimal quantities and then ignoring them now has a strictly logical formulation. But what is an infinitesimal?

An example of a system with infinitesimals

At the time of beginning this work in 1978, Dr Tall was only aware of one example of a simple system including infinitesimals (Robinson's system involves a logical construction). In a simple system, now to be described, infinitesimals are represented not as points on a line, but as rational functions. An understanding of this is crucial in what follows, so we spend a little time considering it.

Let $\boldsymbol{R}(t)$ be the field of rational functions $f(t)/g(t)$ where $f(t)$ and $g(t)$ are polynomials in an indeterminate t. One defines a non-zero rational function $\alpha(t)=f(t)/g(t)$ to be "positive" or "negative" as follows. The non-zero polynomials $f(t)$, $g(t)$ have only a finite number of zeros, so for real x exceeding these zeros, the sign of $\alpha(x)$ is strictly positive or negative and this is defined to be the sign of $\alpha(t)$. For instance,

$(t^2-3t)/(t+1)$

is "positive" according to this definition because

$(x^2-3x)/(x+1) = x(x-3)/(x+1) > 0$ for real $x > 3$.

Next we order rational functions by defining

$\alpha(t) > \beta(t)$ if and only if $\alpha(t) - \beta(t)$ is "positive".

For instance $t > 27$ because the rational function $t–27$ is positive. In fact $t>a$ for any real number a, which leads to the reason why this example is so crucial. In $\boldsymbol{R}(t)$ the element t is greater than any real number a. In general we say $\alpha(t)$ is *positive infinite* if $\alpha(t) > a$ for all real numbers a. The element t is "infinite" in this sense.

Geometrically one may visualise the comparison between rational functions by drawing their graphs. A rational function is "positive" if its graph is above the axis for large values of t and the relation $\alpha(t) > \beta(t)$ holds when the graph of $\alpha(t)$ is above that of $\beta(t)$ for large values of t. In Figure 3 we see that t is positive and $t > a$ for every real a, so t is positive infinite.

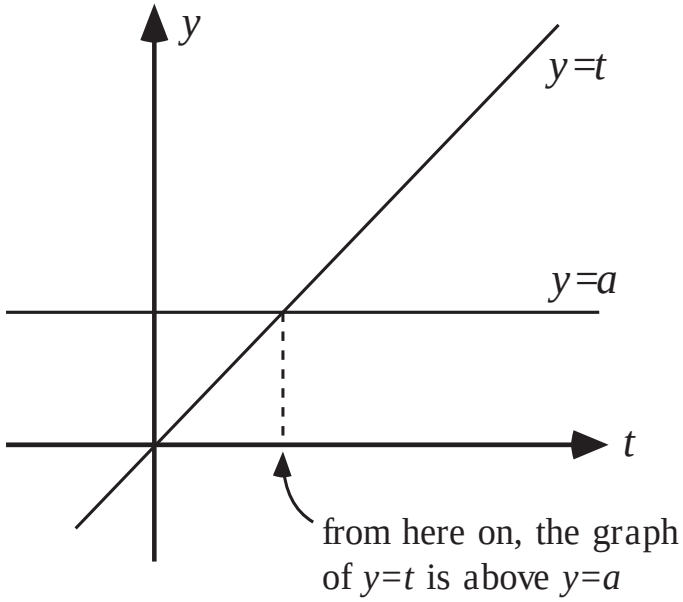


Figure 3

In Figure 4 we see that $0 < 1/t < a$ for all positive real a, simply because the graph of $y=1/t$ is between $y=0$ and $y=a$ for large positive values of t

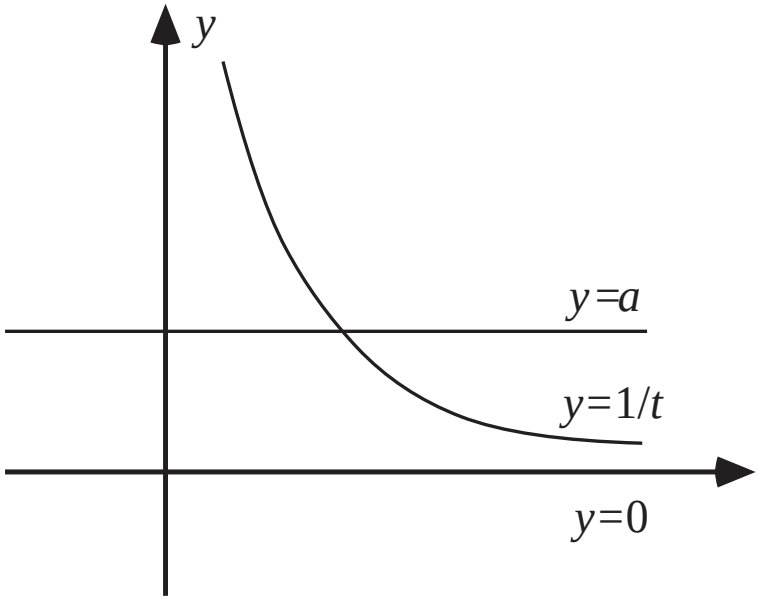


Figure 4

Thus 1/*t* is a *positive infinitesimal* in the sense that it is smaller than any positive real *a* and larger than 0.

Dr. Tall's ambition was to use this field $\boldsymbol{R}(t)$ as the basis for the theory of calculus with infinitesimals. His problem was that these infinitesimals were *functions*, not *points.* In addition, he needed to be able to compute $f'(a+\varepsilon)$ for real *a* and infinitesimal ε to be able to compute $f'(x)$. He could see how to do this when *f* was a rational function because when $a \in \boldsymbol{R}$ and $\varepsilon \in \boldsymbol{R}(t)$, then $f(a+\varepsilon)$ may be considered as a rational function of a rational function and the composition of two functions gives $J(a+\varepsilon) \in \boldsymbol{R}(t)$. He could not see how to extend the ideas to more general functions. He gave up.

The events leading to the first insight

In the summer of 1978 Dr Tall attended a conference on the psychology of learning mathematics in which Professor Fischbein described his experiments on children's intuition of infinity, mentioned above. He noted that Professor Fischbein's interpretation of infinity was essentially a *cardinal* infinity and he resolved to explain the alternative notions of infinity to broaden his perspective.

On November 6th 1978 Dr Tall flew to Israel to spend a month with Dr Vinner at the Hebrew University of Jerusalem, with visits planned to Tel-Aviv (to talk to Professor Fischbein) and Haifa.

In the first week he drafted a paper on "calculations and canonical elements" [10] which he had begun earlier with lan Stewart in Britain. This describes the schism that has arisen between the modem theory of equivalence relations and the classical art of computation with

representative (or "canonical") elements from equivalence classes. This proved to be relevant because of an earlier remark of Dr Stewart who had said "the trouble with non-standard analysis is that there aren't any canonical elements". In the field $\boldsymbol{R}(t)$ Dr Tall considered that there *was* a "natural" choice of canonical infinitesimal, the rational function $1/t$.

He also studied a paper [2] on mathematics education and history which mentioned infinitesimals, so they were at the forefront of his mind.

On the afternoon of Tuesday 7th November he brought up the topic of infinitesimals with Dr Vinner and his wife Hava (also a mathematician) and rehearsed the infinitesimal interpretation of $1/t$. The following week on Thursday 16th November he travelled to Tel-Aviv to talk to Professor Fischbein.

The confrontation

At lunch time on Thursday 16th November, Dr Tall discussed the research of Professor Fischbein and Dina Tirosh [1] on children's intuition of infinity. He tried to explain that, just because the children's intuition did not coincide with that of cardinal number, it did not mean that they were formally wrong. The cardinal explanation is that there are the same number of elements in AB as CD because the correspondence between the point P on AB, distance x from A, and Q on CD, distance $2x$ from C, is a one-one correspondence between the line-segments AB and CD.

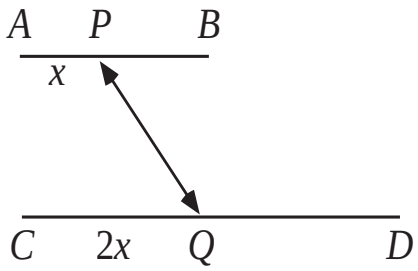


Figure 5

There is also a known 1-1 correspondence between AB and the square $WXYZ$ of Figure 2, so, in the cardinal number sense, there are as many points on the line segments as in the square. Many children in the experiments considered there were twice as many points in CD as AB and many more points in the square.

Dr Tall suggested that if AB has length ℓ and a "point" had infinitesimal size ε, there would be ℓ/ε points in AB and $(2\ell)/\varepsilon$ in CD, twice as many. Similarly there were $(\ell/\varepsilon)^2$ points in the square, far *more*

than in *AB*. In fact, for infinitesimal the ratio ℓ/ε is infinite, so there are *infinitely* many more points in the square than in the line segment.

Professor Fischbein was horrified at such an explanation and insisted on having a clear definition of the notions of infinitesimal and infinite elements. In fact, he wanted more, he wanted to be *shown* an infinitesimal explicitly.

Dr Tall slipped into his glib explanation that the graph of $1/t$ could be considered as an infinitesimal. There was an impasse. Professor Fischbein, as a psychologist, was not prepared to accept such an idea. A formal definition of an infinitesimal cut no ice with him. Dr Tall drew a picture something like Figure 4. He used this diagram in an attempt to demonstrate that for $a > 0$ the graph of $y=1/t$ is ultimately below $y=a$ for large enough values. Professor Fischbein was still not satisfied, he wished to see an infinitesimal as a *small quantity*, not as a graph. Dr Tall countered this by pointing out that any vertical line $x=k$ met $y=1/t$ above $y=0$ and below $y=a$ for sufficiently large k. (Figure 6.)

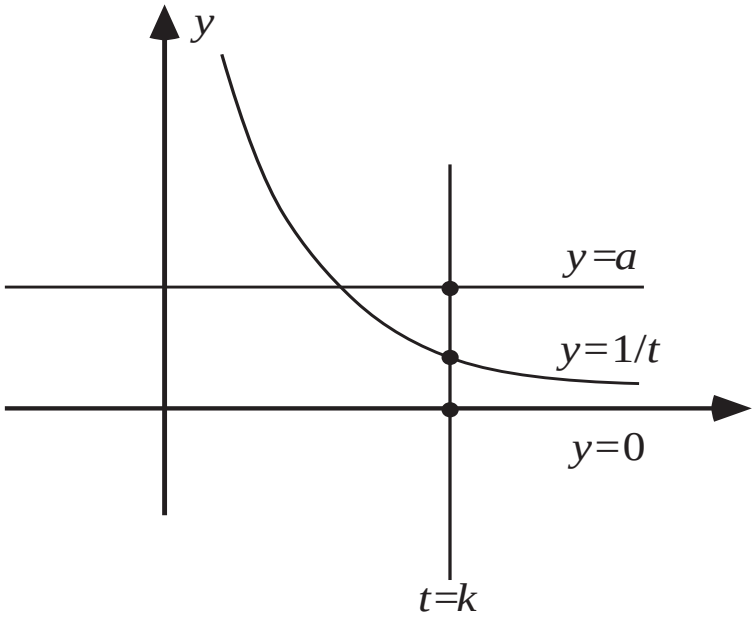


Figure 6

Of course, for smaller values of a, larger values of k are required. Dr Tall suggested that Professor Fischbein imagine a very large value of k; in fact to be able to handle *all* positive a, the best way would be to imagine a vertical line at infinity. Horizontal lines $y=a$ meet this line at a height a, and the graph $y=1/t$ meets the "line at infinity" at an infinitesimal height above the x-axis. (Figure 7.)

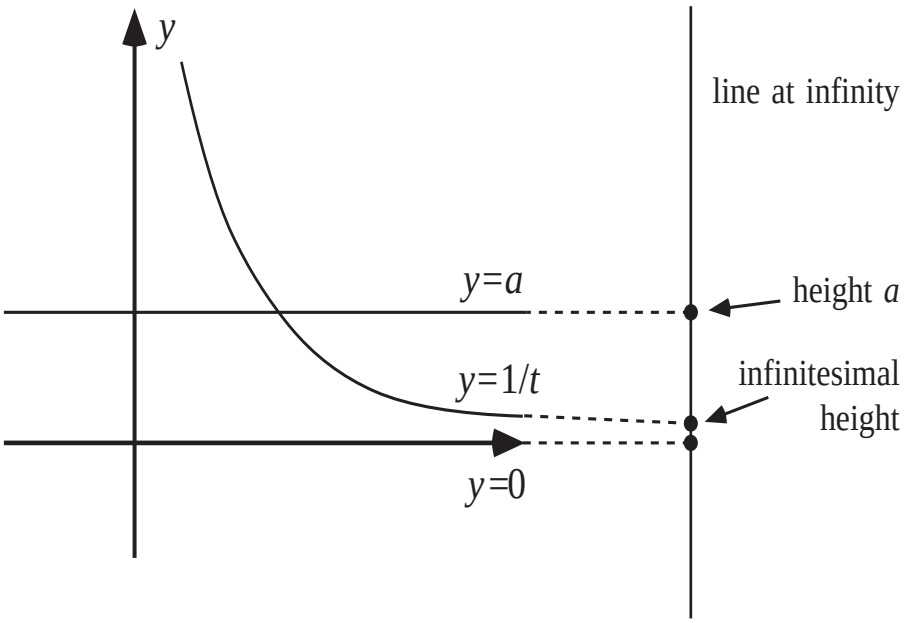


Figure 7

This did not satisfy Professor Fischbein either, but the idea had a profound effect on Dr Tall, who had never considered it before! Professor Fischbein's probing had forced him into describing figure 7 and the new insight intrigued him, forcing him to consider it in detail.

A fanciful idea and many "theorems"

Returning to Jerusalem at around 5:00 in the evening with darkness falling and a young soldier's gun near his left ear, he had a sudden inspiration. He let $\varepsilon=1/t$ and thought of $1/\varepsilon$ as an infinite point on the t-axis. He then realized that if he took his vertical line at infinity through the point $1/\varepsilon$, then this would meet the graph of $f(t)$ at a point which was at a height $f(1/\varepsilon)$ above the axis. In this way he found he had a direct correspondence between rational *functions* $f(t)$ and *points* on the vertical line through $1/\varepsilon$ where the graph meets the line. (Figure 8.)

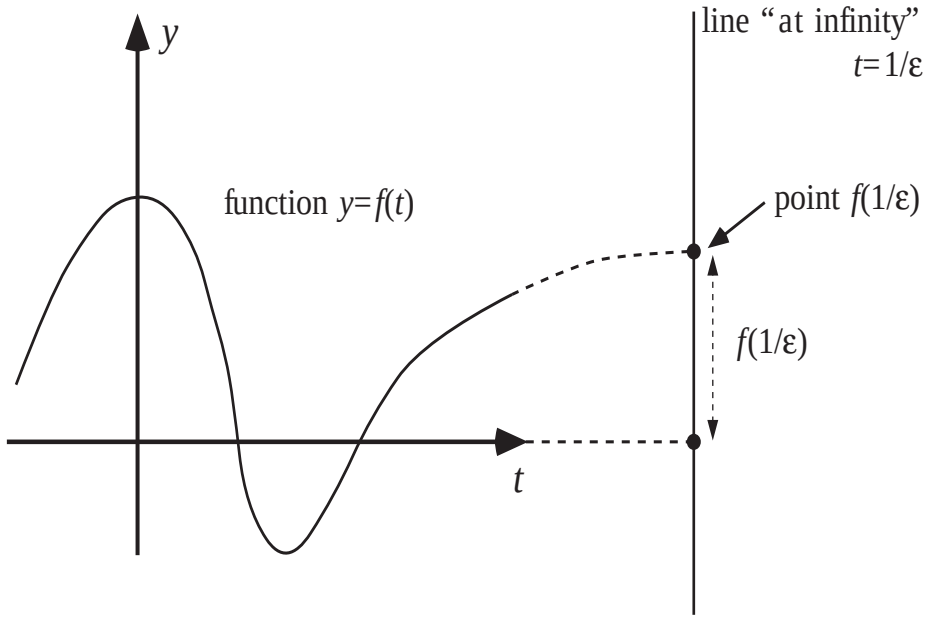


Figure 8

He arrived at the fanciful idea that an *infinitesimal* was a *point* on a certain line at *infinity*! This fascinated and excited him, spurring him to feverish activity. He still could only explain where *rational* functions met this line. Could he “complete” it so that all other functions would meet it in a sensible way? In particular, could he define f(ε) for the particular infinitesimal ε=1/*t* and any function *f*? He began with the function *f*(*x*) = sin*x* and feverishly scribbled on an old brown envelope. He considered

$$\sin\varepsilon = \varepsilon - \varepsilon^3/3! + \ldots$$
$$= \lim_{n\to\infty} s_n$$

where

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

Substituting ε=1/*t* he saw that sinε was a “limit” of elements in $\boldsymbol{R}(t)$.

Now his troubles began. He could only fitfully remember how to handle these kinds of limits (which he had studied fifteen years before as a postgraduate). He thought he should let I^n be the set of all multiples of ε^n and then the notion of convergence $s_n \to s$ should be:

“given *N*, there exists *M* such that $n>M$ implies $s_n - s \in I^N$.”

Similarly a “cauchy sequence” should be

“given *N*, there exists *M* such that $m,n >$M implies $s_m - s_n \in I^N$.”

(These concepts were later re-christened “super-convergence” and “super-cauchy sequence” because the terms became infinitesimally close to the limit or to each other.)

Meanwhile worries niggled at the back of his brain because of his inability to remember the theory accurately. (One worry, for instance, was because I^n was not an ideal in $\boldsymbol{R}(t)$, the latter being a field, and having no proper ideals. This worried him for days, until he realized he was confusing the roles of $\boldsymbol{R}(t)$ and the subring F of finite elements in $\boldsymbol{R}$(t). The subset I^N was an ideal in F!)

Back in his room at the Hebrew University where he was now based, he tried to consolidate his thoughts in writing.

Using the definitions he had just established, he let C be the set of (super-)cauchy sequences and N the subset of sequences which (super-) converged to zero. Following the standard completion process he defined his desired extension ring to be C/N.

At this stage he was elated at what he had done but nervous of the faulty details. However, rather than check the details, he returned to secure ground and writing up earlier results. He felt that his thoughts were unstable and he might lose the thread if he tried to sort out the difficulties. Far better to write down the basic material that he knew well to increase his security before attacking the frontier!

He outlined the story so far up to the definition of C/N (without checking the details). [His notation changed regularly at this stage. The field C/N would later be denoted by $\mathfrak{R}$. To simplify matters now all the various intermediate notations will also be denoted by this symbol.]

Now he began on the frontiers of his knowledge, noting a number of “theorems” without proof. He believed these to be “true”, with a great emotional involvement invested in their truth. [In many details they later proved to be inaccurate. To assist the reader, and to prevent this account growing overlong, four strands will be identified and followed through. It should be emphasised that this is a rationalisation of what happened in the sense that these strands were all inextricably interwoven.]

There were three themes that arose at this time which occupied him in the ensuing days and a fourth arose later.

Theme I concerned an attempt to characterise $\mathfrak{R}$ axiomatically, in a manner similar to the axioms for the real numbers. The real numbers are uniquely described as an archimedean, complete ordered field (where archimedean is equivalent to “there are no (non-zero) infinitesimals” and “complete” means “every cauchy sequence converges”). Dr Tall noted:

> *Theorem.* $\mathfrak{R}$ is a non-archimedean (super)-cauchy complete ordered field.

Believing that he had a working definition for his field with infinitesimals he moved on to the question of extending real functions to have a meaning over $\Re$. He concentrated on power series functions $f(x) = \Sigma a_n x^n$ and wrote:

> *Theorem* If $\Sigma a_n x^n$ is (a real power series) convergent for $|x|<K$, then $\Sigma a_n x^n$ is convergent for $x \in \Re$, $|x|<K$.

He believed that this theorem (once proved) would allow him to extend $f(x) = \Sigma a_n x^n$ to take on values in $\Re$ for $|x| < K$ and, taking $x = a+\varepsilon$ for infinitesimal ε, he would then be able to attack differentiation as described.

The third theme was a growing interest in the structure of $\Re$. The hyperreals of Robinson's theory include a set of "hyperintegers" which extend the properties of the integers. By analogy he felt that $\Re$ must include a concept of "integer". He toyed around with some computations, noting that a finite element of $\Re$ is infinitesimally close to the real number which is its standard part, so every finite element of $\Re$ must lie between integers n and $n+1$. Given any infinite element $\alpha \in \Re$, he divided by a power of the infinite element α to get a finite element α/t^n which must then lie between integers k and $k+1$. Hence α lies between kt^n and $(k+1)t^n$. This delighted him and he concluded that elements of the form kt^m (for integers k, m) acted like "integers" in $\Re$.

(Other ideas noted included an infinitesimal definition of continuity following Robinson's theory and a first attempt at integration using thin elemental strips of equal infinitesimal width. These played no significant part in the next few days, so they are omitted.)

A fourth theme, involving a reassessment of the original definition of order in the field $\boldsymbol{R}(t)$ will arise later. His pages of notes and theorems pleased him. He relaxed, satisfied.

A second day of activity

It was on Friday, the following day, that Dr Tall realised it might be of interest for him to write down his feelings and conscious thought processes in addition to the mathematical notes he had to date. This proved to be a harrowing experience, for in writing down his thoughts his mind was faster than his pen and he became confused trying to grasp the mental processes as they flew by. As well as "blow by blow" action notes he found it necessary to sit down in the evening and summarise the events of the day. At 6:30 in the evening he recalled the day's activity as follows, beginning with the period in the morning:

> "I recall that my mind was buzzing with ideas – I still wasn't clear about the archimedean bit, nor completeness. However I spent an hour photocopying music, including "Virginia, don't go too far" (a Gershwin song). I thought about the hyperreals of Robinson "going too far":

extending too many functions. After coffee I wanted to work but the tension was unbearable, so I read a novel about Jerusalem which I'd been reading over the last few days. As I read, mathematical ideas floated past me, I couldn't seem to grab them, but the shape of $\mathfrak{R}$ was becoming clearer.

In retrospect (6 hours later) I can't remember what the ideas were. But at lunch I sat down and during the meal, the name "superreals" came to me. The "rational functions" in $\boldsymbol{R}(\varepsilon)$ were "superrationals". I realised that any superreal could be written as

$$a_{-n}\varepsilon^{-n} + a_{-1}\varepsilon^{-1} + a_0 + a_1\varepsilon + a_2\varepsilon^2 + \ldots$$

i.e. as an infinite "epsilonicimal"

$$a_{-n}\ldots a_{-1}a_0 \cdot a_1a_2\ldots$$

(This was by analogy with an infinite decimal

$$b_{-n}10^{-n} + b_{-1}10^{-1} + b_0 + b_1 10 + b_2 10^2 + \ldots$$

which stands for

$$b_{-n}\ldots b_{-1}b_0 \cdot b_1b_2\ldots)$$

His notes continued:

… I toyed with the idea of using δ instead of ε and calling the expression a "deltacimal", then decided on "epsimal". I was suddenly sure that the superrationals were "eventually repeating epsimals". If

$$r = \frac{a_m\varepsilon^m + \ldots + a_0}{b_n\varepsilon^n + \ldots + b_0}$$

I could see the long division process of dividing $b_n\varepsilon^n + \ldots + b_0$ into $a_m\varepsilon^m + \ldots + a_0$ would eventually repeat giving a (super)convergent *rep*eating epsimal.

EXCITEMENT !

I mused about the "superintegers" again (for this is what I had called the "integers" in $\mathfrak{R}$ and decided they were of the form $a_{-n}\ldots a_{-1}a_0 \cdot 000\ldots$ (where $a_i \in \boldsymbol{Z}$), so they depended on the choice of ε This didn't worry me too much because I felt sure that any two sets of superintegers were order isomorphic.

He then reviewed the "blow by blow" notes that he had written that afternoon as he was actually doing the research work. After the above mentioned ideas had occurred to him he could not stand the tension and had taken a bus into Jerusalem. There he wandered about the streets with his mind leaping about excitedly like a butterfly. His notes taken at the time were a garbled mixture of travelogue and mathematics which might make little sense to the reader. However, they may be summarised (and rationalised!) as follows.

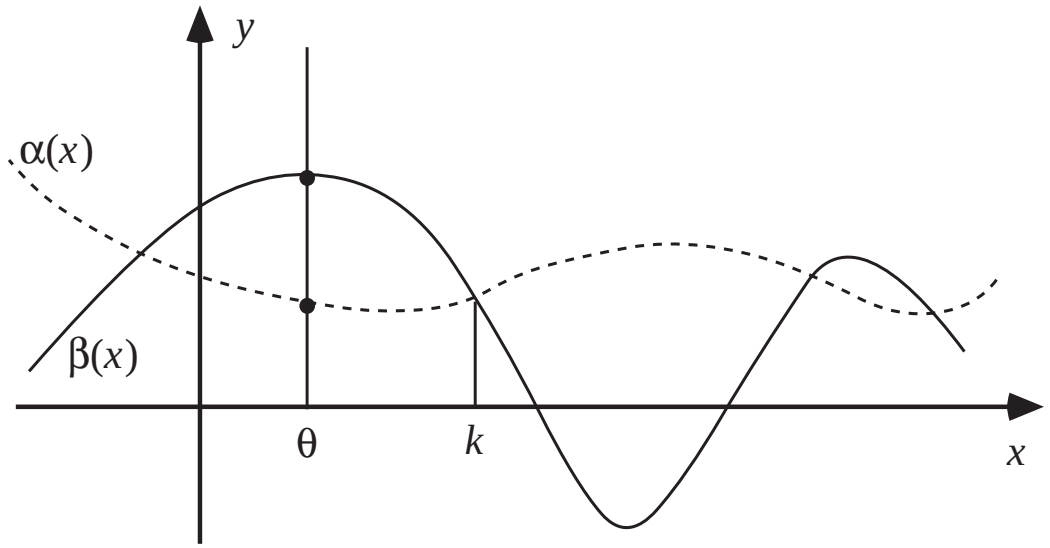
His first theme (to characterise $\mathfrak{R}$ axiomatically) became interwoven with his new thoughts on the structure of $\mathfrak{R}$. (His new words "superreal", "superrational", "superinteger" intrigued him so much that they took on a life of their own and suggested properties analogous to

the corresponding ordinary concepts.) In ordinary analysis the archimedean property can be characterised as "the integers are not bounded above in the reals" so he toyed with the idea that his "superintegers" were not bounded above in the superreals. He was also growing concerned about a fourth theme: the growing conviction that he would need to rephrase his original explanation of the order on $\boldsymbol{R}(t)$. He realised that he was only stressing *infinitesimal* elements in his theory of calculus whilst his example emphasised the *infinite* nature of t. Substituting $x = 1/t$ he reformulated the definition in terms of x, at first haltingly, but later smoothly in the following form:

> If $\alpha(x)$, $\beta(x)$ are rational functions in x, then we say
> $\alpha(x) < \beta(x)$
> if for some real $k > 0$, $\alpha(\theta) < \beta(\theta)$ for all real θ in $0 < \theta < k$.

(Figure 9.)

(For further details the reader may consult [12] or [15].)



$\alpha(x)<\beta(x)$ because, for some $k>0$, $\alpha(\theta)<\beta(\theta)$ for $0<\theta<k$

Figure 9

He realised that putting $x = \varepsilon$ in this version gives a 1-1 correspondence between $\boldsymbol{R}(x)$ (as rational functions) and $\boldsymbol{R}(\varepsilon)$ (as points on the vertical line $x=\varepsilon$) where the graph of the rational function meets the vertical line. (Figure 10.)

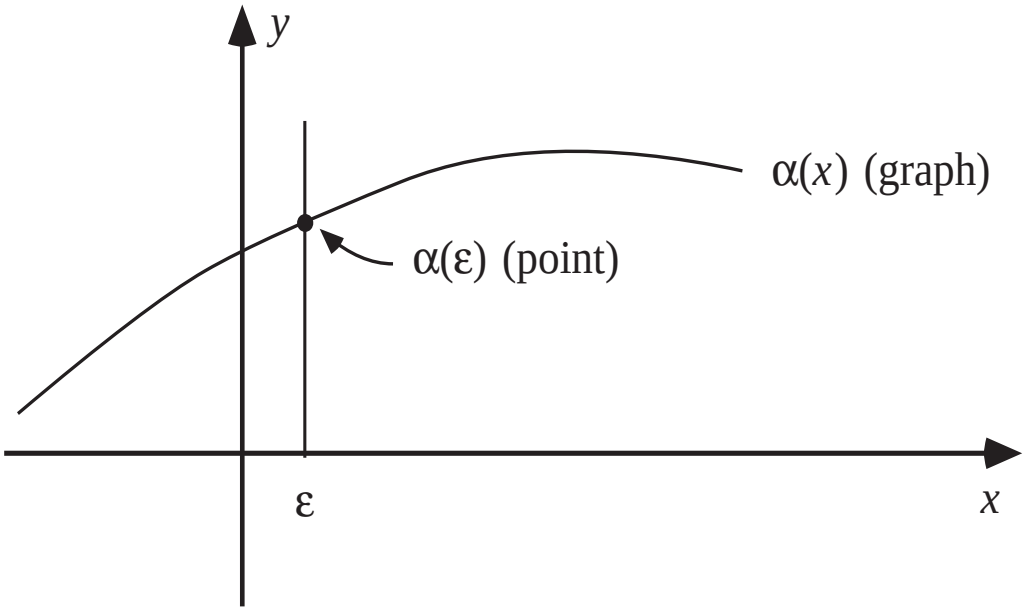


Figure 10

(He saw that the completion processes for $\boldsymbol{R}(x)$ and $\boldsymbol{R}(\varepsilon)$ must correspond, but he hadn't sorted these ideas out precisely, he just saw some dim vision.) At this time he relaxed and had a snack.

Relaxation and conflict

After his eventful walk round Jerusalem and his session recalling the events in writing, as given in the last section, he relaxed and prepared to go out to dinner. At 8:30 he was just about to leave his room at the university to go to Dr Vinner's flat when he had a strange vision which is difficult to explain to the reader. He imagined the set F of finite elements and visualised various "infinite levels" $(1/\varepsilon)F$, $(1/\varepsilon^2)F$ obtained by multiplying by successive powers of the infinite element $(1/\varepsilon)$. Perhaps increasing sequences in one level would be bounded above (by the next level) and yet the sequence not tend to a limit. Would he need to have more elements "between" the levels? Was his definition of the superreals adequate? Walking to Dr Vinner's flat, he thought of power series in ε^2. Putting $\delta = \varepsilon^2$, power series in δ were an isomorphic copy of those in ε, but the former "levels" F, $(1/\delta)F$, $(1/\delta^2)F$, ...were really F, $(1/\varepsilon^2)F$, $(1/\varepsilon^4)F$,...and these had *many* elements missing between them, namely $(1/\varepsilon)F$, $(1/\varepsilon^3)F$, … .

The meaning of this obscure vision was not really clear to him and at dinner the research was not discussed. Dr Tall needed more time to think things out.

A day of mixed blessings

On Saturday morning at 8:20 he was cleaning his teeth, thinking seemingly about nothing in particular when it hit him. His ideas about the "superintegers" were all wrong! To be useful, every element of $\mathfrak{R}$ needed to be between superintegers k and k+1, but what about l/2ε? Another thought struck him. He had expected to have "superintegers" by analogy with the "hyperintegers" of Robinson's theory. But Robinson's theory was a more all-embracing theory. It allowed one to consider sequences as functions on the natural numbers and extend them to functions on the hypernatural numbers (positive hyperintegers). His simpler theory only had a hope of extending power series, which were defined on open intervals. It was totally unreasonable to expect that sequences could be extended in his theory. It was therefore asking a lot to expect his theory to have any "superintegers". *There were no superintegers.* He ruefully recalled his Gershwin song, "Virginia, don't go too far!" Yet his feelings were not all bad. His long struggle with the "superintegers" had come to nothing, the concept had no merit. But Robinson's hyperintegers needed something like the axiom of choice to construct them. He should not have expected to get such a rich notion from a simple algebraic construction like the superreals. To get "everything" in the superreals was a logical impossibility and now he had found something—the superintegers—which were not possible. What was more, he had a heuristic "explanation" as to why they would not work. He felt a strange mixture of disappointment at his failure and elation because he felt he knew that he had failed. The remainder of the Sabbath he relaxed, playing the piano and reading. En route to visit a friend he had various thoughts about the superreals but was unable to recall anything precise later on.

The activity slows down

On Sunday morning, being unable to recall much of the thoughts he had had on the way to his friend's house, he looked at the notes from earlier and mused about his theme of characterising the superreals $\mathfrak{R}$ axiomatically. Recalling that power series in ε^2 were isomorphic to those in ε, he realised he may have to specify a "first order" infinitesimal ε to make sure he didn't have elements "missing" between the "levels". Matters were now getting complicated. He didn't really understand what he was talking about and he was pressed for time to prepare a seminar for that afternoon (Sunday being a normal working day in Israel). He spent most of the time preparing the seminar and his concern over the confused axiomatic description of $\mathfrak{R}$ faded. He made some brief notes, including the following:

> "I'm not in the same state of super-tension about the superreals as I've decided to call them. I've many other things to do, especially a lecture to write. I feel relaxed about it: it can be left for a while. Other things are more pressing."

Interlude in Haifa

On Monday morning (5:30 a.m.) Dr Tall travelled to Haifa where he lectured in the afternoon at the University and later explained his ideas to his host, Dr Nesher. Whilst explaining that the superrationals were "repeating epsimals", and hence analogues of rational repeating decimals, warning bells seemed to ring at the back of his mind. He sensed that something may be wrong but, out of embarrassment, said nothing. For the next few days he had a full itinerary and apart from a few ideas about integration (prompted by a lecture he gave at the Technion) he thought no more about the superreals, or about the possible problem with repeating epsimals. On Friday he returned to Jerusalem. He had no feeling of excitation over the superreal theory and concentrated on things other than mathematics.

First draft of the theory

In the late afternoon on Sunday, in an easy atmosphere, he began writing the first draft of a paper on the superreals. (Though on earlier occasions he had written down what he knew when frontier activity became too tense.) Now he was relaxed and, although there were several ideas to sort out still, he felt that the basis was there and he should establish the foundations of the subject. He planned the paper in the following sections:

I Introduction
2 Ordered extension fields of $\boldsymbol{R}$
3 The superrationals
4 Extending rational functions to the superrationals
5 Superconvergence and supercauchy sequences
6 Constructing the superreals
7 Extending analytic functions to the superreals
8 Continuity
9 Differentiation
10 Integration
11 Broader horizons.

Sections l and 2 flowed easily. The introduction contained a brief outline of intended content and, when he mentioned the superreals, for the first time he noted that each non-zero superreal $\alpha = a_r\varepsilon^r + a_{r+1}\varepsilon^{r+1} + \ldots$ $(\alpha_r \neq 0)$ had a specified *order* $o(\alpha) = r$. He had this idea in embryo from the start, but had never written it down. In writing the paper he now decided to define superconvergence using the order instead of his original idea. The sequence s_n was to be superconvergent to the limit s if,

given M, there exists N such that $n \geq N \Rightarrow o(s_n - s) \geq M$.

Similarly a supercauchy sequence (s_n) satisfied

given M, there exists N such that $m,n \geq N \Rightarrow o(s_n - s_n) \geq M$.

Section 2 was a précis of straightforward material from [5] then, in section 3, he hit his first obstacle. He wasn't sure how to present the definition of superrationals as rational functions $\boldsymbol{R}(t)$. Should he define t to be infinite, and use his fanciful inspiration that an infinitesimal was a point on the line at infinity, or should he use the version with t infinitesimal, which he was sadly beginning to realise might be more practical? His first version was a mishmash, including both!

Another blow

Working on the superrationals again recalled his worries over "repeating epsimals" which he had not considered for over a week. He now realised that long division of one polynomial into another might not repeat. After briefly thinking he could handle the problem if the denominator was only first degree, he flirted with the idea of introducing the complex numbers so that all denominators could be resolved into linear factors and the division could be reduced using partial fractions. Then he realised that his analogy between rationals and superrationals was faulty. In dividing one natural number q into another p to compute the fraction p/q, the remainder at each stage would be less than q and so only a finite number of remainders were possible, leading to a repeating decimal. In dividing a polynomial q into a polynomial p to compute a rational function p/q, all one could say at each stage was that the remainder had *degree* less than that of q, and there were an infinite number of possibilities in general.

Now the superrationals were dying on him in the same way that he had lost the superintegers. He realised that $1/(1+ a_0\varepsilon + \ldots + a_n\varepsilon^n)$ could be written as $1/(1-\delta)$ where δ was an infinitesimal $-a_0\varepsilon - \ldots - a_n\varepsilon^n$ and $1/(1-\delta) = 1+\delta+\delta^2+\ldots$ was a repeating δ-mal. But this was little consolation. A beautiful analogy had failed.

A fear

Now a worse fear hit him: not only might the theory be trivial, it might also be well-known! He had remembered some work from his graduate days. He wrote:

> Thinking about the "order $o(\alpha)$ of an infinitesimal α", I began dimly to remember about valuation rings. I felt disappointed that the whole theory might be known. I recalled the quotation "when the spring comes the violets grow on all sides of the hill" and imagined many people in different parts of the world rushing to get the theory published. Or worse, the theory might already be gathering dust in long-forgotten tomes. He did not wish to speak to anyone until he had worked the ideas out fully for himself.

[He later discovered that the mathematical theory was already well-known although his visual insights would play a useful role in cognitive theory.]

A fresh start and a pleasing discovery

The next morning he awoke refreshed and yet again began writing the paper from the beginning. The first three sections now came much easier. Then he became side-tracked by the problem that he had not yet resolved how to define $f(\alpha+\varepsilon)$ for infinitesimal ε! It would have to be solved sooner or later. In a state of tension again, he had a bath and mused about all sorts of things freely. He did not sort out $f(\alpha+\varepsilon)$, instead something entirely unexpected hit him. After his bath he wrote down that rational functions $\boldsymbol{R}(\delta)$ in an element $\delta \in \boldsymbol{R}(\varepsilon)$ need not generate the whole of $\boldsymbol{R}(\varepsilon)$. Even if is first order, say $\delta = \varepsilon + \varepsilon^2$, then $\boldsymbol{R}(\delta)$ is only a proper subset of $\boldsymbol{R}(\varepsilon)$. But the situation with power series is different. For instance, the equation $\varepsilon^2 + \varepsilon = \delta$ could be solved in the superreals to give

$$\varepsilon = -\tfrac{1}{2} \pm \tfrac{1}{2}(1+4\delta)^{\frac{1}{2}}$$

and the latter could be expanded using the binomial theorem to express ε as a power series in δ. He became greatly excited, for he believed that *any* first order infinitesimal $\delta = \sum_{n=1}^{\infty} a_n \varepsilon^n$ could be manipulated to express δ as a power series in ε. This meant that although he was using one particular infinitesimal ε to construct the superreals, having done so, all the first order elements were as good as each order. Recalling Ian Stewart's remark about canonical elements, he realised that the element ε was just a *canonical* choice out of the collection of first order elements which *all* had equal status. Even though he used the specific element ε to construct $\mathfrak{R}$, it was not so special after all. He did not really *need* an axiomatic description of $\mathfrak{R}$. It simply consisted of power series in a first order infinitesimal of the form

$a_{-n}\varepsilon^{-n} + \ldots + a_{-1}\varepsilon^{-1} + a_0 + a_1\varepsilon + a_2\varepsilon^2 + \ldots$

and *any* first order infinitesimal would do. His problem of finding an axiomatic description of the superreals did not get solved, it simply … evaporated.

Another detail to sort out

Of the four themes mentioned earlier, one, the axiomatic description for $\Re$ had become unimportant to him; another, the structure of the superreals, in terms of superrationals and superintegers had become emasculated; the later question of describing $\boldsymbol{R}(t)$ with t infinite or infinitesimal was in the throes of being decided. Only one theme remained intractable – the details of the definition of $f(\alpha+\varepsilon)$ for infinitesimal ε.

In terms of the previous discussion, it was now clear that a power series in any infinitesimal δ was a superreal. Thus an analytic function

$$f(x) = \sum a_n x^n$$

gave rise to a superreal $f(\delta) = \sum a_n \delta^n$ for any infinitesimal δ. But what about other values of $f(\alpha)$ for superreals α less than the radius of convergence of the power series? Dr Tall felt matters might go astray near the radius of convergence. He considered the particular example

$$f(x) = \sum (-1)^n x^n / n \quad (|x|<1).$$

Rearranging the terms of $f(x+\delta)$ as

$$\begin{aligned} f(x+\delta) &= \sum (-1)^n (x+\delta)^n / n \\ &= \sum (-1)^n (x/n + x^{n-1}\delta + \ldots), \end{aligned}$$

he took $x+\delta = 1-\varepsilon$ and found that the coefficient of ε in the power series became

$$1 - 1 + 1 - 1 \ldots .$$

This bothered him. In Robinson's non-standard analysis, if a function f was defined for real x satisfying $|x| < K$, then it extended to a function for hyperreal x satisfying $|x| < K$. His counterexample showed that it would not work for superreal theory.

With the question still open he had to leave the work as his wife arrived in Israel for a week's stay. His spare time was now limited to occasional short periods and he decided (yet again) to begin rewriting the article, hoping that inspiration would strike him about how to extend functions before he reached that section of the paper.

Technical complications

On this further rewrite there were still technical problems to overcome. His treatment of the order on $\boldsymbol{R}(t)$ remained a mishmash of t as an infinite element side by side with t as an infinitesimal. He could not bear to give up his very first revelation of an infinitesimal on the line at infinity, even though he now realised that he might have to.

He hit the problem of extending functions again. He tried a completely different approach using functor theory, but that proved irrelevant and was later discarded. Then a technical complication hit him straight in the face. If he wished to define $f(x+\varepsilon)$ simply as

$$f(x+\varepsilon)=\sum a_n(x+\varepsilon)^n,$$

then the partial sums

$$s_n=\sum_{r=0}^{n} a_r(x+\varepsilon)^r$$

did not form a supercauchy sequence (because the difference s_n-s_m was not necessarily infinitesimal for large m,n).

He toyed with alternative concepts of convergence, blending superconvergence with ordinary convergence in $\boldsymbol{R}$ [which we omit here] and eventually left the matter unresolved once more.

In a state of disarray, with a partly conceived theory and a number of outstanding technical difficulties, his period of research in Israel came to an end.

Time out

The end of his time in Israel coincided with a temporary halt in his activities. He returned to England on December 5th to find a mountain of mail. He finished typing up the paper “Calculations and canonical elements” because he wanted something tangible to show for his month in Israel. (It was later published. [10]) The month of December was fraught with illness (returning to British damp and cold). He discussed his infinitesimal ideas with colleagues at Warwick but wrote no more.

The theory matures

On January 6th he flew with his family to Montreal for a term at Concordia University. It was all new, with many arrangements to be made and two courses to teach. In January and early February he put aside Tuesdays and Fridays to complete the paper. Once more he started writing at the beginning.

This time his ideas were more settled. He now eliminated his original conception of $\boldsymbol{R}(t)$ with t infinite, regretfully losing the idea which caused the initial breakthrough. His fanciful idea that an infinitesimal was a point on a line at infinity has never featured in print (except in

this paper which explains how it happened). He used a notion of infinitesimal microscope from [5] to be able to "see" infinitesimals. (At this stage he simply copied the ideas of [5], as he published them in [15], but he later adapted them for publication in [12].)

The problem of $f(x+\varepsilon)=\sum a_n(x+\varepsilon)^n$ was simply side-stepped. Although he had worked out a different theory of convergence to cover it, this was never published. Instead he expanded $f(a+\varepsilon)$ as a power series about $x= a$,

$$f(a+h) =\Sigma b_n h^n$$

and then put $h = \varepsilon$ to get a superreal number $\Sigma a_n \varepsilon^n$.

He even generalised this slightly to the case where f might have a pole:

$$f(x+h)=b_{-m}h^{-m}+\ldots+b_{-1}h^{-1}+\sum_{n=0}^{\infty}b_n h^n .$$

This also is clearly a superreal when one substitutes $h=\varepsilon$.

Finally integration was handled using area functions, by analogy with [5].

The paper was completed on February 12th. It was turgid and ugly and remains unpublished [11]. Subsequent papers were better; [12], [13], [14] and have appeared in print. In the ensuing months he refined the notion of integration and arc-length in a form eminently suited to the published theory in [12].

Leibniz

Dr Tall realised the importance of the notion of "order of an infinitesimal" in the history of pre-nineteenth century calculus while reading a paper of Lakatos. [6] For example, in Leibniz's theory, the differential of a product is written

$$d(uv) = vdu + udv.$$

Modern commentators note that

$$d(uv) = vdu + udv + dudv$$

and refuse to neglect the $dudv$ term. However, if u,v are finite, du is an infinitesimal of order m, and dv an infinitesimal of order n, then $dudv$ is of order $m+n$ which is infinitesimally small in size compared with $vdu + udv$, and may therefore be neglected. He realised that in non-standard analysis infinitesimals do not have specified orders, and so the superreals were a *closer* match for Leibniz's calculus than non-standard analysis [13].

Infinite measuring numbers

At last, in June 1979, there came a time when he was due to meet Efraim Fischbein once more and he returned to contemplating the original questions. Thinking of a point marked with a "pencil of

infinitesimal width ε", the number of such points needed to fill in a line *AB* of finite length ℓ is ℓ/ε which is, of course, infinite. To fill a line *CD* of length 2ℓ requires precisely twice as many, $(2\ell)/\varepsilon$. The "cardinal number argument" mapping *P*, distance *x* from *A*, onto *Q*, distance 2*x* from *C* (Figure 5) is exposed! It *doubles* the scale, so a mark of width ε on *AB* is stretched into one of width 2ε on *CD*. If "points" on *CD* have width 2ε, the number in *CD* is $(2\ell)/(2\varepsilon)$ which is, of course ℓ/ε, the same number as in *AB*. The only way that *AB* and *CD* can have the same (measuring) number of points is if those in *CD* are twice as big as those in *AB*!

If the line *AB* is itself drawn with a theoretical pencil width ε, then the square *WXYZ* is made up of horizontal lines width ε. There are ℓ/ε of them and each one contains ℓ/ε points. The total number of points is $(\ell/\varepsilon)^2$ which is an infinite number of *higher order* than ℓ/ε. To be precise there are ℓ/ε times as many points in *WXYZ* as in *AB*. In this sense there are infinitely many more points in *WXYZ* than in *AB*! These ideas were published in [14].

Reflections

Reconsidering the theory as a whole, it now all seems so inevitable. These ideas were not *invented,* they were *discovered*. Reading about the process of discovery written in these pages, it is amazing to see the number of errors made and the false intuitions which had the ring of truth. Yet such was the intensity of excitement at the time that these temporary setbacks were insufficient to cause permanent blockages.

Post-rationalising the discovery, it is clear that the researcher already had most components deep in his psyche. The break-through came, after several years of interest and lack of understanding of infinitesimals, when a psychologist asked for a description that could be properly understood. That description was not satisfactorily given at the time, but it provoked in the researcher the chain of events here recorded.

The realisation that an infinitesimal could be geometrically imagined as a "point on a line at infinity" was so stimulating that he forgot the psychologist's problem and turned to a more deep-seated quest for a simple theory of calculus with infinitesimals. He often got bogged down with technicalities which did not figure in his final publication.

On the other hand, various non-mathematical resonances in his mind amused him ("rational" functions being thought of as "superrational" numbers; a Gershwin song, "don't go too far"). Many "breaks-through" occurred at a subconscious level. Positive breaks-through were accompanied by pleasurable feelings, though actual verification (or realisation of error) took longer. Negative breaks-through usually

occurred, first as a vague feeling of unease, with the conscious rationalisation of the error sometimes taking days or even months to register. (The folly of the "super integers" occurred overnight, the fact that "superrationals" were not "repeating epsimals" took six days from a feeling of unease to a formal understanding, the fears concerning the correct extension of a function were not sorted out for several months.) That is not to suggest that subconscious activity was going on for all the intervening period, but there were times of intense mental activity of some kind which showed that *something* was going on. Often it was an external input or the act of explaining ideas to someone else that triggered off this internal activity once more. But it continued for a time after external impulses had ceased: it even seemed preferable on occasion to have a conscious distraction (reading a book) so as not to disturb the intensity of subconscious thought processes by conscious probing into the ideas. (A very strange feeling indeed!)

The mind did not seem at all happy to be on the frontiers all the time. A strong resonance was a boost to maintaining frontier thought (even if it later proved to be fallacious) but mental conflicts and unease at too high a level provoked withdrawal into secure regions. One only needs to look at the number of times the rewriting of the paper returned to the beginning again to see the extent of the insecurity.

It was nearly a year, until an impending meeting with Efraim Fischbein, that the original problem of the questionnaire was satisfactorily resolved.

A classic description of "problem-solving" involves "conjectures" which are then checked out. Here the researcher never felt that he made "conjectures"; what he saw were "truths" evidenced by strong resonances in his mind. Even though they often later proved to be false, at the time he felt much emotion vested in their truth. These were no cold, considered possibilities, they were intense, intuitive certainties. Yet at the same time his contact with them often seemed tenuous and transient; initially he had to write them down, even though they might be imperfect, before they vanished like ghosts in the night.

When such "truths" later proved false, it was rarely because of a coolly considered counter-example. That usually came later still after a period of mental unease already mentioned. In fact the researcher, when in a state of mental excitement, *did not wish to check the detail at all,* lest he lose the thread of the overall idea. It is remarkable the number of times that there were small errors which went unnoticed at the time but later produced unease, then correction.

These events fit very well with the descriptions given of the classic sequence of activities in research given by Poincaré, and reproduced in [3]. Poincaré reports four basic stages: preparation, incubation,

illumination, verification. The only difference here is that so many facets are being intertwined together that one stage in one part of the theory occurs simultaneously with a different stage in another part. Poincaré suggests that, during a period of incubation, it is the *aesthetic* manner in which ideas fit together which cause them to surface to the conscious mind. I would put it in a more mundane manner. Brain activity is an electrochemical phenomenon, which I believe to work through electrical resonances. I conjecture that it is the strength of subconscious resonances that cause them to surface. The emotions tell us the state of the brain: pleasure with strong resonances, unease with conflicting resonances. I also conjecture that these emotions are aroused by the physiological conditions of the brain and that the reason why we cannot pinpoint the feeling of unease is that there is only a physiological dissonance at this time; there is not necessarily a subconscious *formal* understanding at this stage at all. That comes about by a much slower (chemical?) change in the nature of the resonating circuits, hence the time lapse.

Having gone through the tortuous thought processes outlined in this paper, a process of post-rationalisation takes place. History is mentally re-written. For instance, I was amazed nearly a year later to realise that I had defined "super-cauchy" sequences using the ideal I^n. I subsequently had come to believe that I had worked with the "order of an element" from the start. It is for this reason that I am very suspicious of mathematicians who recall how they did research without taking careful notes at the time. We forget the twists and the minor errors, but we remember the pleasure of our successes and the embarrassment of our major mistakes. This is why I wished to set out the story described here, warts and all.

A strange thing has now happened. I remember some of the pleasures (and a few uneasy times) in the development of these ideas, but now the theory has a life of its own. There is no doubt any more that it is correct. Small details may still be astray, but the whole system has a ring of truth about it. Despite the roundabout route to its conception, other mathematicians could (and have) come independently to the same ideas. It now seems an independent platonic entity, quite separate from the individual mind that fumbled to come to grips with it, a piece of mathematics, a corporate property of mathematicians at large.

References

[1] E. Fischbein, D. Tirosh and P. Hess, "The intuition of infinity", *Educational Studies in Mathematics.*10, 3–40, (1979).

[2] I. Grattan-Guiness, “On the relevance of the history of mathematics to mathematical education”. *Int. J. Math. Ed. Sci. Technol.*, 9 (3), 275–285, (1978).
[3] J. Hadamard, *The Psychology of Mathematical Invention*. Dover, 1945.
[4] T. Kuhn, *The Structure of Scientific Revolutions*. Chicago, 1962.
[5] J. Keisler, *Foundations of Infinitesimal Calculus*. Prindle, Weber and Schmidt, 1976.
[6] I. Lakatos, “Cauchy and the continuum”. *Mathematical Intelligencer*, 2,151–161 (1978).
[7] A. Robinson, *Non-standard Analysis*. North Holland, 1966.
[8] R. L. E. Schwarzenberger and D. O. Tall, “Conflicts in the learning of real numbers and limits”, *Mathematics Teaching*, 82 (1978).
[9] I. N. Stewart and D. O. Tall, *Foundations of Mathematics*. Oxford, 1977.
[10] I. N. Stewart and D. O. Tall, “Calculations and canonical elements”. *Mathematics in School*, 8 (4), 2–5; 8 (5), 5–7, (1979).
[1l] D. O. Tall, *Standard infinitesimal calculus using the superreal numbers* (unpublished notes).
[12] D. O. Tall, “Looking at graphs through infinitesimal microscopes, windows and telescopes”. *Mathematical Gazette*, 64 22–49, (1980).
[13] D. O. Tall, “The calculus of Leibniz – a modern approach”. *The Mathematical Intelligence*r, 2 (1) 54–5, (1979).
[14] D. O. Tall, “The concept of infinite measuring number and its relevance in the intuition of infinity”. *Educational Studies in Mathematics*, 11, 271–284, (1980)..
[15] D. O. Tall, “Infinitesimals constructed algebraically and interpreted geometrically”. *Mathematical Education for Teaching* 4 (1), 34–53, (1981*)*.