INTUITIONS OF INFINITY

by David Tall, Mathematics Education Research Centre, University of Warwick

What is infinity? It is an extrapolation of our finite experience. As such our intuitions of infinity depend very much on the kind of experience which is extrapolated. For instance, if we consider the "samesize" property of sets which have a bijection* between them then we get the kind of infinity called a *cardinal number*. This is the most widely considered notion of infinity amongst mathematicians yet it has properties which usually seem most unintuitive to the uninitiated. If we look at the experiences of pupils in school, they rarely concern the "samesize" aspect of sets so this lack of intuitive appeal is hardly surprising. Secondary pupils have different experiences of infinity which give them a totally different feel for the idea.

For instance, in drawing the graph of

$$f(x) = 1/(1-x)$$

they find that as x approaches 1 from the left, $f(x)$ gets very large and positive but as x approaches 1 from the right, $f(x)$ gets very large and negative. The expression

$$(x^2+4)/(2x^3+3)$$

approaches 0 as x grows large, so although x^2+4 and $2x^3+3$ both grow without limit as x gets large, $2x^3+3$ grows even faster than x^2+4.

How can one reconcile these two examples? In the first case $f(x)$ tends to a "plus infinity" from one side and a "minus infinity" from the other. Is "plus infinity" the same as "minus infinity"? In the second case if x^2+4 tends to "plus infinity" as x grows large, how can $2x^3+3$ tend to something even larger? Can we have something that is even bigger than infinity?

One thing is certain. Talking about cardinal infinity, the "samesize" of sets, will not resolve this problem in a natural way. These ideas of "growing large", "tending to infinity" and so on have absolutely nothing to do with comparing the size of sets.

As another example, consider that hoary character the decimal expansion

$$0.999\ldots999\ldots$$

"nought point nine recurring". Does it equal one, or is it just less? Ask school pupils this question and the vast majority will say "less". I suspect that the majority of teachers would say the same. Mathematicians assert that this decimal expansion *equals* one. For them the decimal expansion means the actual real number which is the *limit* of the decimal approximations. The decimal expansion

$$0.333\ldots333\ldots$$

equals a third. (On that matter most pupils and teachers might agree.) Multiplying this expression by three gives the proverbial "nought point nine recurring" which therefore should equal one.

* *A bijection is a one-one, onto function.*

But for many this leaves a bitter taste in the mouth. Something has gone wrong somewhere. How can "nought point nine recurring" equal one? Surely when we subtract it from one we get a single digit left over, way off in the "infinitieth place":

$$1-0.999\ldots999\ldots=0.000\ldots000\ldots\ldots\ 1.$$

These problems are usually explained away by quoting the current mathematical dogma relating to the appropriate situation. In handling the limit of $1/(1-x)$ as x tends to 1 we might say that the limit from the left is $+\infty$ and from the right is $-\infty$ and $1/(1-x)$ is "not defined at $x=1$". The problem as to whether $2\infty^3+3$ is a bigger infinity than ∞^2+4 is responded to by a rap over the knuckles "don't do arithmetic with infinity, it leads to contradictions", "a symbol like ∞/∞ is meaningless" and so on. The problem of the "infinitieth place" argument also has a slick response. "Let

$$s_n = 0.999\ldots9 \text{ (to } n \text{ decimal places)},$$

so that $s_1 = 0.9$, $s_2 = 0.99$, and so on. Then

$$1-s_n = 1/10^n$$

and taking the limit as n tends to infinity gives

$$1-\lim_{n\to\infty} s_n = 0,$$

so

$$1-0.999\ldots999\ldots = 1-\lim_{n\to\infty} s_n = 0$$

and there is no digit left over in the infinitieth place."

We often learn not to "understand" these arguments but we do "get used to them". So we become enlightened in the true ways of modern mathematics.

However, there are two fundamental problems with this eventual acceptance of the wisdom of our elders and betters. The first is *why* do we so persistently obtain these early intuitions of infinity as a direct product of school experience? The second is a more fundamental one. If everyone seems to get such wild ideas, in what sense is the accepted mathematical definition so much better?

When we look back at the history of mathematics we find a new twist. Three hundred years ago when the calculus was invented the accepted mathematical notion of infinity was quite different from what it is now. In fact it is closer to the notions of infinity given by the limiting processes which are discussed in school. The joke is that many of the intuitions that we sense nowadays would have been perfectly acceptable in the theories of three centuries ago. Moreover a modern invention called "non-standard analysis" permits the use of these old ideas of infinity and it is possible to conceive of such ideas in a much simpler context, as I tried to demonstrate in a recent article in the *Mathematical Gazette*[1].

My purpose in discussing these matters is not to suggest to readers that they must now learn a revolutionary new piece of mathematics so that they may achieve mathematical salvation — far from it. What I would like to do is to give an explanation as to why these intuitions of infinity are produced in schoolwork. Then I would like to demonstrate that these intuitions are not as stupid as the accepted mathematical definitions make out. The formal mathematical definition is perfectly alright in a context where number means a comparison of size of sets and cardinal number gives a theoretical extension to the counting concept. However, in other contexts, such as limiting processes, the cardinal concept is singularly inappropriate to explain intuitions of infinity which arise. To demonstrate this I shall introduce a simple number system which contains infinite elements and infinitesimally small elements as well as the usual real numbers. I hope the average reader won't be frightened by my intention to do this. I guarantee that (s)he will have met the number system before but will just have called it by a different name. We shall see that it *is* possible to have a number system with infinities of different sizes which can be added, multiplied, subtracted and divided in a sensible way, and where the multiplicative inverse of an infinite element is an infinitesimal. This is quite different from the case of cardinal numbers which can only be added and multiplied. Subtraction and division of infinite cardinal numbers cannot be defined, a fact which Cantor took as "proof" that infinitesimals could not exist. I shall deal with these properties of cardinal infinity first since they are instructive in showing how different kinds of infinity have different properties. Then I shall go on to describe the simple system which allows a full arithmetic of infinite elements of different sizes. After that I shall look at the way intuitions of infinity arise in school and consider which mathematical theory is more appropriate for the understanding of these intuitions. The problem of "nought point nine recurring" requires an allusion to non-standard analysis which I shall briefly mention. Once again we shall see that the intuition, though incompatible with the modern theory of cardinal infinity does fit very well with an alternative mathematical theory. Finally I shall consider the consequences of what I have had to say for teachers. It will not mean the learning of new mathematics, rather a greater respect for the mathematics that is used in school today and an even greater respect for the calculus that was taught yesterday.

1. Cardinal Infinities

Suppose that for each set X we have a symbol called the *cardinal number* of X so that if there is a bijection between X and Y then X and Y have the same cardinal number, but if no such bijection exists then X and Y have different cardinal numbers. For finite sets we can do this straight away by taking the cardinal number to be the number of elements in the set. For infinite sets we need to invent new symbols. For instance the cardinal number of the set $\mathbb{N} = \{1, 2, 3, \ldots\}$ is usually denoted by $\aleph_0$ ("aleph zero") where $\aleph$ is the first letter of the Hebrew alphabet, signifying the first infinite cardinal.

Let E be the set of even numbers $\{2, 4, 6, \ldots\}$ and O the set of odd numbers $\{1, 3, 5, \ldots\}$, then there are bijections

$$f: \mathbb{N} \to E, f(n) = 2n,$$
$$g: \mathbb{N} \to O, g(n) = 2n - 1,$$

so the cardinal number of both E and O is $\aleph_0$.

If A and B are finite sets with no elements in common and have m and n elements respectively, then $A \cup B$ has $m+n$ elements. For general cardinal numbers α, β we choose sets X, Y, which have no elements in common, with cardinal numbers α, β respectively and define $\alpha+\beta$ to be the cardinal number of $X \cup Y$.

For instance E and O have no elements in common, so $\aleph_0 + \aleph_0$ is the cardinal number of $E \cup O = \mathbb{N}$ giving

$$\aleph_0 + \aleph_0 = \aleph_0. \qquad (1)$$

For any natural number n, the set $\{-1, -2, \ldots, -n\}$ has cardinal number n and has no element in common with $\mathbb{N}$, so $\aleph_0 + n$ is the cardinal number of the set

$$\mathbb{N} \cup \{-1, -2, \ldots, -n\}$$

But the map $f: \mathbb{N} \to \mathbb{N} \cup \{-1, -2, \ldots, -n\}$ given by

$$f(r) = -r \text{ for } r \leqslant n$$
$$f(r) = r - n \text{ for } r > n$$

is a bijection, so

$$\aleph_0 + n = \aleph_0 \qquad (2)$$

for any finite cardinal n.

Equations (1) and (2) tell us that it is hopeless to try to define $\alpha - \beta$ for cardinals α, β; subtracting $\aleph_0$ from both sides of (1) gives $\aleph_0 = 0$ and subtracting $\aleph_0$ from both sides of (2) gives $n = 0$ for any natural number n!

Analogous definitions and difficulties hold for the product of two cardinal numbers α, β. Here we choose sets X, Y with cardinal numbers α, β respectively and define the product cardinal $\alpha\beta$ to be the cardinal of the cartesian product $X \times Y$. For instance if $\alpha = 2$, $\beta = \aleph_0$, we can take $X = \{0, 1\}$, $Y = \mathbb{N}$ then $X \times Y$ is the set of ordered pairs of the form $(0, n)$ or $(1, n)$ for n a natural number. The function

$$f: X \times Y \to \mathbb{N}$$

given by

$$f((0, n)) = 2n - 1$$
$$f((1, n)) = 2n$$

is a bijection, so

$$2\aleph_0 = \aleph_0.$$

This immediately tells us that we cannot define division in general because dividing both sides by $\aleph_0$ would give $2 = 1$.

In recognising the problems of subtracting and dividing cardinals Cantor decreed that infinitesimals (found by dividing a finite number by an infinite one) could not exist[2]. His conclusion is incorrect. It should have been that infinitesimals do not exist *within the context of cardinal numbers*. In a similar way solutions of the problems $3 \div 4$ or $5 - 8$ do not exist in the context of natural numbers $\mathbb{N}$. However, we can extend $\mathbb{N}$ into a larger number system, the rational numbers for instance, in which solutions of both problems exist. For cardinal numbers there is no way that the system can be extended to include infinitesimals; what is needed is an entirely different system.

2. The Superrational Numbers

In this section we introduce a system that has got infinite elements and infinitesimals in it, when suitably interpreted. The system consists of quotients of polynomials,

$$\frac{a_n x^n + \ldots + a_0}{b_m x^m + \ldots + b_0}$$

with real coefficients and $b_m \neq 0$. These can be added and multiplied in the usual way. However there is no apparent system of order on them. This we introduce as follows. Let $\rho(x)$, $\sigma(x)$ be two quotients of polynomials. First draw the graphs of ρ and σ. For at most a finite number of points each graph is defined and has a finite value. The two graphs are equal when $\rho(x) = \sigma(x)$ which, on simplification, becomes a polynomial in x and so has only a finite number of solutions (except of course the trivial case when the graphs are identical). Sometimes the graph of ρ may be above that of σ and sometimes vice versa. It is no good saying one is "bigger" than the other if it has its graph always above that of the other, for that rarely happens. We make a more modest definition. For $\rho \neq \sigma$ there is always an interval $\{x \in \mathbb{R} \mid 0 < x < k\}$ in which the graphs do not cross. We can find such an interval by finding all the points $x_1, \ldots, x_m$ for which $\rho(x_i) = \sigma(x_i)$ and taking k to be the smallest strictly positive value amongst $x_1, \ldots x_m$. We define $\rho > \sigma$ if the graph of ρ is entirely above the graph of σ in this interval between

0 and k. For instance, if $\rho(x) = 1/x$ and $\sigma(x) = x^2$, then ρ is defined everywhere but zero, and $\rho(x) = \sigma(x)$ where

$$1/x = x^2 \text{ (but } x \neq 0)$$

that is

$$x^3 = 1 \text{ (but } x \neq 0)$$

so $x = 1$. In the interval from 0 to 1 the graph of ρ is above that of σ so we have $\rho > \sigma$ (Fig. 1).

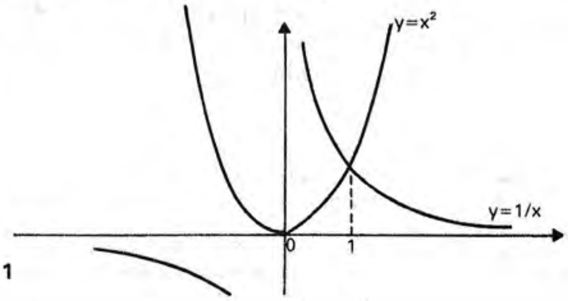

Fig. 1

Amongst these quotients of polynomials are the constants of the form a_0/b_0. For any real number α we imagine this to correspond to the quotient $\alpha/1$. In this way we can think of the real numbers as a subset of the set of quotients of polynomials. Let $\varepsilon(x) = x$ then for any positive real number α we have the graph of $y = \varepsilon(x)$ below the graph of $y = \alpha$ for x between 0 and α (Fig. 2).

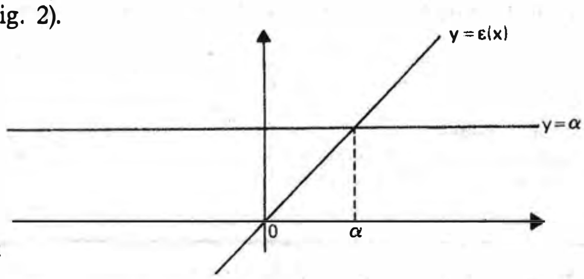

Fig. 2

Thus we can write $\alpha > \varepsilon$. Similarly we have $\varepsilon(x)$ above 0 for positive x, so $\varepsilon > 0$. We end up with the property

$$\alpha > \varepsilon > 0 \text{ for all positive real numbers } \alpha.$$

In this sense we see that ε is a positive quantity (that is $\varepsilon > 0$) which is smaller than every positive real number. This is the historical definition of an infinitesimal. By the same token, if we write $1/\varepsilon$ for the function

$$(1/\varepsilon)(x) = 1/x,$$

we find that

$$1/\varepsilon > \alpha \text{ for every real number } \alpha.$$

In this sense $1/\varepsilon$ is an infinite element since it is bigger than all real numbers.

A useful notation is to write $f + g$ for the sum of two functions f, g using the formula

$$(f+g)(x) = f(x) + g(x)$$

wherever both $f(x)$ and $g(x)$ are defined. Similarly we define $f - g$, $f \times g$, f/g by

$$(f-g)(x) = f(x) - g(x)$$
$$(f \times g)(x) = f(x)g(x)$$
$$(f/g)(x) = f(x)/g(x)$$

wherever the right-hand sides are defined. For example

$$(\varepsilon \times \varepsilon)(x) = x^2$$

which will be written as

$$\varepsilon^2(x) = x^2.$$

More generally we shall write

$$\varepsilon^n(x) = x^n.$$

In the same way a polynomial function f where

$$f(x) = a_n x^n + \ldots + a_0$$

can be written as

$$f(x) = (a_n \varepsilon^n + \ldots + a_0)(x)$$

and a quotient of polynomials

$$\rho(x) = (a_n x^n + \ldots + a_0)/(b_m x^m + \ldots + b_0)$$

can be written as

$$\rho = (a_n \varepsilon^n + \ldots + a_0)/(b_m \varepsilon^m + \ldots + b_0).$$

The set of all quotients of polynomials in ε will be called the *superrational numbers*. The superrational numbers can be thought of as functions or just as formal expressions. They may even be imagined as points on a number line (see the article "Looking at graphs through infinitesimal microscopes, windows and telescopes"[1] where these ideas are explored geometrically). To prevent this article becoming too long we shall be content with observing that the superrationals contain infinitesimals (such as ε) and infinite elements (such as $1/\varepsilon$). These may be added, multiplied, subtracted and divided in the usual arithmetic way. It is also easy to see that the inverse of an infinitesimal is an infinite element and vice versa.

Given a rational function, such as $f(x) = 1/x$, then we may compute the value of this function on an infinitesimal,

$$f(\varepsilon) = 1/\varepsilon.$$

Thus the function takes on infinite values for x infinitesimal. More than that, we find that ε^2 is a smaller infinitesimal than ε, whilst $f(\varepsilon^2) = 1/\varepsilon^2$ is a *bigger* infinite element than $f(\varepsilon) = 1/\varepsilon$.

Referring back to the examples in the introduction, we now see that we still cannot define $f(1)$ when $f(x) = 1/(1-x)$, but if we take a number infinitesimally smaller than 1, say $1 - \varepsilon$, we find that $f(1-\varepsilon) = 1/\varepsilon$ is a positive infinite element and similarly $f(1+\varepsilon) = -1/\varepsilon$ is a negative infinite element. The problem of the introduction concerning the relative sizes of $x^2 + 4$ and $2x^3 + 3$ for arbitrarily large x may be solved by putting an infinite superrational instead of x, for instance $x = 1/\varepsilon$. Then we find that $(1/\varepsilon)^2 + 4$ and $(2/\varepsilon^3) + 3$ are both infinite, but the latter is a larger infinite element than the former.

3. Intuitions of Infinity in Limiting Processes

As I alluded earlier, intuitions of infinity in school tend to arise in the consideration of limiting processes rather than comparison of sets. Certain early intuitions concern cardinality. The realisation that the counting process is unending is perhaps the first of these. The young child realises the *potential* infinity of the natural numbers and that the process of counting can never cover all of them. It is an interesting fact that much later when set theoretic notation is considered and the symbol $\mathbb{N}$ is introduced for the natural numbers then there arises the impression that one *can* consider the totality of the natural numbers. In a cognitive sense most students at university are no longer aware of the intuitive notion of potential infinity, they believe in the *actual* infinity of the set $\mathbb{N}$, a belief supported by the introduction of Cantor's theory of infinite cardinals.

However, with limiting processes the dynamic way in which limits are expressed, for instance, the fact that

$$\lim_{x \to a} f(x) = l$$

is interpreted as "$f(x)$ tends to l as x tends to a", leads to a cognitive belief that limits are approached but not actually reached. I first noticed this strong tendency during joint work with Rolph Schwarzenberger[3] and have reported it in several other contexts during subsequent investigations into students' beliefs[4]. I have also encountered many students who believe that an expression like $x^n/n!$ tends to zero as n tends to infinity because the bottom becomes a "larger infinity" than the top. Such notions arise from the mathematical experiences they receive in school and university concerning limiting processes.

It is clear that the theory of infinity which is more consonant with these intuitions is that of section 2 rather than the cardinal infinity of section 1. In saying this I am not asserting that the students have anything like the superrationals in mind. Far from it. But the kind of intuitions that they have are similar in kind to those experienced by mathematicians of earlier ages who conceived of infinitesimals as "variables which approached zero".

What are the infinitesimals in the superrationals? If we regard them as functions once more and draw their graphs, it is a routine matter to check that an infinitesimal is precisely a function which tends to zero as x tends to zero. In the theory of Tall[1] an infinitesimal corresponds to the notion of a function which tends to zero and in the wider theory of non-standard analysis the same property is true (though references to this fact are heavily buried in the research literature).

In this way I claim that the intuitions of modern schoolchildren who imagine infinitesimal quantities (though they do not always use this term) to be quantities which grow very small and infinite quantities to be those which grow very large, are in a long tradition of mathematics with a notion of infinity quite different from that of cardinal infinity. (These differences between this intuitive notion of infinity and cardinal infinity are explored in greater depth by Tall[5].)

The problem of "nought point nine recurring" being just less than one remains to be discussed. That has already been touched upon[6]. First there is the intuition of students that the process of approaching the limit is never completed. (Notice the contrast between the *actual* infinity of sets imagined by university students and the *potential* completion of infinite processes.) Thus they infer that 0.999 . . . 999 . . . is never actually equal to 1. A phenomenon like this occurs in non-standard analysis. Here it is noted that 0.99 . . . 9 to n places equals $1 - 1/10^n$. Non-standard analysis guarantees the existence of an *infinite* N so that 0.999 . . . 999 . . . to N places equals $1 - 1/10^N$. Here $1/10^N$, being the inverse of an infinite element, is an infinitesimal. So in a genuine mathematical sense 0.999 . . . 999 . . . to an infinite number of places is infinitesimally smaller than 1. The *limit* of the sequence of decimal approximations is defined by removing the infinitesimal part, leaving 1 as the limit of the sequence. Thus in non-standard analysis one has it both ways. The limit is 1, but to an infinite number of places nought point nine recurring is just less than 1.

4. Consequences for Teachers

The conclusion that one should draw from the previous discussion is not that everyone should rush out and buy books on non-standard analysis to learn the new gospel. In general the subject matter hasn't shaken down to a form where it is easily digestible, apart from a good textbook by Keisler[7]. This was used for students in America taking a first course in calculus and first reports were very encouraging[8], though subsequent sales of the book indicate that it has failed to really take off. Another book[9] intended for instructors on the course gives a self-contained account of the theory that is probably the best available.

The real bonus for teachers is that they should trust their well founded intuitions more and not be brow-beaten by the "correct" mathematics. For years a beautifully serviceable version of calculus was taught in schools based on the dynamic ideas of variables "becoming small", appealing naturally to the intuitions of teacher and pupil alike. For all its logical rigour formal analysis is inappropriate for a first course in calculus and instead of feeling guilty about "not doing calculus properly", the teacher should hold his head up high in the knowledge that a calculus based on intuitions of dynamical growth is a perfectly viable beginning for an alternative formal theory.

Perhaps I should come clean about the fact that this article was written as an enrichment and a supplement to "One hundred and one ways to infinity" by Tony Gardiner which has recently been serialised in this journal[10]. There the author claimed that his notes "were written . . . to facilitate the formation of a coherent, if incomplete, idea of 'infinity' in their pupils' minds". It should be clear from what I have said that there is not a single coherent idea of infinity at present available for students to grasp and that the kind of infinity purveyed by Tony Gardiner for the greater part of his article (cardinal infinity) is quite alien to their intuitive notions developed in school. "One hundred and one ways to infinity" failed to mention the kind of infinity more consonant with their intuitive thoughts. It also gave the impression that Cantor "got it right" when previous generations had it all wrong. That is a perversion of history seen by a generation persuaded by the theories of Cantor. Earlier mathematicians who developed the calculus were pretty canny fellows too.

My advice to teachers then is to enjoy your study of calculus and your intuitive notions of infinity and don't let these ideas be confused by the completely separate theory of infinite cardinals. The latter theory is a thing of exquisite mathematical beauty to be studied in depth in its own right. But it has no place in the beginning theory of calculus and no relevance in the type of infinity which arises in association with the limiting processes of school mathematics.

References

1. Tall, D. O. (1980) "Looking at graphs through infinitesimal microscopes, windows and telescopes", *Mathematical Gazette*, **64**, 22-49.
2. Grattan-Guiness, I. (Editor) (1980) *From Calculus to Set Theory*, Duckworth, pp. 216-219.
3. Schwarzenberger, R. L. E. S. and Tall, D. O. (1978) "Conflicts in the learning of real numbers and limits", *Mathematics Teaching*, **82**, 44-49.
4. Tall, D. O. and Vinner, S. (1981) "Concept Image and Concept Definition in Mathematics with Particular Reference to Limits and Continuity", *Educational studies in Mathematics* (to appear).
5. Tall, D. O. (1980) "The notion of Infinite Measuring Number and its Relevance in the Intuition of Infinity", *Educational Studies in Mathematics*, **11**, 271-284.
6. Tall, D. O. (1980) "Infinitesimals constructed algebraically and interpreted geometrically", *Mathematical Education for Teaching* (to appear).
7. Keisler, H. J. (1976) *Elementary Calculus*, Prindle, Weber and Schmidt, Boston.
8. Sullivan, K. (1976) "The Teaching of Elementary Calculus: An Approach using Infinitesimals", *American Mathematics Monthly*, **83** (5), 370-375.
9. Keisler, H. J. (1976) *Foundations of Infinitesimal Calculus*, Prindle, Weber and Schmidt, Boston.
10. Gardiner, A. (1980) "One hundred and one ways to infinity", *Mathematics in School*, **9**, 1, 2-4; **9**, 2, 26-27; **9**, 3, 16.