Rethinking Set Theory

Tom Leinster

Abstract. Mathematicians manipulate sets with confidence almost every day, rarely making mistakes. Few of us, however, could accurately quote what are often referred to as 'the' axioms of set theory. This suggests that we all carry around with us, perhaps subconsciously, a reliable body of operating principles for manipulating sets. What if we were to take some of those principles and adopt them as our axioms instead? The message of this article is that this can be done, in a simple, practical way (due to Lawvere). The resulting axioms are ten thoroughly mundane statements about sets.

As mathematicians, we often read a nice new proof of a known theorem, enjoy the different approach, but continue to derive our internal understanding from the method we originally learned. This paper aims to change drastically the way mathematicians think [. . .] and teach.

—Sheldon Axler [**1**, Section 10]

Mathematicians manipulate sets with confidence almost every day of their working lives. We do so whenever we work with sets of real or complex numbers, or with vector spaces, topological spaces, groups, or any of the many other set-based structures. These underlying set-theoretic manipulations are so automatic that we seldom give them a thought, and it is rare that we make mistakes in what we do with sets.

However, very few mathematicians could accurately quote what are often referred to as 'the' axioms of set theory, short of looking them up. We would not dream of working with, say, Lie algebras without first learning the axioms. Yet many of us will go our whole lives without learning 'the' axioms for sets, with no harm to the accuracy of our work. This suggests that we all carry around with us, more or less subconsciously, a reliable body of operating principles that we use when manipulating sets.

What if we were to write down some of these principles and adopt *them* as our axioms for sets? The message of this article is that this can be done, in a simple, practical way. We describe an axiomatization due to F. William Lawvere [**3, 4**], informally summarized in Figure 1. The axioms suffice for very nearly everything mathematicians ever do with sets. So we can, if we want, abandon the classical axioms entirely and use these instead.

Why rethink? The traditional axiomatization of sets is known as Zermelo–Fraenkel with Choice (ZFC). Great things have been achieved on this axiomatic basis. However, ZFC has one major flaw: Its use of the word 'set' conflicts with how most mathematicians use it.

The root of the problem is that in the framework of ZFC, the elements of a set are always sets too. Thus, given a set X, it always makes sense in ZFC to ask what the elements of the elements of X are. Now, a typical set in ordinary mathematics is $\mathbb{R}$. But ask a randomly-chosen mathematician, 'what are the elements of π?', and they will probably assume they misheard you, or tell you that your question makes no sense. If forced to answer, they might reply that real numbers have no elements. But this too is

http://dx.doi.org/10.4169/amer.math.monthly.121.05.403
MSC: Primary 03E99

1.	Composition of functions is associative and has identities.
2.	There is a set with exactly one element.
3.	There is a set with no elements.
4.	A function is determined by its effect on elements.
5.	Given sets X and Y, one can form their cartesian product $X \times Y$.
6.	Given sets X and Y, one can form the set of functions from X to Y.
7.	Given $f\colon X \longrightarrow Y$ and $y \in Y$, one can form the inverse image $f^{-1}(y)$.
8.	The subsets of a set X correspond to the functions from X to $\{0, 1\}$.
9.	The natural numbers form a set.
10.	Every surjection has a right inverse.

Figure 1. Informal summary of the axioms. The primitive concepts are set, function, and composition of functions. Other concepts mentioned (such as element) are defined in terms of the primitive concepts.

in conflict with ZFC's usage of 'set': If all elements of $\mathbb{R}$ are sets, and they all have no elements, then they are all the empty set, from which it follows that all real numbers are equal.

Could we, perhaps, continue to use ZFC while quietly ignoring the requirement that the elements of a set must be sets too? No; this would leave us unable to state the ZFC axioms. For example, one axiom states that every nonempty set X has some element x such that $x \cap X = \emptyset$, which only makes sense if the elements of X are sets. When X is an ordinary set such as $\mathbb{R}$, few would recognize this axiom as meaningful: What is $\pi \cap \mathbb{R}$, after all?

I will anticipate an objection to these criticisms. The traditional approach to set theory involves not only ZFC, but also a collection of methods for encoding mathematical objects of many different types (real numbers, differential operators, random variables, the Riemann zeta function, ...) as sets. This is similar to the way in which computer software encodes data of many types (text, sound, images, ...) as binary sequences. In both cases, even the designers would agree that the encoding methods are somewhat arbitrary. So, one might object, no one is claiming that questions like 'what are the elements of π?' have meaningful answers.

However, the criticisms made in earlier paragraphs have nothing to do with the matter of encoding. The bare facts are that in ZFC, it is always valid to ask of a set 'what are the elements of its elements?', and in ordinary mathematical practice, it is not. Perhaps it is misleading to use the same word, 'set', for both purposes.

Three misconceptions. The axiomatization presented below is Lawvere's Elementary Theory of the Category of Sets, first proposed half a century ago [**3, 4**]. Here it is phrased in a way that requires no knowledge of category theory whatsoever.

Because of the categorical origins of this axiomatization, three misconceptions commonly arise.

The first is that the underlying motive is to replace set theory with category theory. It is not. The approach described here is not a rival to set theory: It *is* set theory.

The second is that this axiomatization demands more mathematical sophistication than others (such as ZFC). This is false, but understandable. Almost all of the work on Lawvere's axioms has taken place within topos theory, a beautiful and profound subject, but not one easily accessible to outsiders. It has always been known that the axioms could be presented in a completely elementary way, and although some authors have emphasized this [**3, 5, 6, 10, 11**], it is not as widely appreciated as it should be. This paper aims to make it plain.

The third misconception is that because these axioms for sets come from category theory, and because the definition of category involves a collection of objects and a collection of arrows, and because 'collection' might mean something like 'set', there is a circularity; in order to axiomatize sets categorically, we must already know what a set is. But although our approach is categorically inspired, it does not depend on having a general definition of category. Indeed, our axiomatization (Section 2) does not contain a single instance of the word 'category'.

Put another way, circularity is no more a problem here than in ZFC. Informally, ZFC says 'there are some things called sets, there is a binary relation on sets called membership, and some axioms hold.' We will say 'there are some things called sets and some things called functions, there is an operation called composition of functions, and some axioms hold.' In neither case are the 'things' required to form a set (whatever that would mean). In logical terminology, both axiomatizations are simply first-order theories.

1. PRELUDE: ELEMENTS AS FUNCTIONS. The working mathematician's vocabulary includes terms such as set, function, element, subset, and equivalence relation. Any axiomatization of sets will choose some of these concepts as primitive and derive the others. The traditional choice is sets and elements. We use sets and functions.

The formal axiomatization is presented in Section 2. However, it will be helpful to consider one aspect in advance: how to derive the concept of element from the concept of function.

Suppose for now that we have found a characterization of one-element sets without knowing what an element is. (We do so below.) Fix a one-element set $\mathbf{1} = \{\bullet\}$. For any set X, a function $\mathbf{1} \longrightarrow X$ is essentially just an element of X, since, after all, such a function f is uniquely determined by the value of $f(\bullet) \in X$ (Figure 2(c)). Thus:

Elements are a special case of functions.

This is such a trivial observation that one is apt to dismiss it as a mere formal trick. On the contrary, similar correspondences occur throughout mathematics. For example (Figure 2):

- a loop in a topological space X is a continuous map $S^1 \longrightarrow X$;
- a straight line in $\mathbb{R}^n$ is a distance-preserving map $\mathbb{R} \longrightarrow \mathbb{R}^n$;

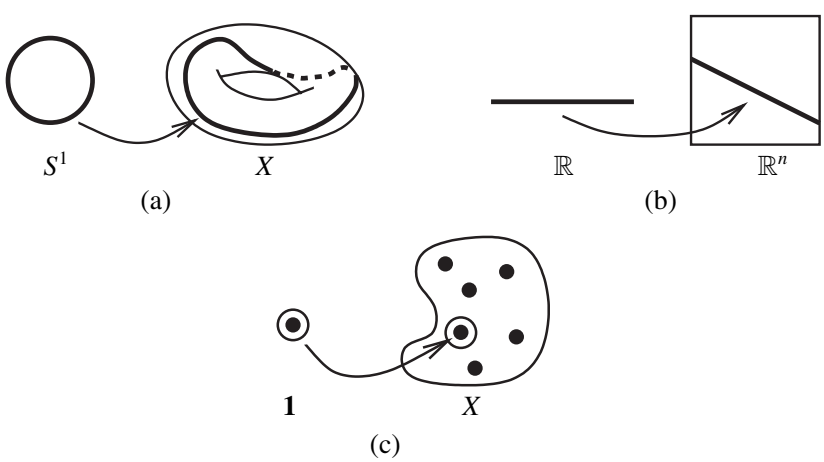

Figure 2. Mapping out of a basic object (S^1, $\mathbb{R}$, or $\mathbf{1}$) picks out figures of the appropriate type (loops, lines, or elements).

- a sequence in a set X is a function $\mathbb{N} \longrightarrow X$;
- a solution (x, y) of the equation $x^2 + y^2 = 1$ in a ring A is a homomorphism $\mathbb{Z}[X, Y]/(X^2 + Y^2 - 1) \longrightarrow A$.

In each case, the word 'is' can be taken either as a definition or as an assertion of a canonical, one-to-one correspondence. In the first, we map out of the circle, which is a 'free-standing' loop; in the second, $\mathbb{R}$ is a free-standing line; in the third, the elements $0, 1, 2, \ldots$ of $\mathbb{N}$ form a free-standing sequence; in the last, the pair (X, Y) of elements of $\mathbb{Z}[X, Y]/(X^2 + Y^2 - 1)$ is the free-standing solution (x, y) of $x^2 + y^2 = 1$. Similarly, in our trivial situation, the set $\mathbf{1}$ is a free-standing element, and an element of a set X is just a map $\mathbf{1} \longrightarrow X$.

We could write $\bar{x}$, say, for the function $\mathbf{1} \longrightarrow X$ with value $x \in X$. However, we will write $\bar{x}$ as simply x, blurring the distinction. In fact, we will later *define* an element of X to be a function $\mathbf{1} \longrightarrow X$.

This will make some readers uncomfortable. There is, you will agree, a canonical one-to-one correspondence between elements of X and functions $\mathbf{1} \longrightarrow X$, but perhaps you draw the line at saying that an element of X literally *is* a function $\mathbf{1} \longrightarrow X$. If so, this is not a deal-breaker. We could adapt the axiomatization in Section 2 by adding 'element' to the list of primitive concepts. Then, however, we would need to complicate it further by adding clauses to guarantee that (among other things) there is a one-to-one correspondence between elements of X and functions $\mathbf{1} \longrightarrow X$, for any set X. It can be done, but we choose the more economical route.

We have seen that elements are a special case of functions. There is another fundamental way in which functions and elements interact: Given a function $f : X \longrightarrow Y$ and an element $x \in X$, we can evaluate f at x to obtain a new element, $f(x) \in Y$. Viewing elements as functions out of $\mathbf{1}$, this element $f(x)$ is nothing but the composite of f with x. That is, $f(x) = f \circ x$, as illustrated below.

$$\begin{tikzcd} \mathbf{1} \arrow[r, "x"] \arrow[dr, "f(x)"'] & X \arrow[d, "f"] \\ & Y. \end{tikzcd}$$

Hence:

Evaluation is a special case of composition.

2. THE AXIOMS. Here we state our ten axioms on sets and functions, in entirely elementary terms.

The formal axiomatization is in a different typeface, to distinguish it from the accompanying commentary. Some diagrams appear, but they are not part of the formal statement.

First we state the data to which our axioms will apply:

- some things called *sets*;
- for each set X and set Y, some things called *functions from X to Y*, with functions f from X to Y written as $f : X \longrightarrow Y$ or $X \xrightarrow{f} Y$;
- for each set X, set Y, and set Z, an operation assigning to each $f : X \longrightarrow Y$ and $g : Y \longrightarrow Z$ a function $g \circ f : X \longrightarrow Z$;
- for each set X, a function $1_X : X \longrightarrow X$.

This last item can be included in the list or not, according to taste. See the comments after the first axiom, which now follows.

Associativity and identity laws.

Axiom 1. For all sets W, X, Y, and Z, and all functions

$$W \xrightarrow{f} X \xrightarrow{g} Y \xrightarrow{h} Z,$$

we have $h \circ (g \circ f) = (h \circ g) \circ f$. For all sets X and Y and functions $f\colon X \longrightarrow Y$, we have $f \circ 1_X = f = 1_Y \circ f$.

If we wish to omit the identity functions from the list of primitive concepts, we must replace the second half of Axiom 1 by the statement that for all sets X, there exists a function $1_X\colon X \longrightarrow X$ such that $g \circ 1_X = g$ for all $g\colon X \longrightarrow Y$ and $1_X \circ f = f$ for all $f\colon W \longrightarrow X$. These conditions characterize 1_X uniquely.

One-element set. We would like to say 'there exists a one-element set', but for the moment we lack the expressive power to say 'element'. However, any one-element set T should have the property that for each set X, there is precisely one function $X \longrightarrow T$. Moreover, *only* one-element sets should have this property. This motivates the following definition and axiom.

A set T is *terminal* if for every set X, there is a unique function $X \longrightarrow T$.

Axiom 2. There exists a terminal set.

It follows quickly from the definitions that if T and T' are terminal sets, then there is a unique isomorphism from T to T'. (A function $f\colon A \longrightarrow B$ is an *isomorphism* if there is a function $f'\colon B \longrightarrow A$ such that $f' \circ f = 1_A$ and $f \circ f' = 1_B$.) In other words, terminal sets are unique up to unique isomorphism. It is therefore harmless to fix a terminal set $\mathbf{1}$ once and for all. Readers concerned by this are referred to the last few paragraphs of this section.

Given a set X, we write $x \in X$ to mean $x\colon \mathbf{1} \longrightarrow X$, and call x an *element* of X. Given $x \in X$ and a function $f\colon X \longrightarrow Y$, we write $f(x)$ for the element $f \circ x\colon \mathbf{1} \longrightarrow Y$ of Y.

Empty set.

Axiom 3. There exists a set with no elements.

Functions and elements. A function from X to Y should be nothing more than a way of turning elements of X into elements of Y.

Axiom 4. Let X and Y be sets and $f, g\colon X \longrightarrow Y$ functions. Suppose that $f(x) = g(x)$ for all $x \in X$. Then $f = g$.

Axioms 1, 2, and 4 imply that a set is terminal if and only if it has exactly one element. This justifies the usage of 'one-element set' as a synonym for 'terminal set'.

Cartesian products. We want to be able to form cartesian products of sets. An element of X together with an element of Y should uniquely determine an element of $X \times Y$. More generally, for any set I, a function $f_1\colon I \longrightarrow X$ together with a func-

tion $f_2\colon I \longrightarrow Y$ should uniquely determine a function $f\colon I \longrightarrow X \times Y$, given by $f(t) = (f_1(t), f_2(t))$. (To see that this really is 'more generally', take $I = \mathbf{1}$.) We can recover f_1 from f by composing with the projection $p_1\colon X \times Y \longrightarrow X$, and similarly f_2, as in the following definition.

Let X and Y be sets. A *product* of X and Y is a set P together with functions $X \xleftarrow{p_1} P \xrightarrow{p_2} Y$, with the following property (Figure 3):

> For all sets I and functions $X \xleftarrow{f_1} I \xrightarrow{f_2} Y$, there is a unique function $(f_1, f_2)\colon I \longrightarrow P$ such that $p_1 \circ (f_1, f_2) = f_1$ and $p_2 \circ (f_1, f_2) = f_2$.

$$\begin{tikzcd}
 & I \arrow[dl, "f_1"'] \arrow[d, dotted, "{(f_1,f_2)}"] \arrow[dr, "f_2"] & \\
X & P \arrow[l, "p_1"] \arrow[r, "p_2"'] & Y
\end{tikzcd}$$

Figure 3. The characteristic property of products

Axiom 5. Every pair of sets has a product.

Strictly speaking, a product consists of not only the set P, but also the projections p_1 and p_2. Any two products of X and Y are uniquely isomorphic: Given products (P, p_1, p_2) and (P', p_1', p_2'), there is a unique isomorphism $i\colon P \longrightarrow P'$ such that $p_1' \circ i = p_1$ and $p_2' \circ i = p_2$. As in the case of terminal sets, this makes it harmless to choose once and for all a preferred product $(X \times Y, \mathrm{pr}_1^{X,Y}, \mathrm{pr}_2^{X,Y})$ for each pair X, Y of sets. Again, this convention is justified at the end of the section.

Sets of functions. In everyday mathematics, we can form the set Y^X of functions from one set X to another set Y. For any set I, the functions $q\colon I \times X \longrightarrow Y$ correspond one-to-one with the functions $\bar{q}\colon I \longrightarrow Y^X$, simply by changing the punctuation:

$$q(t, x) = (\bar{q}(t))(x) \tag{1}$$

$(t \in I, x \in X)$. For example, when $I = \mathbf{1}$, this reduces to the statement that the functions $X \longrightarrow Y$ correspond to the elements of Y^X.

In (1), we are implicitly using the evaluation map

$$\begin{array}{rccc} \varepsilon\colon & Y^X \times X & \longrightarrow & Y \\ & (f, x) & \longmapsto & f(x). \end{array}$$

Then (1) becomes the equation $q(t, x) = \varepsilon(\bar{q}(t), x)$, as in the following definition.

Let X and Y be sets. A *function set* from X to Y is a set F together with a function $\varepsilon\colon F \times X \longrightarrow Y$, with the following property (Figure 4):

> For all sets I and functions $q\colon I \times X \longrightarrow Y$, there is a unique function $\bar{q}\colon I \longrightarrow F$ such that $q(t, x) = \varepsilon(\bar{q}(t), x)$ for all $t \in I$ and $x \in X$.

Axiom 6. For all sets X and Y, there exists a function set from X to Y.

$$\begin{tikzcd} I \times X \arrow[d, dotted, "\bar{q}\times 1_X"'] \arrow[dr, "q"] & \\ F \times X \arrow[r, "\varepsilon"'] & Y \end{tikzcd}$$

Figure 4. The characteristic property of function sets

Inverse images. Ordinarily, given a function $f\colon X \longrightarrow Y$ and an element y of Y, we can form the inverse image or fiber $f^{-1}(y)$. The inclusion function $j\colon f^{-1}(y) \hookrightarrow X$ has the property that $f \circ j$ has constant value y. Moreover, whenever $q\colon I \longrightarrow X$ is a function such that $f \circ q$ has constant value y, the image of q must lie within $f^{-1}(y)$; that is, $q = j \circ \bar{q}$ for some $\bar{q}\colon I \longrightarrow f^{-1}(y)$ (necessarily unique).

Let $f\colon X \longrightarrow Y$ be a function and $y \in Y$. An *inverse image* of y under f is a set A together with a function $j\colon A \longrightarrow X$, such that $f(j(a)) = y$ for all $a \in A$ and the following property holds (Figure 5):

> For all sets I and functions $q\colon I \longrightarrow X$ such that $f(q(t)) = y$ for all $t \in I$, there is a unique function $\bar{q}\colon I \longrightarrow A$ such that $q = j \circ \bar{q}$.

$$\begin{tikzcd} I \arrow[dr, dotted, "\bar{q}"] \arrow[drr, bend left] \arrow[ddr, bend right, "q"'] & & \\ & A \arrow[r] \arrow[d, "j"] & \mathbf{1} \arrow[d, "y"] \\ & X \arrow[r, "f"'] & Y \end{tikzcd}$$

Figure 5. The characteristic property of inverse images

Axiom 7. For every function $f\colon X \longrightarrow Y$ and element $y \in Y$, there exists an inverse image of y under f.

Inverse images are essentially unique: If $j\colon A \longrightarrow X$ and $j'\colon A' \longrightarrow X$ are both inverse images of y under f, then there is a unique isomorphism $i\colon A \longrightarrow A'$ such that $j' \circ i = j$.

Characteristic functions. Sometimes we want to define a function on a case-by-case basis. For example, we might want to define $h\colon \mathbb{R} \longrightarrow \mathbb{R}$ by $h(x) = x\sin(1/x)$ if $x \neq 0$ and $h(0) = 0$. A simple instance is the definition of characteristic function.

Fix a two element-set $\mathbf{2} = \{t, f\}$ (for 'true' and 'false'). The characteristic function of a subset $A \subseteq X$ is the function $\chi_A\colon X \longrightarrow \mathbf{2}$ defined by $\chi_A(x) = t$ if $x \in A$, and $\chi_A(x) = f$ otherwise. It is the unique function $\chi\colon X \longrightarrow \mathbf{2}$ such that $\chi^{-1}(t) = A$.

This is how characteristic functions work ordinarily. To ensure that they work in the same way in our set theory, we now demand that there exist a set $\mathbf{2}$ and an element $t \in \mathbf{2}$ with the property just described: Whenever X is a set and $A \subseteq X$, there is a unique function $\chi\colon X \longrightarrow \mathbf{2}$ such that $\chi^{-1}(t) = A$.

Since we do not yet have a definition of subset, we phrase the axiom in terms of injections instead. (The thought here is that every subset inclusion $A \hookrightarrow X$ is injective, and, up to isomorphism, every injection arises in this way.)

An *injection* is a function $j\colon A \longrightarrow X$ such that $j(a) = j(a') \implies a = a'$ for $a, a' \in A$.

A *subset classifier* is a set $\mathbf{2}$ together with an element $t \in \mathbf{2}$, with the following property (Figure 6):

> For all sets A and X and injections $j\colon A \longrightarrow X$, there is a unique function $\chi\colon X \longrightarrow \mathbf{2}$ such that $j\colon A \longrightarrow X$ is an inverse image of t under χ.

$$\begin{tikzcd} A \arrow[r] \arrow[d, "j"'] & \mathbf{1} \arrow[d, "t"] \\ X \arrow[r, dotted, "\chi"'] & \mathbf{2} \end{tikzcd}$$

Figure 6. The characteristic property of subset classifiers

Axiom 8. There exists a subset classifier.

The notation $\mathbf{2}$ is merely suggestive. There is nothing in the definition saying that $\mathbf{2}$ must have two elements, but, nontrivially, our ten axioms do in fact imply this.

Natural numbers. In ordinary mathematics, sequences can be defined recursively: Given a set X, an element $a \in X$, and a function $r\colon X \longrightarrow X$, there is a unique sequence $(x_n)_{n=0}^{\infty}$ in X such that

$$x_0 = a \text{ and } x_{n+1} = r(x_n) \text{ for all } n \in \mathbb{N}.$$

A sequence in X is nothing but a function $\mathbb{N} \longrightarrow X$, so the previous sentence is really a statement about the set $\mathbb{N}$. It also refers to two pieces of structure on $\mathbb{N}$: the element 0 and the function $s\colon \mathbb{N} \longrightarrow \mathbb{N}$ given by $s(n) = n + 1$.

A *natural number system* is a set N together with an element $0 \in N$ and a function $s\colon N \longrightarrow N$, with the following property (Figure 7):

> Whenever X is a set, $a \in X$, and $r\colon X \longrightarrow X$, there is a unique function $x\colon N \longrightarrow X$ such that $x(0) = a$ and $x(s(n)) = r(x(n))$ for all $n \in N$.

$$\begin{tikzcd} \mathbf{1} \arrow[r, "0"] \arrow[d, "1_{\mathbf{1}}"'] & N \arrow[r, "s"] \arrow[d, dotted, "x"] & N \arrow[d, dotted, "x"] \\ \mathbf{1} \arrow[r, "a"'] & X \arrow[r, "r"'] & X \end{tikzcd}$$

Figure 7. The characteristic property of natural number systems

Axiom 9. There exists a natural number system.

Natural number systems are essentially unique, in the usual sense that between any two of them there is a unique structure-preserving isomorphism. This justifies speaking of *the* natural numbers $\mathbb{N}$, as we invariably do.

Choice. A function with a right inverse is certainly surjective. The axiom of choice states the converse.

A *surjection* is a function $s\colon X \longrightarrow Y$ such that for all $y \in Y$, there exists $x \in X$ with $s(x) = y$.

A *right inverse* of a function $s\colon X \longrightarrow Y$ is a function $i\colon Y \longrightarrow X$ such that $s \circ i = 1_Y$.

Axiom 10. Every surjection has a right inverse.

A right inverse of a surjection $s\colon X \longrightarrow Y$ is a choice, for each $y \in Y$, of an element of the nonempty set $s^{-1}(y)$.

This concludes the axiomatization.

The meaning of 'the'. It remains to reassure any readers concerned by the liberty taken in Axioms 2 and 5, where we chose once and for all a terminal set and a cartesian product for each pair of sets.

This type of liberty is very common in mathematical practice. We speak of *the* trivial group, *the* 2-sphere, *the* direct sum of two vector spaces, etc., even though we can conceive of many trivial groups or 2-spheres or direct sums, all isomorphic but not equal. Anyone asking 'but *which* trivial group?' is likely to be met with a hard stare, and for good reason: No meaningful statement about groups depends on what the element of the trivial group happens to be named.

However, we should be able to state the axioms with scrupulous rigor, and we can. One way to do so is not to single out a particular terminal set or particular products, but instead to adopt some circumlocutions. For example, we replace the phrase 'for all elements $x \in X$' by 'for all terminal sets T and functions $x\colon T \longrightarrow X$.'

More satisfactory, though, is to extend the list of primitive concepts. To the existing list (sets, functions, composition and identities) we add:

- a distinguished set, $\mathbf{1}$;
- an operation assigning to each pair of sets X, Y a set $X \times Y$ and functions

$$X \xleftarrow{\mathrm{pr}_1^{X,Y}} X \times Y \xrightarrow{\mathrm{pr}_2^{X,Y}} Y. \tag{2}$$

Axiom 2 is replaced by the statement that $\mathbf{1}$ is terminal, and Axiom 5 by the statement that for all sets X and Y, the set $X \times Y$ together with the functions (2) is a product of X and Y.

This approach has the virtue of reflecting ordinary mathematical usage. We usually speak as if taking the product of two sets (or spaces, groups, etc.) were a procedure with a definite output: *the* product, not *a* product. But since products are in any case determined uniquely up to unique isomorphism, whether or not we nominate one as special makes no significant difference.

3. DISCUSSION. The ten axioms are familiar in their intuitive content, but less so as an axiomatic system. Here we discuss the implications of using them as such.

Building on the axioms. Any axiomatization of anything is followed by a period of lemma-proving. The present axioms are no exception. Here is a very brief sketch of the development.

It is convenient formally to define a *subset* of a set X as a function $X \longrightarrow \mathbf{2}$, but we constantly use the correspondence between functions $X \longrightarrow \mathbf{2}$ and injections into

X, provided by Axiom 8. Two injections j, j' into X correspond to the same subset of X if and only if they have the same image (that is, there exists an isomorphism i such that $j' = j \circ i$).

The main task is to build the everyday equipment used for manipulating sets. For example, given a function $f\colon X \longrightarrow Y$, we construct the image under f of a subset of X and the inverse image of a subset of Y. An equivalence relation $\sim$ on a set X is defined to be a subset of $X \times X$ with the customary properties, and the axioms allow us to construct the quotient set $X/{\sim}$. Some constructions are tricky. For instance, the axioms imply that any two sets X and Y have a disjoint union $X \sqcup Y$, but no known proof is simple.

We then define the usual number systems. Addition, multiplication, and powers of natural numbers are defined directly using Axiom 9. From $\mathbb{N}$, we successively construct $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ in the standard way. For example, $\mathbb{Z} = (\mathbb{N} \times \mathbb{N})/{\sim}$, where $\sim$ is the equivalence relation on $\mathbb{N} \times \mathbb{N}$ given by $(m, n) \sim (m', n')$ if and only if $m + n' = m' + n$. As this illustrates, past a certain point, the development is literally identical to that for other axiomatizations of sets.

How strong are the axioms? Most mathematicians will never need more properties of sets than those guaranteed by the ten axioms. For example, McLarty [**13**] argues that no more is needed anywhere in the canons of the Grothendieck school of algebraic geometry, the multi-volume works *Éléments de Géométrie Algébrique* (EGA) and *Séminaire de Géométrie Algébrique* (SGA).

To get a sense of the reach of the axioms, let us consider infinite cartesian products. Let I be a (possibly infinite) set and $(X_i)_{i \in I}$ a family of sets. Can we form the product $\prod_{i \in I} X_i$? The answer depends on what is meant by 'family'. We could define an I-indexed family to be a set X together with a function $p\colon X \longrightarrow I$, viewing the fiber $p^{-1}(i)$ as the ith member X_i. In that case, $\prod X_i$ can be constructed as a subset of X^I. Specifically, p induces a function $p^I\colon X^I \longrightarrow I^I$, and $\prod X_i$ is the inverse image under p^I of the element of I^I corresponding to 1_I.

However, we could interpret 'I-indexed family' differently, as an algorithm or formula that assigns to each $i \in I$ a set X_i. It is not obvious that we can then form the disjoint union $X = \coprod_{i \in I} X_i$, which is what would be necessary in order to obtain a family in the previous sense. In fact, writing $\mathcal{P}(S) = \mathbf{2}^S$ for the power set of a set S, the ten axioms do *not* guarantee the existence of the disjoint union

$$\mathbb{N} \sqcup \mathcal{P}(\mathbb{N}) \sqcup \mathcal{P}(\mathcal{P}(\mathbb{N})) \sqcup \cdots \tag{3}$$

unless they are inconsistent [**8**, Section 9].

If we wish to change this, we can add an eleventh axiom (or properly, axiom scheme). It is called 'replacement', formally stated in [**12**, Section 8], and informally stated as follows. Suppose that we have a set I and a first-order formula so that each $i \in I$ specifies a set X_i up to isomorphism. Then we require that there exist a set X and a function $p\colon X \longrightarrow I$ such that $p^{-1}(i)$ is isomorphic to X_i for each $i \in I$. This guarantees the existence of sets such as (3).

The relationship between our axioms and ZFC is well understood. The ten axioms are weaker than ZFC, but when the eleventh is added, the two theories have equal strength and are bi-interpretable (the same theorems hold). This extra strength is sometimes needed; for example, replacement is important in parts of infinitary combinatorics. It is also known to which fragment of ZFC the ten axioms correspond: 'Zermelo with bounded comprehension and choice'. The details of this relationship

were mostly worked out in the early 1970s [**2, 14, 15**]. Good modern accounts are in [**7**, Section VI.10] and [**9**, Chapter 22].

A broader view. Our ten axioms are a standard rephrasing of Lawvere's Elementary Theory of the Category of Sets (ETCS), published in 1964. It was some years before ETCS found its natural home, and that was with the advent of topos theory.

The notion of topos was invented by Grothendieck for reasons that had nothing to do with set theory. For Grothendieck, a topos was a generalized topological space. Formally, a topos is a category with certain properties, and a topological space X is associated with the topos whose objects are the sheaves of sets on X.

Lawvere and Tierney swiftly realized that, after a slight loosening of Grothendieck's definition, the ETCS axioms could be restated neatly in topos-theoretic terms [**16, 17**]. Indeed, ETCS says exactly that sets and functions form a topos of a special sort: a 'well-pointed topos with natural numbers object and choice'. So a topos is not only a generalized space; it is also a generalized universe of sets.

An attractive feature of ETCS is that each of the axioms is meaningful in a broader context than set theory. For example, Axiom 1 states that sets and functions form a category. The job of the remaining axioms is to distinguish sets from other structures that form categories. Axioms 2 and 5 state that the category of sets has finite products. This important property is shared by (for example) the categories of topological spaces and smooth manifolds, which is exactly what makes it possible to define 'topological group' and 'Lie group'. But for one detail, Axioms 1, 2, 5, 6, 7 and 8 state that sets and functions form a topos.

The axiom of choice as formulated in Axiom 10 highlights a special feature of sets. In most other categories of sets-with-structure, it fails, and its failure is a point of interest. For instance, not every continuous surjection between topological spaces has a continuous right inverse, a typical example being the nonexistence of a continuous square root defined on the complex plane.

What kind of set theory should we teach? As Figure 1 indicates, we already teach a diluted form of the ten axioms, even in introductory courses. For example, we certainly tell our students that an element of $X \times Y$ is an element of X together with an element of Y, and we routinely write a function f taking values in $\mathbb{R}^2$ as (f_1, f_2), although we are less likely to state explicitly that, given functions $f_1 : I \longrightarrow X$ and $f_2 : I \longrightarrow Y$, there is a unique function $f : I \longrightarrow X \times Y$ with f_1 and f_2 as components.

When it comes to teaching *axiomatic* set theory, the approach outlined here has advantages and disadvantages. The great advantage is that such a course is of far wider benefit than one using the traditional axioms. It directly addresses a difficulty experienced by many students: the concept of function (and worse, function space). It also introduces in an elementary setting the idea of universal property. This is probably the hardest aspect of the axioms for a learner, but since universal properties are important in so many branches of advanced mathematics, the benefits are potentially far-reaching.

The disadvantages are perhaps only temporary. There is at present a lack of teaching materials (the book [**5**] being the main exception). For example, the axioms imply that any two sets have a disjoint union, and most books on topos theory contain an elegant and sophisticated proof of a generalization of this fact, but to my knowledge, there is only one place where a purely elementary proof can be found [**18**]. A second disadvantage is that any student planning a career in set theory will need to learn ZFC anyway, since almost all research-level set theory is done with the iterated-membership conception of set. (That is the current reality, which is not to say that set theory *must*

be done this way.) Such a course might usefully include a comparison of the types of set theory available.

Reactions to an earthquake. Perhaps you will wake up tomorrow, check your email, and find an announcement that ZFC is inconsistent. Apparently, someone has taken the ZFC axioms, performed a long string of logical deductions, and arrived at a contradiction. The work has been checked and re-checked. There is no longer any doubt.

How would you react? In particular, how would you feel about the implications for your own work? All your theorems would still be true under ZFC, but so too would their negations. Would you conclude that your life's work had been destroyed?

An informal survey suggests that most of us would be interested but not deeply troubled. We would go on believing that our theorems were true in a sense that their negations were not. We are unlikely to feel threatened by the inconsistency of axioms to which we never referred anyway.

In contrast, the ten axioms above are such core mathematical principles that an inconsistency in them would be devastating. If we cannot safely assume that composition of functions is associative, or that repeatedly applying a function $f : X \longrightarrow X$ to an element $a \in X$ produces a sequence $(f^n(a))$, we are really in trouble.

The difference in reactions is telling. Our response to an inconsistency in an axiomatization of set theory reflects our degree of belief that it describes the operating principles we actually employ, in ordinary mathematical practice.

In summary, simply by writing down a few mundane, uncontroversial statements about sets and functions, we arrive at an axiomatization that fits well with how sets are really used in mathematics.

ACKNOWLEDGMENTS. I thank François Dorais, Colin McLarty, Todd Trimble, the patrons of the n-Category Café, and the anonymous referees. This work was partially supported by an EPSRC Advanced Research Fellowship.

REFERENCES

1. S. Axler, Down with determinants! *Amer. Math. Monthly* **102** (1995) 139–154.
2. J. C. Cole, Categories of sets and models of set theory, in *Proceedings of the Bertrand Russell Memorial Logic Conference.* Uldum 1971. Bertrand Russell Memorial Conference, Leeds, 1973. 351–399.
3. F. W. Lawvere, An elementary theory of the category of sets, *Proc. Natl. Acad. Sci. USA* **52** (1964) 1506–1511.
4. ———, An elementary theory of the category of sets (long version) with commentary, *Repr. Theory Appl. Categ.* **12** (2005) 1–35, available at http://www.tac.mta.ca/tac/reprints/articles/11/tr11abs.html.
5. F. W. Lawvere, R. Rosebrugh, *Sets for Mathematics*. Cambridge University Press, Cambridge, 2003.
6. S. Mac Lane, *Mathematics: Form and Function*. Springer, New York, 1986.
7. S. Mac Lane, I. Moerdijk, *Sheaves in Geometry and Logic*. Springer, New York, 1994.
8. A. Mathias, The strength of Mac Lane set theory, *Ann. Pure Appl. Logic* **110** (2001) 107–234.
9. C. McLarty, *Elementary Categories, Elementary Toposes*. Oxford University Press, Oxford, 1992.
10. ———, Numbers can be just what they have to, *Noûs* **27** (1993) 487–98.
11. ———, Challenge axioms, final draft, email to Foundations of Mathematics mailing list, 6 February 1998, available at http://www.cs.nyu.edu/pipermail/fom.
12. ———, Exploring categorical structuralism, *Philos. Math.* **12** (2004) 37–53.
13. ———, A finite order arithmetic foundation for cohomology (2011), available at http://arxiv.org/abs/1102.1773.
14. W. Mitchell, Boolean topoi and the theory of sets, *J. Pure Appl. Algebra* **2** (1972) 261–274.
15. G. Osius, Categorical set theory: a characterization of the category of sets, *J. Pure Appl. Algebra* **4** (1974) 79–119.
16. M. Tierney, Sheaf theory and the continuum hypothesis, in *Toposes, Algebraic Geometry and Logic*. Lecture Notes in Math., Vol. 274, Springer, Heidelberg, 1972. 13–42.

17. ———, Axiomatic sheaf theory: some constructions and applications, in *Proceedings of CIME Conference on Categories and Commutative Algebra.* Varenna, 1971. Edizione Cremonese, Rome, 1973. 249–326.
18. T. Trimble, ETCS: building joins and coproducts (2008), available at http://ncatlab.org/nlab/show/Trimble+on+ETCS+III.

TOM LEINSTER studied at Oxford and Cambridge, and held positions at Cambridge, the Institut des Hautes Études Scientifiques, and the University of Glasgow before taking up his current job in Edinburgh. He is interested in category theory and its applications, especially some of the more unusual ones. He is a professionally qualified masseur.
School of Mathematics, University of Edinburgh, Edinburgh EH9 3JZ, United Kingdom.
Tom.Leinster@ed.ac.uk

An Elementary Application of Brouwer's Fixed Point Theorem to Transition Matrices

Transition matrices play an essential role in the study of Markov processes ([**3**], ch. 11), which have many important practical applications in business, finance, medicine, etc. ([**2**], ch. 8). A transition matrix T is a $n \times n$ stochastic matrix such that each entry p_{ij} lies between 0 and 1, the sum of each column of T equals 1, and each element of the matrix represents the probability of transitioning from state i to state j. Moreover, the equilibrium state of the process represents a fixed point vector (i.e., $Tp = p$), where p is a probability vector such that $\sum p_i = 1$. What is interesting here is that the set of probability vectors form a simply connected compact convex set, because given two probability vectors p and q, we have that $\lambda p + (1 - \lambda)q (0 \leq \lambda \leq 1)$ is also a probability vector, so we may apply the well-known Brouwer fixed point theorem to the transition matrix T (see ([**1**], pp. 251–255) for an elementary discussion of this theorem and ([**4**], pp. 42–45) for the general case). Consequently, we are guaranteed a fixed point by the theorem as well as an eigenvalue of 1. Note that when considering infinite dimensional spaces, the theorem fails since infinite dimensional bounded sets are not necessary compact such as the unit ball in l^2. However, it does hold if the set is convex and compact ([**4**], pp. 45–46). Nevertheless, in the finite dimensional case, this is certainly an elegant application of the theorem.

REFERENCES

1. R. Courant, H. Robbins, *What is Mathematics?* Oxford University Press, New York, 1941.
2. L. J. Goldstein, D. I. Schneider, M. J. Siegel, *Finite Mathematics and its Applications*. Eighth edition. Pearson, Upper Saddle River, NJ, 2004.
3. C. M. Grinstead, J. L. Snell, *Introduction to Probability*. Second edition. American Mathematical Society, Providence, RI, 1997.
4. T. L. Saaty, J. Bram, *Nonlinear Mathematics*. Dover, New York, 1964.

—Submitted by Allan J. Kroopnick,
University of Maryland University College

http://dx.doi.org/10.4169/amer.math.monthly.121.05.415
MSC: Primary 60J10