DOES MATHEMATICS HAVE ELEMENTS?

P.R. HALMOS

For Empedocles, a little over 2400 years ago, there were four chemical elements, fire, water, earth, and air, and they were continually brought together and torn apart by two opposing forces, harmony and discord. For Aristotle, a hundred years later, two binary classification schemes took the place of one: instead of harmony-discord, he had wet-dry and hot-cold. The serious alchemists of the middle ages found that nature was even more complicated than that; they classified matter by its luster, heaviness, combustilibity, solubility, and so on. Boyle, in the 1600's, offered a definition similar to that of a prime number - an element is a substance from which other substances can be made but which cannot be separated into different substances - and analytical chemistry was off and running. A hundred years later Lavoisier (inspired in part by Newton's insight into the paramount importance of weight) could formulate a modified definition that led for the first time to quantitative tables of chemical elements similar to our modern ones.

A chemist asks what things are made of and how they are put together. A systematic listing of chemical elements, incomplete and awkward at first, is not by itself a good answer, but it helps to organize what's known and to indicate what may still be missing.

No doubt many mathematicians have noted that there are some basic ideas that keep cropping up in widely different parts of their subject, combining and re-combining with one another in a way faintly reminiscent of how all matter is made up of elements. A subconscious intuitive awareness

Received 21 September 1981. This article originally appeared in The Mathematical Intelligencer Vol. 3, No. 4, December 1981. We thank Springer-Verlag for permitting its republication here.

of these "elements" of mathematics probably contributes to (possibly it constitutes) the research insight that distinguishes great mathematicians from ordinary mortals. What are they - what are the elements of mathematics?

Examples

Here are three possible examples of the sort of basic concepts that might be considered mathematical elements: geometric series, quotient structures, and eigenvectors. They seem to me to be of three different kinds, perhaps at three different levels of depth; as a first step towards classification the words computational, categorical, and conceptual (respectively) could be used to describe them.

GEOMETRIC SERIES. In a not necessarily commutative ring with unit (for example, in the set of all 3×3 square matrices with real entries), if 1 - ab is invertible, then 1 - ba is invertible. However plausible this may seem, few people can see their way to a proof immediately; the most revealing approach belongs to a different and distant subject.

Every student knows that

$$1 - x^2 = (1+x)(1-x)$$

and some even know that

$$1 - x^3 = (1+x+x^2)(1-x)$$

The generalization

 $1 - x^{n+1} = (1 + x + ... + x^n)(1-x)$

is not far away. Divide by 1 - x and let n tend to infinity; if |x| < 1, then x^{n+1} tends to 0, and the conclusion is that

$$1/(1-x) = 1 + x + x^2 + \dots$$

This simple classical argument begins with easy algebra, but the meat of the matter is analysis: numbers, absolute values, inequalities, and convergence are needed not only for the proof but even for the final equation to make sense.

In the general ring theory question there are no numbers, no absolute

values, no inequalities, and no limits - those concepts are totally inappropriate and cannot be brought to bear. Nevertheless an impressivesounding classical phrase, "the principle of permanence of functional form", comes to the rescue and yields an analytically inspired proof in pure algebra. The idea is to pretend that 1/(1-ba) can be expanded in a geometric series (which is utter nonsense), so that

$$(1-ba)^{-1} = 1 + ba + baba + bababa + \dots$$

It follows (it doesn't really, but it's fun to keep pretending) that

$$(1-ba)^{-1} = 1 + b(1 + ab + abab + ababab + ...)a$$
,

and, after one more application of the geometric series pretense, this yields

$$(1-ba)^{-1} = 1 + b(1-ab)^{-1}a$$

Now stop the pretense and verify that, despite its unlawful derivation, the formula works. If, that is, $c = (1-ab)^{-1}$, so that (1-ab)c = c(1-ab) = 1, then 1 + bca is the inverse of 1 - ba. Once the statement is put this way, its proof becomes a matter of (perfectly legal) mechanical computation.

Why does it all this work? What goes on here? Why does it seem that the formula for the sum of an infinite geometric series is true even for an abstract ring in which convergence is meaningless? What general truth does the formula embody? I don't know the answer, but I note that the formula is applicable in other situations where it ought not to be, and I wonder whether it deserves to be called one of the (computational) elements of mathematics.

QUOTIENT STRUCTURES. The symmetric difference of two subsets A and B of, say, the plane, sometimes denoted by $A \Delta B$, is the set of those points that belong either to A or to B but not to both. Question: is the operation Δ associative?

The answer is accessible by just plain work, but it's not pleasant that way: too many sets have to be kept track of and too many cases must be considered. There is a much better way to guess the answer and, at the same time, to prove that the guess is true, a way that embeds the question into a powerful general theory of structures.

Integers can be added and subtracted; better said, with respect to addition they form an abelian group. Integer-valued functions on an arbitrary non-empty set can be added and subtracted just as well, and they too form an abelian group. There is a well-known and useful correspondence between the subsets of a set (and, in particular, the subsets of the plane) and certain special integer-valued functions: to each set A there corresponds the function χ_A , its characteristic function, that takes the value 1 on A and 0 on the complement of A. When two sets are combined by Δ , what happens to the corresponding functions? The best way to find out is to calculate their sum:

$$\chi_{A}(x) + \chi_{B}(x) = \begin{cases} 0 & \text{if } x \notin A , x \notin B , \\ 1 & \text{if } x \notin A \land B , \\ 2 & \text{if } x \notin A , x \notin B . \end{cases}$$

Otherwise expressed: $\chi_A(x) + \chi_B(x)$ is odd or even according as $x \in A \Delta B$ of not. In still other words: $\chi_{A\Delta B}(x)$ is congruent modulo 2 to the sum of $\chi_A(x)$ and $\chi_B(x)$.

Sudden insight and complete victory: the formation of symmetric differences of sets is the same as addition modulo 2 for their characteristic functions. Since addition modulo 2 is associative, so is the formation of symmetric differences.

In abstract terms the argument is one in the category of abelian groups. What the argument shows is that the collection of subsets of the plane endowed with the operation Δ is (isomorphic to) the quotient of the abelian group of all integer-valued functions on the plane modulo the subgroup of all even-integer-valued functions. (It is obvious, isn't it?, that the plane plays no special role: any non-empty set can be used similarly to construct a "Boolean group".)

The mathematical element that clamors for attention here is the formation of quotient structures. Integers modulo even integers are a familiar manifestation of the concept, and so are integers modulo multiples of 12. Many other manifestations lead to deep theories; a couple of famous analytic examples are measurable sets modulo sets of measure zero.

and operators on Hilbert space modulo compact operators.

EIGENVALUES. After gathering a pile of coconuts one day, five sailors on a desert island agree to divide them evenly next morning. During the night one sailor secretly gets up and tries to divide the coconuts into five equal piles. Finding that there is a remainder of one coconut, he tosses the odd one to a monkey, and, secreting his pile, mixes up the others and goes back to sleep. The second sailor does the same thing, with the same result, and so do the third, fourth, and fifth. In the morning the remaining pile of coconuts (less one) is again divisible by 5. What is the smallest number of coconuts that the original pile could have contained?

Anybody can solve this old puzzle with nothing but paper, pencil, and patience; just "work backwards". An alternative is to examine, for each possible number x of coconuts, the number S(x) that a typical sailor's activities leave in the pile. The formula is simple enough: S(x) = (4/5)(x-1). Call an integer x a "solution" if the result of starting with x and performing the operation S six times is an integer; in this language the problem is to find the smallest positive solution. Since S(x) = (4/5)x - (4/5), so that

$$S^{6}(x) = (4/5)^{6}x - ((4/5)^{6} + (4/5)^{5} + \dots + (4/5))$$

it follows that

$$S^{6}(x) - S^{6}(y) = (4/5)^{6}(x-y)$$

for all x and y. This implies that if both x and y are solutions, then x and y are congruent modulo 5^6 ; if, conversely, x is a solution and x and y are congruent modulo 5^6 , then y also is a solution.

The simplest conceivable "solution" is the "eigenvector" belonging to the "eigenvalue equation" S(x) = x; that solution is -4. The smallest positive solution, therefore, is $-4 + 5^6 = 15621$. Moral: the concept of invariance (fixed point, eigenvalue, eigenvector) is a possible element of mathematics.

Whatever their provenance, all three of these examples $((1-ba)^{-1})$,

 $A \Delta B$, 5^6) are parts of mathematical folklore by now. The geometric series one is usually ascribed to N. Jacobson, and the eigenvalue one to P.A.M. Dirac.

Universal algebra

I chose the three examples above not only because they illustrate the possibility of three different kinds of elements, but, to be honest, because they have also a quality that is as rare as it is striking: in each of them the element can be used to solve a problem in a manner that is efficient and at least slightly surprising. All other candidates that I have been able to think of are of a more obvious and less spectacular kind. The commonest ones, the most visible ones, are on the level of universal algebra. The purpose of what follows is to describe (or at least to mention) some of the most conspicuous ones, first on the categorical and then on the conceptual level.

STRUCTURE. Mathematicians often (always?) traffic in sets with structures. The structure may come from an internal operation or two (as in groups and fields), or an external function or two (as in metric spaces or analytic manifolds). The structure may involve more than one set (as the scalars that act on and the vectors that belong to a vector space), and it may be defined in terms of classes of subsets (as in topological spaces and in measure theory).

A pertinent observation (which is a friend of every effective teacher and every productive research mathematician, but which seems never to have received official recognition) is that the constituents of a structure cannot, must not, dangle separately, but must be subjected to appropriate structural compatibility conditions. A ring is not just a set in which both addition and multiplication can be performed - it is vitally important that they be connected by the distributive law. A topological group is not just a set that has a topological structure as well as a multiplicative one - it is vitally important that they be connected by continuity.

CATEGORIES. It has been viscerally obvious to the working mathematician for a long time that homomorphisms between groups, continuous functions between topological spaces, and linear transformations between vector spaces are all the same sort of thing, playing the same roles, and behaving in many respects the same way. The concept of a "structure" might be an element of mathematics, or, perhaps better, might be an appropriate classification heading for some of the elements. The subject that discusses structures in varying degrees of generality has been called various names. It can be called universal algebra, and, under that name, it has been loved by some and hated by others; some of the people who profess to despise it have put salt on its tail and called the result category theory. Once that's done, the abstract versions of the mappings that keep arising are dubbed (homo)morphisms, and, motivated by settheoretic notions such as injection, surjection, and bijection, a large Greek lexicon of mono-, epi-, iso-, as well as endo- and auto-morphisms goes to work. Arguably, each one of the concepts here named is an element of mathematics.

ISOMORPHISM. The recognition that two structures are "the same in some sense" - isomorphic - (recall the eigenvalue solution of the coconut problem) is surely a reasonable candidate for an element. There is a general notion in the background here; a vague way to describe it is as a manifestation of the number 1, of uniqueness. (Singly-generated structures, for example, cyclic groups, are also suggested by this circle of ideas.)

The rare and always exciting insight that something is "really" something else - for example, that probability is really measure theory, or that Fourier series are really a part of the study of locally compact abelian groups - all that belongs to the one-ness element.

[I permit myself an autobiographical disgression. When I was studying David Berg's proof that every normal operator on a Hilbert space of dimension \aleph_0 is the sum of a diagonal operator and a compact one, I kept thinking that I had seen something like this before, that the same permutation of ideas belonged to another proof, somewhere else. The insight was slow in coming, but it was worth working for. What I finally remembered was that that "other proof, somewhere else" proved the classical theorem about every compact metric space being a continuous image of the Cantor set. As soon as I saw that, I understood Berg's theorem completely, and, incidentally, I became convinced that the right road to follow was to use the theorem about the Cantor set rather than imitate its proof.] QUOTIENTS. Isomorphism is probably the structural element that occurs most often, that has the greatest visibility, but the one that I think has the greatest depth and least obvious manifestations is epimorphism. The set-theoretic language in which surjections (which is what epimorphisms usually are) are most frequently discussed is that of quotient structures (the concept that solved the associativity problem for symmetric differences). Quotient structures occur everywhere in mathematics and play a vital role: witness quotient groups in algebra, identification spaces in topology (that give us, for instance, circles by bending closed intervals), modular arithmetic in number theory, and the (already mentioned) L^p

Size

Isomorphism, as an element, contrasts with epimorphism; the distinction is between recognizing that two things are the same and identifying them so as to make them the same. From a different point of view isomorphism can be looked at as the emphasis on one-ness, or primeness, and from that point of view the proper contrast is two-ness, or (finite) many-ness, or infinity. These words are meant to point to a few of the less universal (and therefore more important?) categorical elements.

PRIMES. Primes occur in many parts of mathematics. Their first appearance for most of us is in number theory, followed soon afterward by their appearance in algebra (as, for instance, irreducible polynomials), and something like them crops up almost everywhere. The connected components of a topological space may be viewed as its prime constituents; search for "primes" among Beurling's inner functions plays an important role in operator theory; and the determination of all the primes among finite groups (that is, of the simple groups) was a gigantic job that occupied many algebraists for many years.

DUALITY. The most familiar antonym of "one" is "two", and, correspondingly, two-ness or duality should probably be viewed as an element. The word "duality" is used in projective geometry (points and lines in the plane), in category theory, in logic (memento also Boolean algebras and totally disconnected compact Hausdorff spaces), in topology (memento Alexander), in harmonic analysis (Pontrjagin, Tanaka), in Banach space theory (memento reflexivity) - and while it means something different in each of these cases, the underlying element is recognizably the same. Related but different manifestations of two-ness occur in the theory of involutions in groups and algebras, and in the study of order, equivalence, and other types of binary relations.

[A different level at which two-ness enters mathematics is visible in the two-way pull of many well-known examples and theorems. The best theorems don't just say that a strong assumption implies a desirable conclusion; they are more likely to say that despite a strong pull downward something goes upward. The theorem that $\sum_{n=1}^{\infty} \frac{1}{n^{1+\epsilon}}$ converges whenever

 $\varepsilon > 0$ is improved by the knowledge that the series diverges when $\varepsilon = 0$; the wonder of the Cantor function is that while it remains constant almost all the time it manages to grow from 0 to 1.]

In their role in primality and duality the numbers 1 and 2 are, apparently, different from all other positive integers. Bilinear forms can be generalized to trilinear ones, and the achievement is in many ways quite successful; ternary relations exist as well as binary ones (although no good general theory of them has been found yet); and in a group the elements of order three can be looked at as well as the elements of order two - but somehow it all looks unnatural and forced, like a generalization that does nothing but generalize.

PIGEONHOLE. The concept of finiteness is the next element of mathematics that I should like to call attention to. When finiteness enters mathematical reasoning it usually does so in the guise of the famous pigeonhole principle: ten bills delivered to nine mailboxes are bound to make at least one debtor extra grumpy.

The pigeonhole principle is the quintessence of "finite mathematics"; it is at the heart of the very definition (Dedekind's) of the concept of finiteness, and, in particular, at the heart of the burgeoning activity called combinatorics. Its most subtle generalizations are the Ramsey theorems, some of which have shed surprising light on the theory of sentences that are true but unprovable.

Finiteness has echoes in the most infinite parts of mathematics - it is a defensible thesis that in some sense all mathematics is finite. Compactness, for instance, a concept associated with topological spaces (usually infinite), is thought of by many mathematicians as a cleverly generalized kind of finiteness. [Incidentally, it has always seemed to me that set-theoretic topology is the right infinite generalization of combinatorics - I might go so far as to say that it *is* infinite combinatorics. As a curious piece of authoritarian evidence for this view I cite Norman Steenrod: he disliked both subjects equally, and thought lowly of them, apparently in the same way and for the same reasons.]

As another example of the finite in infinity I mention my faith (surely it is not mine alone?) that if we knew all about operator theory in spaces of *finite* dimension, then we could answer all questions about operators, even the ones (such as the invariant subspace problem) that are incontrovertibly infinite. I do not mean something as absurd as a one-to-one correspondence: the solution of the invariant subspace problem for finite-dimensional spaces does not solve the problem in general. What I mean, what \therefore must mean, is that the solution of every operator problem can be approached by finding, formulating, and answering the right, the appropriate, the truly central finite-dimensional questions that are connected with it. Case in point: Enflo's solution of the (strictly infinite) basis problem for Banach spaces showed that the difficulty was really a matter of understanding the structure of finite-dimensional examples.

INFINITY. Finiteness is an element, and so is infinity. By "infinity" I do not mean something as shallow as "the point at infinity" in projective geometry or on the Riemann sphere; there is nothing infinite about that. Just as the typically finite argument is the pigeonhole principle, the typically infinite one is induction. I say "induction" because it's a single word that points in the right direction. What I really have in mind, however, is the simplest theorem about infinite cardinal numbers that is at the same time one of the deepest theorems of mathematics, namely the statement that $1 + \aleph_0 = \aleph_0$. That's what is really at the heart of mathematical induction and is the basis of Dedekind's definition of infinity. It enters in analysis (including ergodic theory and, of course, operator theory), in algebra (for example in the classification of infinite abelian groups), in topology, and in logic - its

171

connotations embrace everything that anyone could properly call infinite.

[Here is a curious aside, parallel to the statement that in some sense all mathematics is finite: in another sense all mathematics is infinite. This is more nearly the classical view. Even an uninteresting mathematical assertion, such as 29 + 54 = 83, is a general theorem with an uncountably infinite set of special cases (or do I mean applications?). There is no contradiction: each statement uses a slightly dramatized stylistic ellipsis to make a point, and with an adequate amount of pedantic padding both statements become true and dull.]

It may be worthy of note that although in natural languages "infinite" is a negative word (not finite), whose definition must presumably come . after the positive one, in mathematics infinity is the free concept, defined in positive terms (something exists), and finiteness is the restrictive, negative one (something cannot be done).

Dedekind defined a set to be infinite if it can be put in one-to-one correspondence with a proper subset, and finite if it cannot. The idea, transplanted to the context of category theory, can be used to define generalizations of mere set-theoretic infinity and finiteness - generalizations whose study could possibly shed some light on those elements. What I have in mind can be illustrated in the category of groups by these definitions: a group is *infinitary* if it is isomorphic to a proper subgroup and *finitary* otherwise. Example: the infinite group of all those roots of unity whose exponent is a power of 2 is finitary. Similar definitions in other categories (for example, metric spaces) yield other suggestive examples of finitary objects (for example, the compact ones).

Composition

A difficult aspect of trying to discover the elements of mathematics is nomenclature. "Infinity" is probably a good name, and "geometric series" is probably a bad one. I have no good name for the next cluster of ideas I should like to recall; till a better name comes along, I'll leave them grouped under the heading *composition*. Here they are, my next batch of possible elements, and these are of the conceptual kind.

ITERATION. It all starts from the trivial observation that mappings can be composed. In other words, if we can perform each of two operations (and if their domains and ranges match properly), then there is nothing to stop us from performing them both, one after the other, and, in particular, there is nothing to stop us from performing one operation over and over again. This concept is at the root of Archimedes' theorem (if K and ε are positive numbers, then some positive integral multiple of ε is greater than K, or, informally, every little bit helps). It is at the root of the idea of integration also (the sum of many "infinitesimals" can be large), and at the root of greatly generalized integrals such as arise in spectral theory.

Another part of analysis where the notion of iteration plays a central role is the method of successive approximations. The "method" is also called Banach's fixed point theorem. Note that the element under discussion makes contact with the element of invariance: fixed points are good to look for and to find, whether the search is iterative or not.

CROSS SECTION. Composition is an important mathematical element even when it is not done infinitely often; one reason is its connection with the element that may be called transversality. The most famous instance of what I want to discuss here is the axiom of choice. Possible formulation: for every disjoint set of sets there exists a set that has exactly one member in common with each of them, or, in other words, every partition has a transversal set. (Disjointness is not always a part of the statement, but it might as well be; the versions with it and, suitably rephrased, without it, are equivalent.) It doesn't take much thought to see that the axiom can be expressed as follows: whenever f is a function from a set Y onto a set X, then there exists a function q from X into Y such that f(g(x)) = x for all x in X. (The idea is that if Y is the disjoint union of sets Y_x , with x in X, and if f(y) = x for each y in Y_{x} , then g is a choice function that assigns to each x in X a member y_r of Y_r .) In other words, every function (better: every surjection) has a right inverse, or, in a different but frequently used terminology, every function has a cross section.

The element "cross section" is a part of many mathematical compounds, subject, of course, in each case, to the additional conditions that the enveloping structure naturally imposes. (The cross section must be not only a function, but an appropriate kind of continuous, or differentiable,

or algebraically well-behaved function.) Cross-sections enter, for instance, in differential geometry (connections), in combinatorics (the marriage theorem), in algebra (semidirect products), in analysis (Stone and von Neumann had occasion to study cross sections of the mapping that associates with each measurable set its equivalence class modulo sets of measure zero), and in topology (which continuous mappings have continuous cross sections?, do they all have Borel cross sections?).

EXPONENTIAL. The idea of composition leads naturally to iteration and thence to exponentiation, and that suggests one more element associated with this circle of ideas, namely the exponential function. We all learn about e^x relatively early in calculus, and many of us are surprised and pleased by its complex properties (for example, its periodicity). A concept that plays a fundamental role not only in interest rates, but also in Banach algebras and Lie theory, is surely deserving of being considered an element of mathematics.

Analogy

There are patterns that occur and recur in mathematics that seem to lie deeper than the first ones (such as morphism) that category theory calls attention to, but that are vaguer, less well known, and less studied (so far) than some of the classical elements (such as geometric series). Worthy of mention among them are commutativity, symmetry, and continuity.

COMMUTATIVITY. Is the concept of commutativity an element? Could be. Surely the insight that what classical double limit theorems assert is in some respects similar to the relation between two translations acting on Euclidean space, and that both of them resemble the behavior of the paths in certain arrow diagrams - surely that is an insight worth having, and, perhaps, the common feature, the concept of commutativity, is an element of mathematics.

SYMMETRY. The word "symmetry" is not usually defined; mathematicians are more likely to use it informally, as they use words such as "analysis". "Proof by symmetry" is a popular phrase in which the word occurs; is there an element hiding there?

The commonly agreed context to which symmetry is assigned is group theory, and there too an element may be hiding. I do not just mean the well-established concept of group, but I mean the more lately arrived tendency to make a group out of everything; phrases such as Grothendieck group and K-theory will indicate the direction I am looking in.

CONTINUITY. Continuity is another word that seems to want to mean something even when it shouldn't, even when there is no topology present (or, in any event, no obvious one). Mathematicians seem to be disposed to believe that everything is continuous. The tendency is visible in the Kodaira-Spencer theory of deformation of manifolds, in the Riesz-Thorin interpolation theorem, and in the Kadison-Kastler study of perturbation of von Neumann algebras. (The least precise and oldest manifestation of the phenomenon is the principle of permanence of functional form already referred to.)

All these phenomena - the element candidates of commutativity, symmetry, and continuity, are instances of what might be called illegitimate generalization, or conclusion jumping by analogy. However shady their logical status, however lacking in rigor they may be, they can be valuable weapons in the arsenal of the truth-hunting mathematician: long before all else fails, be sure to ask "do they commute?", "do they have inverses?", and "do they converge to the right limit?".

Epilogue

Is what I have been saying mathematical mysticism, or is it possible that there really are some underlying guiding principles in mathematics that we should try to learn more about? I think there are such principles, but I do not know; I believe, therefore I think.

I feel like a fumbling freshman disciple of Empedocles, picking "elements" almost at random, not by controlled observation and careful analysis, but based on private experience and personal intuition. The elements I have suggested range from dim analogy (everything is continuous), through conventional wisdom (structural constituents must be compatible) and universal algebra (look for invariants, form quotient structures), to computational trickery (sum the geometric series). My own hesitant first steps were suggested by the geometric series trick, and, more generally, by the conviction that Euler's illegal manipulations with divergent series were not worthless: he had the guidance of genius, an unformulated but sharply focused insight into the fundamental elements of truth. My subsequent thoughts led to a more rarified air than Euler's solid earth, but I continue to feel that the concrete, down-to-earth candidates (exponential function, pigeonhole principle) are more likely to be elements than the abstract, airy ones (duality, infinity).

In any event, I know that, at best, all I have done was to suggest the existence of a question. If that suggests to others wider applications of the elements I have mentioned, or new elements in parts of mathematics I know nothing about, or steps toward a beautiful and useful theory of mathematical elements, then I'll have accomplished my mission.

Department of Mathematics, Indiana University at Bloomington, Bloomington, Indiana 47405, USA.