PROOF IN MATHEMATICS

Gila Hanna and Ed Barbeau, University of Toronto

1. Introduction: The Need for Proof

Observe that $1 = 1^2$, $1 + 3 = 2^2$, $1 + 3 + 5 = 3^2$. It would seem that the sum of the first n odd numbers is equal to n^2. Checking higher values of n does not disappoint us, and soon we are prepared to assert confidently what the sum of the first billion odd numbers is. Our result is supported by all our empirical evidence. Every instance we have looked at confirms it; nothing has been found to contradict it. On this basis, we accept facts in the natural sciences. Should mathematics be any different?

As any science matures, its practitioners are not content merely to record and use their individual observations. They also want to account for them, to develop rules to help understand why things are the way they are, and even to predict the results of observations not yet made. In other words, they see the need for an encompassing theoretical framework. In the natural sciences, this framework does take the form of general laws founded upon observational evidence.

For the early Egyptians, Babylonians and Chinese, in fact, the weight of observational evidence was enough to justify mathematical statements as well. But the classical Greek mathematicians felt that this way of determining mathematical truth or falsehood was less than satisfactory. They saw that mathematics, unlike other sciences, often deals with entities that are infinite in extent or number, such as the set of all natural numbers, or are abstractions, such as triangles and circles. When dealing with such entities, mathematics wishes to make absolute statements, *i.e.*, statements that apply to every instance without exception. No statement about members of an infinite set can possibly be checked in every instance. Only individual approximate representations of an abstract entity can be observed; there are infinitely many possibilities and all are subject to measurement error. Thus the ancient Greek mathematicians and their successors realized that mathematical statements, whose truth depended on empirical confirmation, would always be open to doubt despite the most thorough and acute observation.

2. The Classical Paradigm: Euclidean Geometry

The geometers of classical Greece addressed the need for a firm intellectual foundation. Their achievement is reflected most notably in the *Elements* of Euclid. They needed to avoid a situation where the validity of results depended on the experience, intuition or implicit assumptions of any individual. Geometry should be grounded in a relatively small number of fundamental statements (known as axioms) that could readily be accepted; all other propositions should be proven from these axioms by applying laws of logical reasoning.

The axioms of these geometers were inspired by, and true of, the real world. But they saw that if care were taken in choosing the axioms, defining the abstract geometric entities they dealt with, and applying rigorous reasoning, then a solid tower of new and valid propositions (known as theorems) could be build upon this foundation without further appeal to outside experience. Structures founded on these principles have come to be known as an axiomatic systems. Euclid began with a list of basic assumptions (axioms) in the form of five common notions and five postulates.

The five common notions:

1. Things which are equal to the same thing are also equal to one another.
2. If equals be added to equals, the wholes are equal.
3. If equals be subtracted from equals, the remainders are equal.
4. Things which coincide with one another are equal to one another.
5. The whole is greater than the part.

The five postulates:

1. A straight line segment can be drawn joining any two points.
2. Any straight line segment can be extended indefinitely in an straight line.

3. Given any straight line segment, a circle can be drawn having the segment as its radius and one end point as its centre.

4. All right angles are congruent.

5. Through a point not on a given line there can pass only one parallel to the line. (This is a simpler restatement of the original "parallel postulate.")

In addition, Euclidean geometry adopted 23 definitions, in which entities such as points, lines, plane surfaces, circles, and obtuse and acute angles are described. The following are examples: 1. A point is that which has no part. 2. A line is a length without breadth. 3. The extremities of a line are points. $\cdots$ 15. A circle is a plane figure contained by one line such that all the straight lines falling upon it from one point among those lying within the figure are equal to one another. $\cdots$ 22. Of quadrilateral figures, a square is that which is both equilateral and right-angled; an oblong that which is right-angled but not equilateral; a rhombus that which is equilateral but not right-angled; and a rhomboid that which has its opposite sides and angles equal to one another but is neither equilateral nor right-angled. And let quadrilaterals other than these be called trapezia. 23. Parallel straight lines are straight lines which, being in the same plane and being produced indefinitely in both directions, do not meet one another in either direction.

Though rules of inference are another essential requirement of an axiomatic system, Euclid did not incorporate such rules explicitly into his system of geometry. He assumed, instead, that his readers would be familiar with the rules of logic set out earlier by Aristotle. These rules are not part of geometry itself, but rather constitute valid ways of manipulating statements of any type, so that truth is transmitted from one statement (or set of statements) to another. In this sense they stand above the subject matter to which they are applied.

This can be illustrated by the elementary rule of logic known as *modus ponens*. This rule says simply that if a statement P implies another statement Q, and if P is true, then Q is also true. Let us assume, for example, that we know two things are true: (1) If today is Sunday (P), then I don't have to go to work today (Q), and (2) Today is Sunday (P). Then the modus ponens rule says that a third statement must be true: (3) I don't have to go to work today (Q). This scheme can be written more generally and briefly as follows (where "P" standing by itself means "P is true":

(1) P implies Q
(2) P
(3) Therefore Q

Under this rule Q is a consequence of propositions (1) and (2) only when they are taken together. It is not a consequence of proposition (1) by itself ("P implies Q"), nor of proposition (2) by itself ("P is true"). Note, too, that the modus ponens rule does not say that Q must be false if P is false (even if it is not Sunday, you may not have to go to work anyway for some other reason, perhaps because it is a public holiday).

If proposition (1) is true, however, and you do have to go to work today, you can be sure it is not Sunday. In other words, if it is true that "if it is Sunday, then I don't have to go to work today" and also true that "I have to go to work today," then "it is Sunday" must be false. This illustrates a second rule of deduction known as *modus tollens*:

(1) P implies Q
(2) not-Q
(3) Therefore not-P

The modus tollens rule is the basis for a very common method of proof in mathematics, in which a proposition is proved by showing that its falseness leads to unacceptable consequences. Wishing to prove a statement S, one first assumes that its negation "not-S" is true. One goes on to show that not-S implies Q, where Q is already known to be false. By modus tollens, then, not-S must also be false, which means that its negation, the original proposition S, must be true.

Related to this is *reductio ad absurdum*, which says that a proposition must be false if one can derive a contradiction from it. In other words, if a proposition P entails both the proposition Q and its negation,

then P cannot be true:

(1) P implies Q
(2) P implies not-Q
(3) Therefore not-P

A rule often applied in mathematics is known as *elimination of disjunction*. It sometimes happens that one does not know whether P is true or S is true, or both, but one knows that one of them must be true. If one can show both that P implies Q and that S implies Q, then Q must be true. This can be put concisely as follows:

(1) P or S
(2) P implies Q
(3) S implies Q
(4) Therefore Q

Modus ponens, modus tollens, reductio ad absurdum and elimination of disjunction are but four examples of natural rules of deduction known at the time of Euclid. In modern times they are reflected in rather more formalised systems of logic. The rules cited above, for example, form part of the "propositional calculus," in which each statement is dealt with as an indivisible whole ("P"). Other patterns of valid reasoning worked out in some detail by Aristotle (syllogisms) deal with statements that include quantifiers ("some;" "all"), and would be subsumed in the modern "predicate calculus."

Euclid's axiomatic approach is not only of historical importance. It has become a central motif in mathematics. Modern mathematical structures are often described through sets of axioms, which are required to be consistent (not yielding contradictory statements) and sufficiently rich to admit the derivation of useful and deep theorems. In such a system, a proposition is considered to be true if it can be derived from the axioms by a finite number of logical steps using permitted rules of inference.

Proof is clearly the central process in an axiomatic system, and proof is central to mathematical theory and practice largely because the axiomatic approach was adopted by Euclid and has become a key paradigm in mathematics. But proof how new mathematical truth is validated in general, even beyond the confines of an axiomatic system. Because it is a procedure and not a result, in fact, a proof can be quite valid in itself even if it starts from invalid premises. Typically, of course, a proof starts with propositions known or assumed to be true and culminates in a new proposition which, once proven, becomes a theorem and represents new mathematical knowledge.

For example, let us look at the very first proposition in Euclid's *Elements*, in which he describes a construction for an equilateral triangle, and then proves that the construction achieves its goal. It might help the reader to construct a diagram.

On a finite straight line, to construct an equilateral triangle.

Let AB be the given finite straight line. Thus it is required to construct an equilateral triangle on the straight line AB. With centre A and distance AB let the circle BCD be described [Post. 3]; again, with centre B and distance BA let the circle ACE be described [Post. 3]; and from the point C, in which the circles cut one another, to the points A, B let the straight lines CA, CB be joined [Post. 1].

Now, since the point A is the centre of the circle CDB, AC is equal to AB [Def. 15]. Again, since the point B is the centre of the circle CAE, BC is equal to BA [Def. 15]. But CA was also proved equal to AB; therefore, each of the straight lines CA, CB is equal to AB. And things which are equal to the same thing are also equal to one another; therefore CA is also equal to CB [C.N. 1]. Therefore the three straight lines CA, AB, BC are equal to one another. Therefore the triangle ABC is equilateral; and it has been constructed on the given finite straight line AB. Being what it was required to do.

3. **Discovery and Creativity**

While proof is needed to secure mathematical knowledge, the genesis of mathematical results is more complex. Mathematicians, on the basis of empirical investigations, often know what they would like to prove before they start. The effort of justifying this leads to increasing levels of analysis and familiarity with the situation, so that more possible results come to light. Euclid's result that the base angles of an isosceles triangle are equal seems hardly surprising, but the equality of two angles subtended at the circumference by an arc of a circle, might not have been foreseen by mathematicians on the basis of simple observation. Discovery and verification may influence one another, but are two different processes. Indeed, is it true, as the great French mathematician, Henri Poincaré (1854-1912) claimed, that mathematical discovery is not only independent of the axiomatic approach, but in effect stifled by it?

Intimations of new mathematical truths come to a mathematician not only while working through a proof (though that too may happen), but also from flashes of mathematical insight, from hunches or "intuition," from observation of the physical world, from experimentation with specific cases. Proof comes into its own because mathematicians cannot accept even the most plausible of such intimations until it has been verified, and the touchstone of mathematical verification is proof. To paraphrase Poincaré again, mathematicians invent by intuition and prove by logic.

But proof as well as discovery requires creativity. Even though the rules of inference are given, and even though the proof cannot proceed until the thing to be proven has been stated as a precisely worded proposition, the construction of a proof is far from being straightforward or mechanical. There are many different ways of constructing proofs, and mathematicians must draw not only upon their knowledge of mathematics and sense of logic, but also upon their imagination and inventiveness.

Despite its importance, there is really no satisfactory general account of mathematical creativity. It has been investigated in some depth, however, by both psychologists and mathematicians. Jacques Hadamard (1865-1963), for example, in his book *The psychology of invention in the mathematical field* (1954), summarized a survey he conducted among creative mathematicians, probing their own perceptions of the sources of their creativity. Without purporting to explain it, the survey explores the elusive discovery process in terms such as "noises," "pacing a room," "thinking aside," and "mathematical dreams."

Whatever detailed explanations may be offered, it is clear that mathematical creativity is less related to logic or to proof than to inspired guesswork. This is not to suggest that it is a hit-or-miss affair, a case of imagination running wild. On the contrary, in perhaps inexplicable ways it is at once both nourished and guided by everything the mathematician has ever known. It is a fortuitous juxtaposition of seemingly unrelated ideas drawn from past experience, a making of useful new combinations with known mathematical entities. The core of fruitful inventiveness is the conscious or unconscious exercise of judgment in weighing the products of the imagination. There is discernment even in inventiveness, but it is entirely clear that mathematical creativity cannot be reduced to a system, much less to an axiomatic one.

The distance between discovery and verification, and the reliance on creativity to bridge that gap, can be illustrated by the important role played in mathematics by what are known as conjectures. These are educated guesses, consistent with other mathematical statements known to be true. Conjectures may arise from experimentation, which mathematicians, like other scientists, do a lot of. In that case the conjecture is in the nature of a generalization, a statement that mathematicians believe to be true for all cases because they have seen it to be true for many cases and have never found a case for which it is not true. However, conjectures often play an important role in mathematical development by focussing attention on critical issues.

Conjectures are sometimes proven true, sometimes proven false. But there are a few well known conjectures that have resisted both proof and refutation. A case in point is the famous Goldbach conjecture. In 1742 the Prussian-born mathematician Christian Goldbach (1690-1764) wrote a letter to the Swiss mathematician Leonhard Euler (1707-1783), suggesting that "every integer n greater than 5 is the sum of three primes." Euler expressed his belief in the correctness of this statement, though he was unable to prove it. (Euler added that it can be shown to be equivalent to the statement that "every even integer n greater than 2 is the sum of two primes," and it is the latter statement that is now known as Goldbach's conjecture.)

Mathematicians have tried to resolve this conjecture for some 250 years, but no proof or refutation has been found.

4. **Types of Proof**

There are well-established techniques for constructing a proof. The most common is a *direct* argument that begins with a list of accepted facts and hypothesis and proceeds by use of modus ponens to the desired conclusion. However, there are other types that are in frequent use.

Proof by contradiction. This form of argument, often called *indirect* is based on modus tollens, or reductio ad absurdum. Suppose we wish to show that A implies B. We begin the argument by asserting that B is not true and then show that A or some other known hypothesis must then fail. To prove that the square root of 2 cannot be written as a fraction with integer numerator and denominator, for example, we begin by assuming that it can be so written, so that $2 = (p/q)^2$, where p and q are integers whose greatest common divisor is 1. One then shows, since $2q^2 = p^2$, that p and q have to be both even, thus contradicting our assumption about their greatest common divisor.

Proof by induction. To show that some result holds for an entire set of natural numbers n, a proof by induction first verifies the result for the smallest relevant value(s) of n and then shows its truth for one value of n implies its truth for the next higher value. This may be likened to climbing up a ladder rung by rung, by first making sure we can find the lowest rung, and then ascertaining that no matter which rung we have reached, we can always reach the next one.

Typical is a proof of the result that opens this essay:

$$1 + 3 + 5 + \cdots + (2n - 1) = n^2$$

(the sum of the first n odd integers is equal to n^2) for every positive integer n. Since $1^2 = 1$, the result clearly holds for the smallest pertinent value of n. If we grant this result for $n = m$, then, when $n = m + 1$, the left side becomes

$$1 + 3 + \cdots + (2m - 1) + (2(m + 1) - 1) = m^2 + (2m + 1) = (m + 1)^2,$$

which is the desired result with n replaced by $m + 1$. Thus, we can make our way up to any given value of n, by repeating the proof schemata for $m = 1, 2$ and so on until we have what we require.

A form of a proof by induction, with the flavour of a contradiction argument, is argument by infinite descent, a method pioneered by Pierre Fermat (1601-1665). In showing that there are no positive integers satisfying $x^4 + y^4 = z^2$, one uses a contradiction argument. If we assume a solution, then there is one for which z takes its smallest positive value. One then shows that, given a solution, we can construct another solution with a smaller z (the "descent" part), so that we get a contradiction of the minimality of z. The proof of the irrationality of $\sqrt{2}$ can be framed as a descent argument; given any fraction equal to this number, we can find one with a smaller denominator.

Induction and descent arguments depend on special characteristics of the natural numbers as an infinite set. The natural ordering $1, 2, 3, \cdots$ is such that (a) given any nonvoid subset, there is a least number in it, and (b) given any particular number, we can reach it in a finite number of steps by counting upwards from 1, adding 1 at a time. Property (a) says that the positive integers are *well ordered*.

The rational numbers and the real numbers with the usual ordering (modelled by their representation as points on a line) do not have these properties. However, there is an axiom known as the *Well Ordering Principle* which holds that any nonvoid infinite sets can be well-ordered, and, when this is done, proof by induction can be generalized to a process called *transfinite induction*.

Proof by example and counterexample. Often mathematicians will make a conjecture or ask a question whether some statement is true. Consider the following conjecture: *Every fraction of the form* $4/n$*, where*

n is a positive integer exceeding 2, can be written as the sum of three distinct reciprocals of integers, $1/a + 1/b + 1/c$, where a, b and c are different positive integers. For example, $1/5 = 1/8 + 1/20 + 1/40$ validates the conjecture in the case $n = 5$. It is straightforward to check the conjecture in a great many specific cases, but no one has found a general argument that applies to all values of n. However, to disprove the conjecture, all we would need is one particular value of n for which a representation of the desired type is not possible. Thus, the similar statement, *every fraction of the form $3/n$ with n exceeding 3 can be written as the sum of two distinct reciprocals of positive integers* can be shown as false because 3/7 cannot be so represented.

For another example, we observe that the numbers 31, 331, 3331 are all primes and might conjecture that any number, all of whose digits except the last are 3, is prime. However, we can disprove the conjecture by noting that $333333331 = 17 \times 19607843$ is composite.

To cite an historical example, it seemed to be accepted in the nineteenth century that all continuous functions defined on an interval had a derivative at virtually all the points in the inteval. Karl Weierstrass was able to give an example of a function that was continuous but did not have a derivative anywhere.

In presenting a proof, it is often clearer to illustrate the ideas through a particular case. Technically, a single example does not constitute a proof, and the reader should be able to understand the general argument. Here are two examples.

One can show that the sum of the first five odd numbers is 5^2 by drawing a square array of 25 dots and then indicating by lines how they array is made up of $1 + 3 + 5 + 7 + 9$ dots:

```
*   *   *   *   *
*   *   *   *   *
*   *   *   *   *
*   *   *   *   *
*   *   *   *   *
```

One can easily imagine such a diagram being given with 5 replaced by any other positive integer, to show that $1 + 3 + 5 + \cdots + (2n - 1) = n^2$ for each positive integer n.

In the second example, we consider a set of equations. There is one more square on the left than on the right side of each equation and the sum of the roots flanking the equal sign is itself a perfect square:

$$3^2 + 4^2 = 5^2$$

$$10^2 + 11^2 + 12^2 = 13^2 + 14^2$$

$$21^2 + 22^2 + 23^2 + 24^2 = 25^2 + 26^2 + 27^2$$

The reader can try to find the next few instances of this pattern. Let us show how the third equation can be obtained, by looking at the difference:

$$\begin{aligned}(25^2 + 26^2 + 27^2) - (24^2 + 23^2 + 22^2) \\ &= (25^2 - 24^2) + (26^2 - 23^2) + (27^2 - 22^2) \\ &= (25 + 24)(25 - 24) + (26 + 23)(26 - 23) + (27 + 22)(27 - 22) \\ &= 49 \times (1 + 3 + 5) = 7^2 \times 3^2 = 21^2 \ .\end{aligned}$$

While this establishes only the third equation, the way in which this example is laid out indicates how one could find and verify any of the other similar equations as well. In a formal presentation, one would use variables and actually write out the argument for the general equation with n squares on one side and $n+1$ squares on the other. The general result is expressed as

$$\sum_{i=0}^{n} (2n^2 + 2n - i)^2 = \sum_{i=1}^{n} (2n^2 + 2n + i)^2$$

for each positive integer n. The proof of its truth, given generally for n, is

$$\begin{aligned}\sum_{i=1}^{n}(2n^2+2n+i)^2 - \sum_{i=0}^{n-1}(2n^2+2n-i)^2 \\ &= \sum_{i=1}^{n}[(2n^2+2n+i)^2-(2n^2+2n-i+1)^2] \\ &= \sum_{i=1}^{n}(4n^2+4n+1)(2i-1) = (2n+1)^2\sum_{i=1}^{n}(2i-1) \\ &= (2n+1)^2n^2 = (2n^2+n)^2\ .\end{aligned}$$

Note that many readers will find it more difficult to extract from the general form of the argument the key ideas behind it.

5. **Analysis and synthesis.**

Often in approaching a result to be established, one begins with the result and works backwards to the hypothesis or to something that is already known. This process of *analysis* is very dangerous logically, as the reasoning goes from what is desired to what is given, and it is not always clear that one can validly argue in the other direction. A proper argument would have to conclude with a *synthesis*, which puts together established facts to produce a new result.

One situation in which this occurs is in the solution of equations. We begin by assuming that an equation for some unknown holds, and then reason from the equation to simpler equations from which the unknown can be easily evaluated. In many situations, this process leads to precisely the solutions of the equation, but it is possible that somewhere along the way, extraneous solutions are produced. A proper solution therefore would involve checking the candidate solutions to see which of them actually satisfy the equation. Surd equations provide an example of this phenomenon. In solving the equation, $\sqrt{x^2+3} = 3x-1$, we may square both sides to obtain $x^2+3 = 9x^2-6x+1$, or $0 = 2(4x^2-3x-1) = 2(x-1)(4x+1)$. The last equation is satisfied by $x=1$ and $x=-1/4$, but only $x=1$ satisfies the given equation.

One often approaches the proof of inequalities in this way. For example, to establish that $\sqrt{ab} \le \frac{1}{2}(a+b)$, for $a>0$, $b>0$, we may square the given inequality to find that $4ab \le a^2+2ab+b^2$ or that $0 \le a^2-2ab+b^2 = (a-b)^2$. The last statement is certainly true. However, we must go back and verify that we can reason from the positivity of $(a-b)^2$ to the required inequality. In this case, the reasoning that we used was reversible, so we can synthesize a proper argument.

Similarly, in geometry we may be required to construct some figure using ruler and compasses. This is often approached by assuming that the construction has been carried out and then looking for relationships among the geometric entities that can be exploited to make the construction. In Euclid's first proposition, cited above, we could note that if the triangle ABC were constructed, then B and C would have to be equidistant from A and so on the same circle with centre A. Thus, the analysis of the figure will be a necessary part of our arriving at a proof, but will not be part of the formal proof itself. We must describe the construction in terms of given entities, and then finish with a proof that the construction indeed is valid.

6. **The Many Functions of Proof.**

Proof in mathematics can perform a number of functions, which, though different, often exercise an influence one upon the other:

1. Verification: Validating correctness
2. Explanation: Answering the question "Why?"
3. Conviction: Removing doubt
4. Systematization: Fitting mathematical results into a wider context

5. Discovery: Inventing new results
6. Communication: Transmitting mathematical knowledge and understanding
7. Enjoyment: Meeting an intellectual challenge elegantly

Verification. This is its primary and traditional function, in which all that is needed is a series of arguments sufficiently rigorous to persuade an informed audience. As long as the conclusion follows from the hypotheses, it is correct regardless of its form or aesthetic appeal.

Explanation. A proof is more satisfactory if it not only demonstrates the truth of its assertions, but also helps to understand why the assertions are true. To do so, an explanatory proof makes use of well-known and well-understood properties of the mathematical objects involved. When it is explanatory, a proof can also contribute to systematization by bringing to light underlying relationships that place the result in its broader context. In addition, an explanatory proof may help the reader see why the result of the proof is worth knowing. No less importantly, such a proof has the advantage, because our level of conviction is directly related to our understanding, of making a proof more convincing. Often, a search for explanatory power results in a proof that is economical, that that uses only those hypotheses that are absolutely necessary. Consider the problem of determining the maximum area of an isosceles triangle whose equal sides are equal to 1. This could be done, for example, as a maximization problem in calculus and after some labour one finds that the optimum triangle has its third side equal to $\sqrt{2}$. However, a more transparent way of looking at the situation is to imagine one of the equal sides as the base of the triangle; since the base is constant, area is maximimzed when the height is maximized, and this will occur when the two equal sides are at right angle. The latter argument conveys the inevitability of the optimal configuration while the former does not.

Conviction. A demonstration is sufficient for verification, but even a demonstration in which no flaws can be found does not necessarily produce conviction. It is best if a proof can also convince its readers that it does prove what it purports to. For the purpose of conviction, a proof usually strikes a middle ground between a rigorous argument and a sketch of logical inferences. In a completely rigorous proof, valid as it may be, the essential mathematical argument necessary for conviction may be obscured by the level of detail. On the other hand too brief a sketch, while showing the general structure of the argument, may leave even the sophisticated reader in some doubt as to the validity of some of the steps. To be convincing, then, a proof must be sufficiently clear and complete, without being overly detailed. The proper balance can only be a matter of judgment.

Systematization. Another role of proof is to connect a mathematical result with a larger body of knowledge. In exercising this function, a proof may allow results initially thought to be unrelated to be seen as parts of a greater common structure based upon shared assumptions. In so doing, a proof may help expose an underlying axiomatic structure (and even help weed out logical or mathematical flaws elsewhere in that structure). Investigating the systematic aspects of proof, as we have seen, has lead to the development of alternative axiomatic systems and to insights into axiomatic systems in general. By placing results into a wider context, of course, systematization can also facilitate their communication.

Discovery. Mathematicians usually come across new mathematical truths in ways that have little to do with proof. Nevertheless, constructing a proof can turn out to pave the way to a discovery. In some cases, because the subject matter is not intuitive, the only way to generate an unexpected result through proof may be to forge ahead almost blindly with a series of logical inferences. But when creating a proof does lead a mathematician to new truths, it is almost always because the proof has offered new insight into significant underlying properties of the objects involved. Finally, the attempt to write a certain proof may reveal completely new areas for investigation.

Communication. Presenting and publishing proofs are the primary ways in which one mathematician communicates with others. A proof is useful in this role, because by its very nature it conveys the assumptions made, the definitions and rules of inference used, and the theorem to be proven. Even among mathematicians, a proof is most useful for communication if it is explanatory of its subject matter. In addition, proofs can reveal the habits of mind of their authors and the intellectual tools and resources they use. For this reason a proof can be most useful in communicating with other mathematicians when it provides an element not

required for the narrower goal of demonstration: an understanding of the thinking processes that led to its creation. The different styles of proof, and the different levels of rigour, preserved in the history of mathematics tell us a great deal about the social process of communicating mathematical knowledge.

Enjoyment. Proofs can evoke a response similar to that evoked by good art. Mathematicians particularly enjoy significant ones that fill an important gap in mathematical knowledge. They also consider, however, that a proof which meets an intellectual challenge by verifying a long-standing conjecture is worth knowing, even if the proposition proven does not have major mathematical implications. Apart from their subject matter, though, mathematicians think that some proofs are inherently better than others. This has to do, first of all, with the ability of the proof not only to demonstrate, but also to reveal, explain and ultimately to convince. In addition, mathematicians appreciate a concise proof, and in particular one that is economical, in the sense that it requires fewer assumptions that one might have thought. Mathematicians often speak of the elusive quality known as beauty or elegance. The most enjoyable proof, in the eyes of practising mathematicians, is probably one which is at once important, revealing, succinct and unforeseen.

7. **Foundational questions.**

For two thousand years, Euclid's Elements stood as a model for ascertaining mathematical truth, even in other disciplines. In practice, standards of argument have varied widely over the years and between mathematicians. Though there was still a reverence for Euclid's model, pattern and analogy also played a role in validating results. Some of the great mathematicians to whom we owe much of our mathematical knowledge often used arguments that were quite heuristic. However, as mathematics become deeper and more complex, and paradoxical situations emerged, the need to put mathematics on a solid foundation was increasingly felt. One of the most notable achievements of the nineteenth century was the work on the foundations of the calculus led by Augustin-Louis Cauchy (1789-1857) and Karl Weierstrass (1815-1897). Their work led to a greater reliance on proof in mathematical theory and practice, but at the same time to a recognition among mathematicians that the role and value of the axiomatic method in validating results needed more clarification.

Part of the re-evaluation of role of axioms and proof was spurred by Euclid's *Elements.* Mathematicians, disturbed by the complexity of the parallel postulate, tried to deduce it from the other axioms. Eventually, it dawned on them that perhaps it was independent of these axioms, that one could have consistent geometries for which the postulate was true and also for which the postulate failed (as in spherical geometry, where "lines" are interpreted as great circles). Inspired by this insight, a number of famous mathematicians, such as Adrien-Marie Legendre (1752-1833), Carl Friedrich Gauss (1777-1855), Nikolai Lobachevsky (1793-1856), Janos Bolyai (1802-1860), and Bernhard Riemann (1826-1866), developed more than one "non-Euclidean geometry," in which there are no lines or many lines through a point that do not intersect a given line not containing the point. Certain of these new geometries proved to be useful in the physical sciences. Their significance for the development of axiomatic systems was twofold. They signalled to mathematicians a definitive break from any requirement that axioms be intuitively obvious, and at the same time pointed out to them just how important are the requirements in axiomatic systems for internal consistency and very careful deduction.

A second development called into question the very integrity of Euclid's Elements, when close study revealed that he had made use of unstated assumptions. For example, in Euclid's very first proposition quoted above, his assumptions provide no guarantee that the two circles constructed will indeed intersect in some point C. As a result, mathematicians such as David Hilbert (1862-1942) undertook to bring these assumptions more fully to light and to set geometry upon a firmer foundation. In his seminal work Foundations of geometry (*Grundlagen der Geometrie*, 1899), Hilbert laid out a system of geometry based upon six primitive terms and twenty-one axioms. He was able to convince some other mathematicians that his approach, which he called the "hypothetico-deductive model," not only had allowed him to reconstruct Euclidean geometry, but could also be applied to any branch of mathematics. Indeed, Hilbert's work on foundations of geometry started a new trend to the axiomatization of other parts of mathematics (Eves and Newson, 1965).

One effect of these developments, triggered by a closer examination of Euclid's geometry, was to il-

luminate the connection between mathematics and reality. Mathematical structures, even those originally motivated by outside reality, had to stand or fall on the basis of their own internal consistency. As they became more sophisticated and less directly reflective of physical reality, intuition was more likely to fail. Thus arguments would have to be constructed more carefully, be capable of independent scrutiny and leave no gaps in the reasoning. Since so much of mathematics turns on sets and numbers, at the turn of the century, there was a program to axiomatize the number system, that is, to provide a set of axioms that is both *consistent* and *complete*. The second property requires that, given any proposition, one should be able to, on the basis of the axioms and rule of inference, determine its truth or falsity.

In 1931 a mathematician at Princeton University, Kurt Gödel, demonstrated that this program was not realizable. Given any axiom system strong enough to describe the number system, he showed that there would be *undecidable statements*, statements whose truth or falsity could not be proved from the axioms. Of course we could simply add any such statement to the axioms already in hand, to form a larger system. But then, he also showed, there would be some other undecidable statement and so on *ad infinitum*.

This was a landmark result which meant, in principle, that a mathematician working on a conjecture might be on a wild goose chase, because it might be impossible to prove or refute it. Certainly those whose world view of mathematics embraced the notion that every true statement is capable of being proved were upset by this turn of events. Most mathematicians took little heed of this in their own researches, however, and the decades since Gödel's theorem have seen a dazzling array of deep and intricate new results.

As the twentieth century dawned, there arose a concern about the validity of proofs that purport to show the existence of a mathematical entity. The best way to show that something exists is to show an example, of course. To prove that the equation $x^2 + y^2 = z^2$ has a solution in positive integers, for example, we just have to point to the case where x, y, and z are 3, 4, and 5 respectively.

Existence has often been shown in other ways, however. For example, we could assume that an entity X does not exist and then show that this assumption leads to a contradiction with some known result. In other words, we could argue as follows: either X exists or it does not, we have proven that the non-existence of X is untenable, and therefore X exists. This is an appeal to the *law of the excluded middle*, which says that there is no third alternative to the two extremes of existence or non-existence. There are other more advanced results, such as the Axiomn of Choice or the Baire Category Theorem, which prove existence of some entity while providing no method of describing it or algorithm to obtain it. This caused considerable discomfort among certain individuals such as L.E.J. Brouwer (1882-1966) and Erret Bishop (1928-1983) who tried to develop systems (*intuitionism* amd *constructivism* respectively) that avoided this type of argument. The concern raised by this sort of proof is that it provides no sense of the entity that is purported to exist.

This issue is raised in the following argument: We wish to show that there are two positive irrational numbers a and b for which a^b is rational. Consider the number $\sqrt{2}^{\sqrt{2}}$. Either it is rational or it is not. In the first case, we can take $a = b = \sqrt{2}$. In the second case, we can take $a = \sqrt{2}^{\sqrt{2}}$ and $b = \sqrt{2}$, since then $a^b = (\sqrt{2}^{\sqrt{2}})^{\sqrt{2}} = \sqrt{2}^2 = 2$. We have shown that one of the alternatives must occur without given an indication as to which.

Other proofs of existence where the mathematical entity is not described or constructed depend on such advanced results as the *Baire Category Theorem* or a hypothesis equivalent to the *Axiom of Choice*, which asserts innocuously that given any family of non-void subsets, there exists a set consisting of one element taken from each set in the family.

However, these approaches were not widely accepted and remain on the fringes of mathematical endeavour, alternative systems with their own rules of procedure coexisting with a large body of mainstream mathematics. However, there is a branch of mathematics, *foundations*, devoted to the research of different systems of logic as well as questions of consistency, independence and completeness of axiomatic systems.

8. **Issues.**

The advent of computers. Over the past thirty years, the power of high speed computers has changed the ways in which proof has been regarded. In 1976, Haken and Appel announced that they had solved the Four Colour Conjecture, that a map of connected countries on a globe could be coloured with at most four colours so that any two countries sharing a length of border were coloured differently. This settled a notorious conjecture that was over a century old. The 1976 argument was based on showing that any map had at least one of two thousand configurations that could help in the prosecution of an induction argument on the number of countries; these were sufficiently complicated that an analysis required several thousand hours of computer time. Thus, the work of Haken and Appel was beyond checking "by hand". While computer checking eventually did put the proof beyond doubt, some felt dissatisfied because this approach did not give insight into why the four colour conjecture was indeed true.

While mathematicians have always been like explorers, computers have brought this activity to the fore. Myriads of cases and examples can be processed in a trice, and new areas of mathematics, such as dynamical systems and fractal geometry, have been opened up. The apprehension of new results in some fields have outstripped the ability to prove them in traditional ways. In attempting to draw general conclusions from their explorations, these mathematicians would appear to be turning to the methods of the empirical sciences. However, it does make more urgent the question as to what can reasonably be proved and what methods should be employed. Such a radical shift in mathematical practice is entirely justified, according to Philip Davis, a mathematician who strongly advocates greater use of computer graphics. He argues that the concept of "visual proof" is an ancient one that was unfortunately overshadowed by the rise of deductive logic and deserves to regain its important place in mathematics (Davis, 1993).

But proofs themselves are organized structures, and so amenable to mathematical analysis. Assumptions and rules of inference can be codified, and there is currently a great deal of research being done on the extent to which computers themselves can be used as proof-creators and proof-correctors. One interesting and innovative way, make possible by computers, to check extremely long proofs is that of holographic proof. It consists of transforming the proof into a so-called transparent form that is verified by spot checks, rather than by checking every line. The authors of this concept have shown that it is possible to rewrite a proof in such a way that if there is an error at any point in the original proof it will be spread more or less evenly throughout the rewritten proof (the transparent form). Thus to determine whether the proof is free of error one need only check randomly selected lines in the transparent form. This can yield near-certainty, and in fact the degree of near-certainty can be precisely quantified. Nevertheless, a holographic proof is entirely at odds with the traditional view of mathematical proof, because it does not meet the requirement that every single line of the proof be open to verification.

Fallibility. Much recent discussion has focussed on the fallibility of mathematical arguments, particularly in the wake of the book, *Proofs and refutations*, by Imre Lakatos. A mythical teacher and her class are discussing the Euler relation, $V - E + F = 2$ for the numbers of vertices, edges and faces of a polyhedron. For example, a cube has $V = 8$, $E = 12$ and $F = 6$. In the hands of Lakatos, this apparently solid result seems to evaporate under the onslaught of objections and examples that seem to deny it. The moral to be drawn from Lakatos' work is not that solid proofs of mathematical results are impossible, but that a great deal of throughness, imangination and precise reasoning is necessary in order to determine the precise scope of a result, and to detect and dispose of errors.

Lakatos dramatizes what frequently occurs in mathematics. Researchers treating some body of results find that, to reduce confusion and dissension, they need to take special care in defining concepts and formulating results. Definitions try to capture the essence of what is under discussion and lay out the rules of the game. However, sometimes reasoning from the definitions takes surprising turns. For examples, many mathematicians early in the nineteenth century likely thought that a continuous real function of a real value could be differentiated on its domain of definition or even expanded in a power series. It was quite astonishing that Karl Weierstrass was able to show that there existed such functions that possessed a derivative at *no* point whatsoever. When this happens, one either revises one's perception of the concept or makes a new definition to capture the properties that one wants to deal with. In this case, one could define *smooth* functions as those with derivatives at every point or *analytic* functions as those which have power series expansions.

Social acceptance. There is a social component to the acceptance of a mathematical argument. Because of the difficulty of laying out an argument in full detail, its acceptability depends on the care and experience of the prover and the reader. Researchers spend a great deal of time trying to construct proofs that are not only correct, but appealing, insightful and easily grasped in their entirety. Even so, steps that seem self-evident may be glossed over until someone with sufficient creative imagination points out a subtle flaw or hidden assumption. It is likely that there are many papers published each year with flawed arguments. In many cases, the error, once detected, may be easily corrected and the main flow of the argument is not affected. In other cases, it will be more serious. For significant papers, the error is likely to be found, as those using them for their own research will satisfy themselves that the results are valid. Theorems that make it into the canon of widely used results will be worked over by generations of students and so can be appealed to with confidence.

A recent example of the validation process at work is the proof of Fermat's Last Theorem by the British mathematician, Andrew Wiles (1953-), working at Princeton University. In a famous marginal note, Pierre de Fermat (1601-1665) stated that the equation $x^n + y^n = z^n$ could not be solved for positive integers when n exceeded 2, but did not have room in his margin to give his proof. For a decade from the mid-1980s, Wiles laboured privately to establish this result, checking carefully for errors in his arguments. When he finally sketched out a proof, experts in the field felt that his approach was reasonable and had a good chance of working. However, they fully expected errors to be present and feared that there may be one sufficiently serious to vitiate the whole argument. Before the paper was published, a draft was distributed to a team of referees, each charged with going over some part of the paper minutely. Many errors were found. Most were corrected, but one was more serious. In the end, a significant alteration was necessary before the proof could be counted as correct and be published. Undoubtedly, as scholars continue to study Wiles' work, some improvements and simplifications will be made in the arguments.

It is often the case that theorems, particularly those that are particularly significant, are proved in different ways. While this provides additional validation, the motivation is often aesthetic or to render a result in its most general setting. Mathematicians often speak of *elegance*, the property of an argument that exhibits the beauty of pertinence and economy, methods appropriate to give insight into the problem, and, occasionally, surprise and delight. Sometimes, the theorem turns out to hold with fewer hypotheses than originally envisaged. Often a theorem is proved in a more general, and abstract setting, so that its scope covered a wide variety of similar results. For example, Karl Weierstrass showed that every continuous real function on a closed interval could be approximated as closely as desired by a polynomial. Fifty years later, this was generalized by Stone who replaced the closed interval by a compact Hausdorff space and the set of polynomials by a family of continuous functions that were closed under addition and multiplication, and had an additional "point-separating" property. Moreover, the proof of Stone's result was no more complex that proofs of the Weierstrass theorem and seemed to put its finger on the key ideas that made the result possible.

The future of proof. Holographic proofs and the creation and verification of extremely long proofs such as that of the four-colour theorem have become feasible only because of computers. Yet even these innovative types of proof are traditional, in the sense that they remain analytic proofs. They pose intriguing questions for practitioners and philosophers of mathematics. For example, Babai (1994) asks the following questions: "Are such proofs going to be the way of the future?", "Do such proofs have a place in mathematics? Are we even allowed to call them proofs?".

Others have posed similar questions. Should mathematicians accept mathematical propositions which are only highly probably true as the equivalent of propositions which have been proven true in the usual sense? If not, what is their status? Should mathematicians accept proofs that can be verified only statistically? Can mathematical truths be established by computer graphics and other forms of experimentation? Where should mathematicians draw the line between experimentation and deductive methods?

It is difficult to know just how mathematicians will eventually answer such questions, and in any case one should not expect unanimity. If a loose consensus does evolve, it will undoubtedly redefine the concept of proof to some degree. Perhaps this consensus will recognise a multiplicity of types of justification, perhaps

even one that is hierarchically ordered. But such a new consensus, even one with large remaining areas of disagreement, would not create a situation which differs in principle from that which has prevailed up too now. We have to keep in mind that there has never been a single set of universally accepted criteria for the validity of a mathematical proof. Yet mathematicians have been united in their insistence on the importance of proof. This is an apparent contradiction, but mathematics has lived with this contradiction and flourished. Why would one expect or want this to change?