Part VIII
Final Perspectives

VIII.1 Mathematical Writing
Timothy Gowers

1 Introduction

The purpose of this article is not to offer advice about how to write mathematics well. Such advice can be found in many places. However, I do have three pieces of very general advice, which inform the rest of the article. The first is to be clear about your intended readership; for example, if you want what you write to be understood by an undergraduate, then do not assume knowledge of any terminology that is not standardly taught in undergraduate mathematics courses. The second is to aim for this readership to be as wide as possible. If, with a small amount of extra explanation, you can make what you write comprehensible to an expert in another field of mathematics, then put in that extra effort. Whatever you do, do not worry that experts will not need the explanation; if that is true (which it often is not), then they can easily skip it. The third, which is related to the second, is to set the scene before you start. Some people who read what you have written will do so because they want to understand all the technical details and use them in their own work, but the vast majority will not. Most readers, including people who need to make quick judgments that will profoundly affect your career, will want to read the introduction quickly to see what you have done and assess how important it is. However much you might wish everyone to read what you have written in complete detail, you should be realistic and cater for readers who just want to skim it.

For the rest of this article I shall discuss various choices that one must make when writing a mathematical document. I will not advocate choosing one way rather than another, since the choices you should make depend on what you want to achieve; my final piece of general advice is merely that you should make the choices consciously rather than by accident.

2 Formality versus Informality

There are (at least) two goals that one might have when writing a mathematical document. One is to establish a mathematical result by whatever means are appropriate to the field; in pure mathematics the usual requirement is unambiguous definitions and rigorous proofs, whereas in applied mathematics other forms of evidence, such as heuristic arguments and experimental backing, may be acceptable. The other is to convey mathematical ideas to the reader.

These two goals are often in tension. If a pure mathematician discovers a complicated proof of a theorem, then that proof will be hard to understand. However, sometimes the apparent complication of a proof is misleading; what is really going on is that the author had one or two key ideas, and the complication of the argument is the natural working out of the details of those ideas. For the expert reader, an informal explanation of the ideas that drive the proof may well be more valuable than the proof itself.

Nobody would advocate writing papers with *just* informal explanations of ideas, since plausible-looking ideas often turn out not to work. However, there is still a choice to make, since it *is* considered acceptable to display proofs and not explain the underlying ideas. There may sometimes be circumstances where this is appropriate; for example, perhaps the proof is short, and explaining the ideas that generate it will double the length of what you are writing and put off readers. But usually, the advice I gave earlier—to broaden your readership if it not too difficult to do so—would dictate that technical arguments should be accompanied by informal explanations.

3 Giving Full Detail versus Leaving Details to the Reader

When you are writing you need to decide how much detail to give. If you give too little, then the reader you are aiming at will not be able to understand what you have written. But you also want to avoid making too many points that the reader will find completely obvious.

Of these two potential problems, the first is undoubtedly more serious. It is much easier for readers to skip details that they find too obvious to be worth saying than it is for them to fill in details that they do not find obvious at all.

The question of how much detail to give is related to the question of how formal to be, but it is not the same question. It is true that there is a tendency in informal mathematical writing to leave out details, but with even the most formal writing a decision has to be made about how much detail to give; it is just that in formal writing one probably wants to signal more carefully when details have been left out. This can be done in various ways. One can use expressions such as "It is an easy exercise to check that...," or "The second case is similar," which basically say to the reader, "I have decided not to spell out this part of the argument." One can also give small hints, such as "By compactness," or "An obvious inductive argument now shows that...," or "Interchanging the order of summation and simplifying, we obtain...."

If you do decide to leave out detail, it is a good idea to signal to the reader how difficult it would be to put that detail in. A mistake that some writers make is to give references to other papers for arguments that can easily be worked out by the reader, without saying that the particular result that is needed is easy. This is straightforwardly misleading; it suggests that the best thing to do is to go and look up the other paper when in fact the best thing to do is to work out the argument for oneself.

4 Letters versus Words

The following is problem 10 of book 1 of an English translation of Diophantus's *Arithmetica*.

> Given two numbers, to add to the lesser and to subtract from the greater the same (required) number so as to make the sum in the first case have to the difference in the second case a given ratio.

A modern writer would express the same problem more like this.

> Given two numbers a and b and a ratio ρ, find x such that $a + x = \rho(b - x)$.

The main difference between these two ways of describing the problem is that in the second formulation the numbers under discussion have been given *names*. These names take the form of letters, which allow us to replace wordy expressions such as "the second number" or "the given ratio" by letters such as "b" and "ρ."

The advantage of modern notation is that it is much more concise. This is not just a matter of saving paper; the extra length of "to make the sum in the first case have to the difference in the second case a given ratio" over "such that $a+x = \rho(b-x)$" makes it significantly harder to understand because it is difficult to take in the entire phrase at once.

However, the concision that comes from naming mathematical objects comes at a cost: one has to learn the names. In the example above, that is very easy and the cost is negligible. However, sometimes it is far from negligible. The following proposition comes from a paper in Banach space theory.

Proposition. *Let* $0 \leqslant \alpha \leqslant \frac{1}{2}$ *and* $1/p = \frac{1}{2} - \alpha$. *Then*

$$\mathfrak{P}_2(E,F) \subset \mathfrak{L}^{(a)}_{p,\infty}(E,F)$$

for all Banach spaces F *if and only if* $E \subset \Gamma_\alpha$.

Just before the proposition, the reader has been told that $\mathfrak{P}_2$ is the ideal of 2-summing operators from E to F, which is a standard definition in the area. As for $\mathfrak{L}^{(a)}_{p,\infty}(E,F)$, this has been defined early in the paper as follows. (It is not necessary to understand these definitions to understand the point I am making.)

> Given an operator T, the *approximation number* $a_n(T)$ is defined to be $\inf\{\|T - L\| \colon \operatorname{rank}(L) < n\}$. Then $\mathfrak{L}^{(a)}_{s,w}(E,F)$ is the set of operators T such that the sequence $(a_n(T))_{n=1}^{\infty}$ belongs to the Marcinkiewicz space $\ell_{s,w}$.

The definition of the Marcinkiewicz space is again standard in the area. Finally, the set Γ_α is defined to be the set of all Banach spaces of weak Hilbert type α. That is not a standard definition, but it is given earlier on in the paper.

Thus, another way of stating the proposition is as follows.

Proposition. *Let* $0 \leqslant \alpha \leqslant \frac{1}{2}$, *let* $1/p = \frac{1}{2} - \alpha$, *and let* E *be a Banach space. Then the following two statements are equivalent.*

(1) *For every Banach space* F *and every 2-summing operator* $T \colon E \to F$, *the sequence* $(a_n(T))_{n=1}^{\infty}$ *of approximation numbers belongs to the Marcinkiewicz space* $\ell_{p,\infty}$.
(2) E *is a space of weak Hilbert type* α.

This time, it is the *wordier* definition that places a smaller burden on the reader's memory. If you know what 2-summing operators, approximation numbers, the Marcinkiewicz space, and weak Hilbert type are, most of which are standard definitions in the area, then you can understand without much effort what the proposition is claiming. With the first formulation, there is an extra step you have to perform to unpack the notation into those standard definitions. Another advantage of the second formulation is that a sufficiently expert reader who is skimming the paper will be able to understand it without having to look back in the paper to find out what everything means. The first formulation does not leave that option open.

Thus, in more complicated mathematical writing, there is another source of tension. If you use too little notation, your sentences will become hopelessly clumsy and repetitive, but if you use too much, you are placing excessive demands on the memory of your readers.

This may be a delicate balance to strike, but there is one principle that applies universally: if you do decide to use some nonstandard notation, then make sure that the reader can easily find where it is defined. This can be done by means of a section devoted to preliminary definitions, though it will often be kinder to give definitions just before they are used. If that is not possible, one can give reminders of definitions, or at the very least pointers to where they can be found.

5 Single Long Arguments versus Arguments Broken Up into Modules

If you are trying to justify a mathematical statement and the justification is long and complicated, then what you write may well be hard to understand unless you can somehow break the argument up into smaller "modules" that fit together to give you what you want. In pure mathematics, these modules usually take the form of *lemmas*. If you are proving a theorem and you do not want the proof to become unwieldy, then you try to identify parts of the argument that can be extracted and proved separately. One can then simply quote these results in the main argument. Lemmas play a role in proofs that is similar to the role of subroutines in computer programs.

For breaking up an argument to be a good idea, it greatly helps if the part of the argument you want to extract is not too context dependent. If the statement of a lemma requires a long piece of scene setting, then it is probably better to leave it in the main body of the argument, where the scene has already been set. However, if it can be stated without reference to the particular context, which usually means that it is more general than the particular application needed of it in the main argument, then it is more appropriate to extract it. Again, this is a matter of judgment.

A disadvantage of more modular arguments is that extracting lemmas, or more general modules, forces you to put them somewhere where they do not arise naturally. If you put them before the main argument, so that they will be available when needed, then the reader is presented with statements of no obvious use and is expected to remember them. If they are particularly memorable, then that is not a problem, but often they are not; for instance, they may depend on two or three slightly odd conditions that just happen to be satisfied in the later argument. If you put them after the main argument, then the reader keeps being told, "We will prove this claim later" and reaches the end of the argument with the uneasy feeling that the proof is incomplete. A third possibility is to state and prove lemmas *within* an argument, but nested statements of this kind can be fairly ugly.

With some complicated arguments, there may be no truly satisfactory solution to these problems. In that case, the best thing to do may well be to choose an unsatisfactory solution and mitigate the problems somehow. The default option is probably to state lemmas before they are used. If you choose that option and the lemmas are somewhat complicated and hard to remember, then you can always add a few words of explanation about the role that the lemma will play. If even that is hard to do, then another option is to advise the reader to read the main argument first and return to the lemma only when the need for it has become clear. (An experienced reader may well do that anyway, but it is still helpful to be told by the author that it is a good approach to understanding the argument.)

6 Logical Order versus Order of Discovery

Suppose you wish to present the fact that a sequence of continuous functions that converges pointwise does not have to converge uniformly. Here is one way that you might do it.

Theorem. *There exists a sequence of continuous functions* $f_n\colon [0,1] \to [0,1]$ *that converges pointwise but not uniformly.*

Proof. For each positive integer n and each $x \in [0,1]$, let $f_n(x) = nxe^{-nx}$. Then, for $x > 0$ we have $e^{-x} < 1$, so $ne^{-nx} = n(e^{-x})^n \to 0$ as $n \to \infty$. It follows that $f_n(x) \to 0$ as $n \to \infty$. Also, when $x = 0$ we have $f_n(x) = 0$ for every n, so again $f_n(x) \to 0$ as $n \to \infty$. Therefore, $f_n(x) \to 0$ pointwise.

However, the convergence is not uniform. To see this, observe that $f_n(n^{-1}) = e^{-1}$ for every n. Thus, for every n there exists x such that $|f_n(x) - 0| \geqslant e^{-1}$. □

This proof has a feature that is common in mathematics: it is easier to follow the steps than it is to see where the steps came from. If you are told to try the functions $f_n(x) = nxe^{-nx}$, then checking that they satisfy the conditions is a straightforward exercise, but what made anybody think of that particular sequence of functions?

Here is what we might write if we wanted to make the answer to that last question clearer.

Proof. If $f_n \to f$ pointwise but not uniformly, then $f_n - f \to 0$ pointwise but not uniformly, so we may as well look for functions that converge to zero. In order to ensure that they do not converge uniformly to zero, we need a positive number θ such that for infinitely many n there exists $x \in [0,1]$ with $|f_n(x)| \geqslant \theta$. Since infinitely many of these $f_n(x)$ will have the same sign, and since we can multiply all functions by θ^{-1}, we may as well look for a sequence of functions f_n that converges pointwise to 0 such that for every n there exists x_n with $f_n(x_n) \geqslant 1$.

Now, if $f_n(x_n) \geqslant 1$ and f_n is continuous, then there exists an open interval $I_n = (x_n - \delta_n, x_n + \delta_n)$ around x_n such that $f_n(y) \geqslant \frac{1}{2}$ for every $y \in I_n$. We are going to have to make sure that we do not have infinitely many of these intervals overlapping in some point u, since then we would have $f_n(u) \geqslant \frac{1}{2}$ for infinitely many n, which would imply that $f_n(u)$ does not tend to zero.

How can we find infinitely many open intervals without infinitely many of them overlapping? The simplest way of doing it is to take intervals of the form (a, b_n) for a sequence (b_n) that converges to a. So, for example, we could take I_n to be the interval $(0, 1/n)$.

This suggests that we should let f_n be a continuous function that takes the value 1 somewhere inside the interval $(0, 1/n)$ and is small outside that interval. One way of defining a function that reaches 1 for a small value of x and then quickly drops back down again is to take a function that grows rapidly to 1, such as $g_n(x) = \lambda x$, and multiply it by a function that is roughly 1 for a little while and then decays rapidly, such as $e^{-\mu x}$. The rapid decay of $e^{-\mu x}$ starts when x is around $1/\mu$, which suggests that we should take μ to be around n. Since we want $g_n(x)$ to reach 1 in the interval $(0, 1/n)$, we should probably take λ to be around n as well.

It is now easy to check that the functions $f_n(x) = nxe^{-nx}$ converge pointwise to zero but not uniformly. □

Of course, one might well give a detailed proof that the functions nxe^{-nx} do the job.

As with the other choices, there are advantages and disadvantages that need to be weighed up when deciding how much to explain the origin (or at least a possible origin) of the ideas one presents. If one's main concern is *verification* of a result—that is, convincing the reader of its truth—then it may not matter too much where the ideas come from as long as they work. But if the aim is to *teach the reader* how to solve problems of a certain kind, then presenting solutions that appear out of nowhere as if by magic is not helpful. What is more, demonstrating where the ideas come from gives the reader a much clearer idea of which features are essential and which merely incidental. For example, in the argument above it is clear from the second presentation that there is nothing special about the functions $f_n(x) = nxe^{-nx}$: for $f_n(x)$ one could take any non-negative function such that $f_n(0) = 0$, $f_n(1/n) \geqslant c$ (for some fixed constant c), and $f_n(x)$ is small for every $x \geqslant 2/n$. For instance, one could take a "witch's hat" that equals nx when $0 \leqslant x \leqslant 1/n$, $2 - nx$ when $1/n \leqslant x \leqslant 2/n$, and 0 when $2/n \leqslant x \leqslant 1$.

That is not to say that a diligent reader cannot look at a presentation of the first kind and work out for him/herself where the idea might have come from. In this case, if one sketches the graph of $f_n(x)$, one sees that it grows and shrinks rapidly in a small interval near 0 and is small thereafter, and then it becomes clear why these functions are suitable. However, one needs experience to be able to do this with an argument. So the extent to which you should explain where your arguments come from depends largely on the level of experience of your intended reader—both generally and in the specific area you are writing about.

7 Definitions First versus Examples First

Suppose that one wanted to write an explanation of what a topological manifold is. An obvious approach would be to start by giving the definition. That could be done as follows.

Definition. A *d-dimensional topological manifold* is a topological space X such that every point x in X has a neighborhood that is homeomorphic to a connected open subset of $\mathbb{R}^d$.

Having done that, one would give a few examples of topological manifolds, such as spheres and tori, to illuminate the definition.

An alternative approach is to start with a brief discussion of the examples. One could point out, for instance, that it is easy to come up with a satisfactory coordinate system for any small region of the world but that it is not possible to find a good coordinate system for the world in its entirety; there will always be annoying problems such as the poles not having well-defined longitudes. A discussion of that kind will give the reader the informal concept of a space that is "locally like $\mathbb{R}^d$" and after that the formal definition is motivated: it is the formal expression of an informal idea that the reader already has.

The advantage of the second approach is that an abstract definition is often much easier to understand if one has a good idea of what it is abstracting. One will read the definition with strong expectations of what it will look like, and all one will have to commit to memory is the ways in which the definition does not quite fit those expectations. If the definition is presented first, then one will be expected to hold the whole thing in one's head, rather than what one might think of as the difference between the definition and one's prior expectation of it.

Whether or not this advantage makes it worth presenting examples before giving a definition depends on how difficult you expect it to be for your reader to grasp the definition. To give an example where it might not be worth giving examples first, suppose that you want to introduce the notion of a commutative ring for a reader who is already familiar with groups and fields. A natural way of doing it would be to list the axioms for a commutative ring and make the remark that what you have listed is very similar to the list of axioms for a field but you no longer assume that elements have multiplicative inverses, and sometimes you do not even assume that your rings have multiplicative identities.

Once you have said that, it will still be a very good idea to give some important examples, such as the ring $\mathbb{Z}$ of all integers, the ring $\mathbb{Z}[x]$ of all polynomials with integer coefficients, and the ring $\mathbb{Z}[\sqrt{2}]$ of all numbers of the form $a + b\sqrt{2}$ where a and b are integers. However, the argument for presenting these examples *first* is weaker than it was for topological manifolds, for two reasons.

The first reason is that the definition is easy to grasp: rings are like fields but without multiplicative inverses. Therefore, giving the definition straight away does not place a burden on the reader's memory. Of course, the reader will want reassurance that there are interesting examples, but that can be given immediately after the definition.

The second reason is that the necessity for this particular abstraction is less obvious than it is for manifolds. Given examples such as spheres and tori, it is natural to think that they are all examples of the same basic "thing" and then to try to work out what that "thing" is. But the benefits of thinking of the integers and the polynomials with integer coefficients as examples of the same underlying algebraic structure are not clear in advance; they become clear only after one has developed a considerable amount of theory. So it is more natural in this case to think of the abstraction as primary, at least in the first instance.

As ever, the decision about how to present a new mathematical concept involves a judgment that is sometimes quite delicate. Broadly speaking, the harder a definition is to grasp, the more helpful it will be to the reader to have some examples in mind when reading it. But that depends both on the reader and on the intrinsic complexity of the definition. However, one general piece of advice is still possible here, which is at least to *consider* the possibility of starting with examples. It may not always be appropriate to do so, but many mathematical writers like to start with definitions under all circumstances, and the result is that many expositions are harder to understand than they need to be.

Let me close this section by pointing out that the examples-first device is quite a general one. Indeed, I have used it in a number of places in this article; see the openings of sections 4 and 6 and of this very section.

8 Traditional Methods of Dissemination versus New Methods

A person who wishes to produce mathematical writing today faces a choice that did not exist twenty years ago. Until recently, almost all mathematical writing took the form of books or journal articles. But now the Internet has given us new methods of dissemination, which have already had an impact and are likely to have a much bigger impact in the future.

In particular, the existence of the Internet affects every single one of the considerations discussed in this article. Let me take them in turn.

8.1 Level of Formality

The main task of each generation of mathematicians is to add to the body of mathematical knowledge. However, there is a second task that is almost as important as the first, and not entirely separate from it, which is to digest this new knowledge and present it in a form that subsequent generations will find as easy as possible to grasp. This process of digestion can of course happen many times to the same piece of mathematics.

Sometimes, digesting a piece of mathematics is itself a significant advance in mathematical knowledge. For example, a theory may be developed that yields quite easily a number of already existing and seemingly disparate results. The traditional publication system is well suited to this situation; one can just write an article about the theory and get it published in the normal way.

Sometimes, however, digesting a piece of mathematics does not constitute a mathematical advance. It can be something more minor, such as thinking of a way of looking at an argument that makes it clearer where the ideas have come from or drawing an informal analogy between one piece of mathematics and another that is simpler or better known. Insights of this kind can be hard earned and extremely valuable to other mathematicians, but they do not lead to publishable papers.

With the Internet, there are many ways that more informal mathematical thoughts can be shared. An obvious one is to write a conventional mathematical text and make it available on one's home page. Another option, which an increasing number of mathematicians have adopted, is to have a blog. The advantage of this is that one obtains feedback from one's readers, and experience has shown that the quality of much of this feedback is very high.

There are other forms of mathematical literature that would not be conventionally publishable but that could be extremely valuable. For example, an article about a serious but failed attempt to solve a problem would not be accepted by a journal, and the result is a great deal of duplication of work; if the problem is important and the attempt looks plausible to begin with, then many people will try it. A database of failed proof attempts would be very useful, and in principle the Internet makes it easy to set up, though so far nobody has done so.

In general, the Internet allows us much greater freedom in choosing the level of formality at which we wish to write and allows us to publish documents that do not fit the mould of a standard journal article.

8.2 Level of Detail

Suppose that you use a mathematical result or definition that will be familiar to some readers but not to others. In a print document you have to decide whether to explain it and, if so, how elaborate an explanation to give.

In a hyperlinked document on the web, one is no longer forced to make this choice. One can write a version for experts, but with certain key words and phrases underlined, so that readers who need these words and phrases explained further can click on them and read explanations. This kind of writing has become very common on Wikipedia and other wikis.

It also introduces a new balance that needs to be struck. Sometimes wiki articles are hard to read because the writers use the existence of links to other wiki pages as a license not to explain terms that they might otherwise have explained. The result is that unless one is familiar with most of the definitions in the original article, one can get lost in a complicated graph of linked wiki pages as one finds that the page that explains an unfamiliar concept itself requires one to click through to several other pages. So if you are going to exploit hyperlinks, you need to think carefully about what the experience of following those links will be like for your intended readers.

Another inconvenience of hyperlinks is that they require you to visit an entirely new page, which makes it easy to forget where you were before (especially if you backtrack and then follow some other sequence of links). However, there is plenty of software that gets round this problem. For example, on some sites one can incorporate "sliders," pieces of text that insert themselves into what you are reading when you click on an appropriate box and disappear when you click on it again. So if, for example, one wrote, "by the second isomorphism theorem," one could have a box with the words "What does that say?" on it, so that readers who needed it could click on the box and have a short paragraph about the second isomorphism theorem inserted into the text. One can have sliders within sliders, so perhaps within that slider one could have the option of bringing up a proof of the theorem as well.

The main point is that the Internet has made it possible to write new kinds of documents where one is no longer forced to make choices such as of how much detail to give. One can leave that decision to the reader. Such documents have a huge potential to improve the way mathematics is presented, and this potential will only increase as technology improves.

8.3 Letters versus Words

I will not say much about this, since most of what I have to say is very similar to what I have already said about the level of detail in which a document is written. With the kinds of electronic documents that are now possible, one can save the reader the trouble of searching through a paper to find out what a letter stands for by incorporating a reminder that appears when you click on the letter. Perhaps better still, it could appear in a little box when you hover over the letter. One could also have condensed statements involving lots of letters with the option of converting them into equivalent wordier statements. Again, the point is that there are many more options now.

8.4 Modularity

The kinds of electronic documents I have been discussing make possible a form of top-down mathematical writing that would be far less convenient in a print document. One could write a high-level account of some piece of mathematics, giving the reader the option of expanding any part of that high-level account into a lower-level account that justifies it in more detail. And there could be many levels of this, so that if you clicked on everything you would end up with a presentation of the entire argument in full gory detail.

A less ambitious possibility is one that solves the problem discussed earlier about where to place a lemma. The difficulty was that in a print document you will either put it before the proof where it is used, in which case it is not adequately motivated, or during the proof, in which case it looks ugly, or after the proof, in which case the proof itself leaves you with awkward promises to fill in gaps later. But with an electronic document, putting a lemma exactly where it is needed is no longer ugly. During the proof, one can say, "We are now going to make use of the following statement," and give the reader a button to click on that will bring up a proof of that statement.

8.5 Order of Presentation

If you do not want to decide whether to give an abstract definition first or start with motivating examples, then you can give the reader the choice. Just start with a page of headings and invite the reader to decide whether to click on "Motivating examples" first or "The formal definition" first.

To some extent, the same goes for the decision about whether to present arguments in their logical order or in a way that brings out how they were discovered. If at some point the logical order requires you to draw a rabbit out of a hat, you could at the very least introduce a slider that explains where that rabbit actually came from.

VIII.2 How to Read and Understand a Paper

Nicholas J. Higham

Whether you are a mathematician or work in another discipline and need to use mathematical results, you will need to read mathematics papers—perhaps lots of them. The purpose of this article is to give advice on how to go about reading mathematics papers and gaining understanding from them.

The advice is particularly aimed at inexperienced readers. A professional mathematician may read from tens to hundreds of papers every year, including published papers, manuscripts sent for refereeing by journals, and draft papers written by students and colleagues. To a large extent the suggestions I make here are ones that you naturally adopt after reading sufficiently many papers.

Mathematics papers fall into two main types: primary research papers and review papers. Review papers give an overview of an area and usually contain a substantial amount of background material. By design they tend to be easier to read than papers presenting new research, although they are often longer. The suggestions in this article apply to both types of papers.

1 The Anatomy of a Paper

Mathematics papers are fairly rigid in format, having some or all of the following components.

Title. The title should indicate what the paper is about and give a hint about the paper's contributions.

Abstract. The abstract describes the problem being tackled and summarizes the contributions of the paper. The length and the amount of detail both vary greatly. The abstract is meant to be able to