How to read mathematics

Alet Roux
University of Hull

18 November 2003

Reading mathematics is difficult, because mathematics is difficult. Since mathematics is a subject that builds upon itself, books and papers in mathematics often presuppose different knowledge than what a specific reader actually possesses. Moreover, due to the succinct, dense style of mathematical writing, reading a mathematical text requires more involvement from the reader than most texts in other subjects. Consequently, reading mathematics takes longer than other types of texts; often, attempting to read mathematics too fast results in frustration without comprehension.

According to Adler (1967, p. 119), if a text "is worth reading at all, it is worth three readings at least". By "reading", he implies a reading in manner, namely the emphasising of a particular aspect of a text during the reading process. In short, the three ways of reading a book can be summarised as follows.

> In the first place, you must be able to grasp what is being offered as knowledge. In the second place, you must judge whether what is being offered is really acceptable to you as knowledge. In the other words, there is first the task of understanding the book, and second the job of criticizing it. ... The process of understanding can be further divided. To understand a book, you must approach it, first, as a whole, having a unity and a structure of parts; and, second, in terms of its elements, its units of language and thought. (Adler, 1967, pp. 123–124)

These readings tend to coalesce, especially the first two. In essence, the first two readings lead the reader on a circular route, starting from the whole of the text, dividing into all the parts, putting them back together again, ending with the whole. The third reading is aided by observations during the first two readings; however this reading can never become absolutely simultaneous with the other two, as understanding a text must always precede its criticism.

We now consider each of these readings separately. By "text", we mean any piece of mathematical writing, understood as to include books, papers, theorems, definitions and proofs.

1 Structural or analytic reading

In this reading, the reader proceeds from the whole of the text to its parts. To accomplish the first reading, you must complete the following tasks (adapted from Adler, 1967, pp. 124, 268):

1. Find out what subject matter is covered by the text, i.e. what kind of book it is.
2. Identify the purpose of the text, e.g. what the main problems are that the author is trying to solve. Find out, in essence, what the text is saying.
3. Identify the individual parts or components of the text, and analyse these parts in the same way as the whole.

We now consider each of these tasks in turn.

1.1 Classification according to subject matter

Especially at the beginning of your research, the books you read would have been recommended by your supervisor, making this step almost superfluous. However, the classification step becomes very important when you start choosing your own reading, as this may provide the first clue as to the suitability of a piece.

In many journals, articles contain classification numbers (for instance, JEL or MSC followed by a number) pointing to their subject matter. In cases where you are doubtful, it may be worthwhile to look up these numbers. However, some questions should be raised as to the general readability of an article if its subject cannot be deduced from its title, abstract and introduction.

1.2 Capturing purpose and essence

You must be able to summarise, as succinctly as possible, what the text is trying to say. Secondly, you should discern its purpose, for instance, whether it is to solve a specific problem or problems or to provide an exposition of a particular field of mathematics.

In the case of books and papers, the essence may be distilled by scanning the cover, abstract, table of contents, introduction and conclusion, and considering the intended audience (Wicks, 2003). It is well known that mathematics papers are harder to read than books and harder to understand than lecture courses. Körner (1998, pp. 8–9) gives a number of reasons for this, all of which are related to the difference in purpose of the two types of mathematical literature:

1. The aim of a paper is to stake out a claim. Consequently, the writer will state and prove as strong[1] a version of its main theorem as possible. Quite often, the proof of a slightly weaker theorem, while still containing the essence, turns out to be much easier.
2. The emphasis of a paper is on novelty. Although a theorem or its proof is best understood in context, a paper will not discuss what is already known, nor will it contain routine arguments.
3. A paper focuses on a very specific area of mathematics, whereas the aim of a lecture course or book is to give an overview.

[1] The strength of a theorem is measured by the strength of its conclusion and the weakness of its assumptions.

> In a lecture course [or textbook], the lecturer [author] is like a conductor blending the contribution of many individuals into a harmonious whole. A single paper represents the contribution of the double bass or the triangle. (Körner, 1998, pp. 8–9)

4. A paper contains new mathematics. Consequently, the author may not fully understand what is important or what is difficult, and how the exposition is best made.

A proof is a piece of mathematical writing with the purpose of substantiating the claim (theorem or proposition) that it aims to prove (its essence). Simply put, a proof is a mini-paper with a very specific conclusion; in fact, many papers are just long proofs.

Understanding the purpose of a definition or result is intricately linked with deciphering its meaning. Detailed investigation of the content of definitions and results is perhaps best left to the second reading; at this stage, a short note on the purpose of results and definitions suffices.

A definition is "an agreement, by all parties concerned, as to the meaning of a particular term" (Solow, 1990, p. 29). Definitions do not contain any new knowledge or insight; in general terms, the purpose of a definition is to simplify the exposition of what follows.

Results are referred to by using different terms, indicating their purpose. Partially following Solow (1990, p. 37), we distinguish the following types of claims:

1. A *proposition* is a true statement of interest that has been (or is being) proven; this is the most common type of result.

2. Propositions considered to be extremely important are referred to as *theorems*: the number of propositions in any given text is likely to exceed the number of theorems by far.

3. The proof of a theorem can be very long: it is often easier to communicate the proof by means of several supporting propositions, or lemmata. A *lemma*, therefore, is simply a preliminary proposition that is to be used in the proof of a theorem, more often than not in the proof of several theorems.

4. Once a theorem has been established, it is often the case that certain propositions follow almost immediately as a result of knowing that the theorem is true. These are called *corollaries*.

5. Certain statements are accepted without a formal proof; they are called *axioms*. Perhaps the most often-used axiom in geometry is assuming that the shortest distance between two points is a straight line.

6. A *conjecture* is a claim that cannot be substantiated (yet), but is nevertheless believes to be true by the author. Most theorems start their careers as conjectures.

1.3 Enumeration of the major parts

The final step in the analytic reading at each level is to find the major constituent parts of the text, to label them and to analyse their structure in turn. As most modern mathematical textbooks are written in a labelled rather than a narrative style, this step is perceived as straightforward. However, it can be deceptively difficult: watch out for definitions hiding in theorems, and theorems masquerading as definitions.

At this point, before you start interpreting the different elements of the text, it may be useful to summarise the structure of the text. This can be done mentally, verbally (cf. Academic Resource Center, Wheeling Jesuit University, 2001), graphically (cf. Wicks, 2003) or by physically writing it down (cf. Adler, 1941).

2 Interpretative or synthetic reading

In the interpretative or synthetic reading, the reader proceeds from the parts to the whole. Again following Adler (1967, p. 125), this reading consists of the following tasks:

1. Discovering and interpreting the most important words in the text.
2. Discovering and interpreting the most important results in the text.
3. Reconstructing and understanding the arguments in the text.
4. Determining which of the goals of the text were attained.

Notice that the parts you start with to construct the whole in this reading are not necessarily the same as the parts that you come to by analysing the whole in the first reading. In the structural reading, the parts are the building blocks of the author's treatment of the subject matter. In this reading, the parts we consider are the author's ideas, assertions and arguments, as expressed in the definitions, theorems and proofs in the text.

2.1 Coming to terms

Before even attempting to understand mathematical terms, you should learn the mathematical symbols commonly used in your branch of mathematics: the obvious ones, like $\sum$ for summation and $\prod$ for multiplication, but also less well-known symbols, like $\otimes$ and $\vee$. This includes the names of the Greek letters; again not just the obvious ones like α, β, and π, but also the more obscure ψ, ξ, τ and ω. Learn not just to recognise these letters, but also to pronounce them: "it is hard to think about formulas if they are all 'squiggles' to you" (Hubbard and Burke Hubbard, 2002, p. 2).

The language of mathematics is very precise; this implies that a particular word will be used in one sense and one sense only in a particular text. However, bear in mind that the meaning of a word in a mathematical context may differ dramatically from its meaning in everyday use. For instance, most dictionaries define the word "martingale" as follows.

1. The strap of a horse's harness that connects the girth to the noseband and is designed to prevent the horse from throwing back its head.
2. *Nautical.* Any of several parts of standing rigging strengthening the bowsprit and jib boom against the force of the head stays.
3. *Games.* A method of gambling in which one doubles the stakes after each loss.
4. A loose half belt or strap placed on the back of a garment, such as a coat or jacket. (*The American Heritage Dictionary of the English Language*, 2000)

However, studying probability theory with this definition in mind is disastrous, as the usual definition of a martingale in mathematics is in fact the following (greatly simplified from Williams, 1991, p. 94).

Definition 1 (Martingale) *A random process $X \equiv (X_n)$ is called a* martingale *if the following holds true.*

1. *The value X_n is known at time n.*
2. *The value X_n of X at time n is expected to be finite.*
3. *Given the values of X up to time n, the next value X_{n+1} of X is expected to be the same as X_n.*

To add to the confusion, different authors may well use the same words to mean different things. Moreover, different authors may use different words to mean the same thing; for instance, there is not difference between the meaning of the phrases "almost surely" and "with probability one". With this in mind, do look up any unfamiliar words in a dictionary (mathematial or otherwise), but also keep an eye on the context in which they are used. Use the index and the table of contents as well as the references to obtain clarity in problematic cases.

To understand the purpose of a specific definition, you should ask yourself the following questions (adapted from Bender, 1996; Reiter, 1998):

1. To which mathematical objects—or group of mathematical objects—does this definition apply? Also, are there any known sets of objects not covered by the definition?
2. How would we verify whether an object satisfies this condition?
3. Are there necessary or sufficient conditions for it, i.e. are there some known sets of objects that satisfy this definition?
4. Why would we be interested in making this definition? What is the important or useful concept behind it?
5. Is there a convenient way of classifying objects satisfying this definition?

It may not be possible to answer all these questions until you've read further, but in considering them you are alerted to the emphasis of future lines of argument followed in the text.

Halmos (1985) and Samii (1998) suggest substituting notation and terms for your own favourites, rewriting the text as you read. This keeps your concentration levels high, helps you to remember the process and will help you to better judge your understanding of the concepts involved. If you do this consistently, it will be easier to blend your own notes of different texts into a coherent whole later.

2.2 Understanding the claims

When presented with a narrative piece of mathematical text—such as the body of a proof—, examination of the argument is necessary for identification of the most important claims. In the labelled style, this task is much easier.

In general parlance, a theorem is a statement of interest that has been (or is being) proven as true. The first step in finding the essence of a theorem is recognising that the purpose of a theorem is to provide necessary and/or sufficient conditions for its conclusion to hold. To understand the content of a theorem, the following questions are useful (Reiter, 1998; Körner, 1998):

1. Which question does this theorem answer?
2. Where and why are each of the premises needed?
3. What is new in this theorem? How does it relate to known theorems? Is it surprising?
4. Are there some cases in which it is trivial? Which simpler results does this theorem generalise?
5. Can it be generalised? If not, what is the obstacle?
6. Are there any direct and indirect applications? Can it be used to prove other facts?
7. Does this theorem raise any new questions? What is the next step in the development of the subject?

2.3 Reconstructing the arguments

The most common way—the way of least resistance—of reading a mathematical argument or proof is to verify that each step follows from the previous ones. This is an essential and illuminating activit; however, this leads to understanding only at a purely technical level. The only way to understand a proof completely is to attempt it yourself (cf. Körner, 1998; Simonson and Gouveau, 2003). In this way, you will discover where each premise is used, why it can or cannot be relaxed, what the underlying strategy of the proof is and whether the proof can be simplified.

The key step in reconstructing a mathematical proof is to recognise the proof method used. What follows, is a summary—neither mutually exclusive nor completely exhaustive—of common methods for showing that a statement A implies another statement B (adapted from Solow, 1990, Table 5).

1. A *forward-backward* proof works forward from A and backward from B; the proof concludes when these two sequences of statements converge.

2. The proof of the *contrapositive* assumes that B does not hold, and deduces from this that A cannot hold either.

3. Proof by *contradiction* assumes that A holds, but not B, and reaches a contradiction, from which it follows that the assumption about B not being true was erroneous.

4. A *constructive* proof is used when B is a statement about the existence of an object with a certain property C. While assuming A, the object is constructed (or guessed), and then C is demonstrated.

5. When A specifies that a statement C holds true for all objects in a certain class, the proof by *specialisation* simply chooses a particular object in that class, assumes that C holds, and demonstrates B.

6. When B specifies that a statement C holds true for all objects in a certain class, the proof by *choice* assumes A, chooses an object in the class and shows that C holds.

7. A statement C that holds for each integer above an initial number, say n_0, can be substantiated by *induction*. Assuming A, it proceeds by proving C for n_0. Invoking the induction hypothesis—that C holds for n—, it then demonstrates that C holds for $n+1$.

8. When B is a statement about the uniqueness of an object with property C, a *direct uniqueness* proof assumes that there are two such objects, that A holds, and concludes that the two objects are equal. An *indirect uniqueness* proof arrives at a contradiction by assuming A, and the existence of two different objects with the property C.

9. If B states that either C or D holds, then proof by *elimination* assumes either that A holds and C does not, and concludes that D holds, or it assumes that A holds, but D not, and deduces C.

10. If A states that either C or D holds, then proof by *cases* assumes, in turn, statements C and D, and deduces B.

Expect to attempt a proof at least three or four times before starting to make headway. When you get (honestly) stuck, refer back to the text. Chances are that you will make some mistakes; at some point, a new fact will come to light. This may be a reference to a lemma; knowing the point of the lemma, state it and attempt to prove it. If it is a reference to a theorem, look it up. If it is an unusual algebraic manipulation, try it yourself. Only by attempting as much of a proof yourself as you can will you identify the key steps which are not automatic.

Much of mathematics is automatic writing. Although this is a necessarily sequential activity, mathematical understanding is not. You will find that when returning to something left half-understood, it becomes clear in the light of further information that has become available. Annotate both your own notes and the text as you go: doing so records your own train of thought, as well as the arguments in the text.

Of couse, proving every single theorem you encounter is time-consuming and often unnecessary. The best compromise between the ideal of complete understanding and the reality of limited energy and time is to adapt the level of involvement of your reading to the difficulty of the material and how central its subject is to your research.

Finally, should it come to the worst, some comfort may be found in the words of Hubbard and Burke Hubbard (2002):

> keep in mind that there are two parts to understanding a theorem: understanding the statement, and understanding the proof. *The first is more important than the second.*

2.4 Ticking the boxes

In this step, determine which of the aims of the text have been attained, and which have not. Where the text failed to fulfill its purpose, decide whether the author is conscious of this fact. In narrative texts, such as books, papers and proofs, this would hinge on whether the claims have been substantiated, and if they have not, whether the reasons for this are explained. In general, a definition fulfills its purpose if it is applicable to a non-degenerate class of objects, and therefore useful. A theorem must be true; it should also be useful.

3 Critical or evaluative reading

In the third reading, you should judge the text, and decide whether and to what extent you agree or disagree with its account of the subject. While doing so, keep the following guidelines in mind (Adler, 1967, p. 125).

1. Do not start evaluating the text before you have finished the first two readings. You cannot judge a text fairly before you understand it completely.
2. Refrain from disagreeing disputatiously or contentiously.
3. Motivate any critical judgment you make carefully. Respect the difference between knowledge and opinion.

Due to the precise nature of mathematics, points of disagreement can often be lifted directly from your own notes on the text. The specific grounds for disagreement are as follows.

1. The text is *uninformed* if it ignores a known fact that, had it been used by the author, would have improved the text significantly.
2. The text is *misinformed* if it uses other material incorrectly or in an inappropriate way, therefore rendering its own claims invalid.
3. The text is *illogical* if its arguments are invalid.

Another reason for criticism, namely incompleteness, is often easy to substantiate, but difficult to correct. The best practice in this case is to simply write down immediately as completely as possible any unusual or original ideas you may have.

This last mantra is perhaps the most important in mathematical research, and is motivated by the history of Fermat's Last Theorem (more accurately, Fermat's Last Conjecture). Sometime during the late 1630s, Pierre de Fermat, a jurist and amateur mathematician, wrote the following in the margin of his copy of Diophantus' *Arithmetic*, next to problem 8 in Book II which asks "given a number which is a square, write it as a sum of two other squares":

> *Cubum autem in duos cubos, aut quadratoquadratum in dus quadratoquadratos, et generaliter nullam in infinitum ultra quadratum potestatem in duos ejusdem nominis fas est dividere: cujus rei demonstrationem mirabilem sane detexi. Hanc marginis exiguitas non caperet.*[2]
> (Wiles, 1995, p. 443)

In mathematical terms, Fermat's Last Theorem can be stated as follows.

Theorem 2 (Fermat, Wiles) *For any integer $n > 2$, there does not exist any integers x, y and z such that the equation*

$$x^n + y^n = z^n$$

holds true.

However, Fermat neglected to write down his "marvelous" proof of this astonishing claim, in this margin or elsewhere. Indeed, this conjecture remained unproven—a controversial challenge for cranks and serious mathematicians alike—for over three centuries. It has only recently been proved by Wiles (1995).

References

Academic Resource Center, Wheeling Jesuit University (2001), How to read a college textbook, Handout, Wheeling, West Virginia. Accessed 5 December 2003.
URL: *http://www4.wju.edu/arc/handouts/read_text.pdf*

Adler, M. J. (1941), 'How to mark a book', *The Saturday Review of Literature, 6 July* pp. 268–272. Accessed 6 December 2003.
URL: *http://www.maebrussell.com/Articles and Notes/How To Mark A Book.html*

Adler, M. J. (1967), *How to read a book: The Art of Getting a Liberal Education*, Simon and Schuster, New York. Accessed 8 December 2003.
URL: *http://www.questia.com/PM.qst?a=o&d=978814 and http://cyberspacei.com/greatbooks/h2/how2read/h2r.htm*

Bender, E. A. (1996), *Mathematical Methods in Artificial Intelligence*, Wiley-IEEE Computer Society, pp. xviii–xix. Accessed 5 December 2003.
URL: *http://math.ucsd.edu/ ebender/proofs.html*

[2] "On the other hand, it is impossible for a cube to be written as a sum of two cubes or a fourth power to be written as a sum of two fourth powers or, in general, for any number which is a power greater than the second to be written as a sum of two like powers. I have a truly marvelous demonstration of this proposition which this margin is too narrow to contain." (Edwards, 1977, p. 2)

Edwards, H. M. (1977), *Fermat's Last Theorem: A Genetic Introduction to Algebraic Number Theory*, number 50 *in* 'Graduate Texts in Mathematics', Springer-Verlag, New York.

Halmos, P. R. (1985), *I want to be a mathematician: An automathography*, Vol. New York, Springer-Verlag.

Hubbard, J. H. and Burke Hubbard, B. (2002), *Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach*, 2nd edn, Prentice Hall, chapter 0.1: Reading Mathematics. Accessed 10 December 2003.
URL: *http://www.math.cornell.edu/~hubbard/readingmath.pdf*

Körner, T. W. (1998), How to write a Part III essay. Accessed 5 December 2003.
URL: *http://www.dpmms.cam.ac.uk/~twk/Essay.pdf*

Reiter, A. (1998), Helping undergraduates learn to read mathematics. Accessed 5 December 2003.
URL: *http://www.maa.org/t_and_l/exchange/ite3/reading_reiter.html*

Samii, B. (1998), Reading mathematics, Leaflet, California State University, Northridge. Accessed 5 December 2003.
URL: *http://www.csun.edu/devmath/resources/readingmath.html*

Simonson, S. and Gouveau, F. (2003), How to read mathematics. Accessed 5 December 2003.
URL: *http://www.stonehill.edu/compsci/History_Math/math-read.htm*

Solow, D. (1990), *How to read and do proofs: An introduction to mathematical thought processes*, 2nd edn, Wiley, New York.

The American Heritage Dictionary of the English Language (2000), 4th edn, Houghton Mifflin.

Wicks, J. R. (2003), How to read a mathematics article, Handout, North Park University, Chicago. Accessed 5 December 2003.
URL: *http://campus.northpark.edu/math/courses/Math_4010/reading.htm*

Wiles, A. (1995), 'Modular elliptic curves and Fermat's last theorem', *The Annals of Mathematics, 2nd Series* **141**(3), pp. 443–551.

Williams, D. (1991), *Probability with martingales*, Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge. 6th print.

Reading a mathematics text is very different from reading ordinary English. Trying to read math the same way as a novel or a history text is certain to cause you trouble. Math text typically alternates passages of explanation in English with pieces of mathematics. Even its English, however, is of a special and stylized kind. When reading any explanatory material in a math text, the main principle is simple: read every word, one word at a time.Â How to work a solved problem in the textbook: Start with a solved problem, working it through to the end, one step at a time. Then close the book, and try to work it through again on your own. Do the whole problem as many times as necessary, until you can reproduce the whole solution with the book closed.