Rekindling critical thinking: heeding major errors in current *Introduction to Proof* type textbooks

raymond.boute@pandora.be

Abstract Mathematics is slowly recovering from unwanted side-effects of the mathematics reforms that took place during the past decades. Quite encouraging is the revived interest in proofs, evident from the growing number of *introduction to proof* type textbooks in basic curricula. Alas, many of these texts contain self-conflicting definitions, which defeats their very purpose. Sloppiness and uncritical “borrowing” of definitions and comments from other books seem to be the main causes.

Various examples are given, also showing that, remarkably, flawless treatments were commonplace 50 years ago. Especially problematic are the most basic concepts of relations and functions. In the simplest definitions from the literature, a relation is a set of pairs, and a function is a relation that is functional, i.e., where no two pairs have the same first member. Correct accounts for the concept of a function from $A$ to $B$, the notation $f : A \rightarrow B$, onto-ness, composition, inverses and so on follow naturally. Still, current textbooks use a convoluted way to define exactly the same concepts via the Cartesian product and, perhaps in the resulting confusion, make additions that cause logical contradictions. Similar issues are discussed for the concept of truth set, a typical topic in *introduction to proof* books.

Well-chosen problem statements can teach students to read math texts more critically and prevent errors from becoming epidemic. Proper use of symbolic notation and reasoning is most clarifying in this respect.

A philosophical conclusion is that definitions (and ensuing theorems) are tools for studying mathematical concepts and hence, like any tools, deserve diligent design. The rift between formalism and Platonism is artificial, arguably even misleading.

1 Introduction: context, motivation and overview

This note was initially submitted for the MathFest TCPS#11: *Cultivating Critical Thinking through Active Learning in Mathematics*, but was reassigned to the general session *Teaching or Learning Advanced Mathematics*. Yet, its topics are quite basic and simple.

It is evident that, if mathematics is to cultivate critical thinking at all, critical thinking must start within mathematics. This necessity has been made even more compelling by a recent phenomenon: widespread blunders in current textbooks of the *Introduction to Proof* type (henceforth abbreviated ITP). We are not talking about minor flaws, but about logical contradictions within basic definitions. Quite remarkable is the contrast with the flawless accounts in all “classic” texts from the 1960s in a random sample from various areas of mathematics (algebra, analysis/calculus, logic, set theory) and still found in most current analysis/calculus texts.

We will start with a brief discussion in Section 2 of some aspects of mathematical practice and exposition that heighten the alertness for pitfalls during further reading. Arguably, if these aspects had been heeded more carefully in the recent past, many common errors would have been prevented.

Section 3 illustrates how the design of the various definitions in the literature around relations and functions requires critical analysis. In particular for the notion of function or mapping this effort is well-spent since, as Herstein [20] aptly observes, “Without exaggeration this is probably the single most important and universal notion that runs through
all of mathematics”. A well-designed definition for this concept is a cornerstone for every introductory mathematics book, especially if it is intended as an introduction to proof.

By comparison, the topic of Section 4 appears more specialized at first. It is motivated by the fact that many introductory books mentioning sets use the concept of truth set as an example, yet a correct treatment was found in only one reference. In a wider context this illustrates the importance of a proper understanding of variables and bindings, the distinction between bindings and statements, and between statements and predicates.

The conclusion (Section 5) briefly reviews the rôle of Section 2 in retrospect.

2 Some simple measures to avoid blunders

2.1 Recognizing the basic nature of mathematics as a method

As opposed to the layperson’s view of mathematics as a “load of facts”, we recall Morris Kline’s more pertinent view that “More than anything else mathematics is a method” [23]. Although Kline does not make this equally explicit, the purpose of the method is effective reasoning in problem solving. Theorem proving can be seen as the particular kind of problem solving where the answer is told in advance.

A direct consequence for mathematical practice and exposition is that the method and the strategy are at least as important as the result. For practice, the open problem approach provides more insight than goal-directed proofs, leaves more opportunity for exploration, and has higher fecundity, that is: a higher yield in additional results. For exposition, especially in ITP-type textbooks, discussing methods and strategies are important in the writeup of a solution or proof.

Still, when addressing deeply ingrained misconceptions, as becomes necessary in this note, formulating certain points as theorems rather than as open questions may be safer, to avoid well-trodden paths to pitfalls. Of course, when presenting the same issues to unprejudiced novices, the open-ended problem approach retains all its educational value.

2.2 Exploiting all mental tools benefiting the method

Evidently, it is wise to exploit every mental tool that enhances the “mathematical method”.

Well-designed and properly used symbolism is arguably one of the most effective tools. It boosts clarity of expression and provides guidance in reasoning, allowing to draw information from the shape of the expressions in addition to just their meaning.

In mature disciplines, such as algebra and calculus, symbolism is well-developed and, as a result, its possibilities are routinely exploited. Notational abuse is fairly moderate.

Practices are quite different in relatively more recent disciplines dealing with logic, set theory, discrete math and so on. Here poor design and sloppy use of notation continues hampering the clarity of symbolic expression and the reliability of symbolic reasoning. As a result, symbolism is underexploited and has acquired a poor reputation.

The ensuing vicious circle is made clear by the apt criticisms from famous mathematicians half a century ago. For instance, Halmos [18] mentions “unacceptable but generally accepted” notational variants. Instead of this euphemism, junk notation might be more fitting. Quine [30] even talks about “glaring perversity”, “willfulness” and “carelessness” for lesser offenses. Since things have not improved, perhaps “sabotage” should be added.

Indeed, unacceptable yet persistent practices constitute the major (arguably even the only real) cause of “symbol phobia” among students and authors.
In this manner, a crucial opportunity is neglected: the ability to reason with predicates and quantifiers as fluently as is taught for polynomials in high school. The benefits for mathematics in general are briefly outlined in the Appendix.

What is important for this note, and will become evident later on, is that all blunders discussed later would have been avoided by a modicum of symbolism. Indeed, even though the word statements in nearly all references are most precise, they still have caused readers to see things that are clearly excluded, or to miss things that are clearly included. Literal translation of precise word statements into symbolic form provides an effective antidote.

2.3 Systematic design of definitions, with explicit justifications

Definitions do not arise “out of thin air” but are the result of design. Their rôle is capturing mathematical concepts in order to facilitate reasoning about them. The adequacy of a definition depends on various design considerations.

Evident basic requirements are clarity, precision and soundness (absence of logical conflicts). Recognizing possible proof obligations generated by a definition is crucial yet, as we shall see, all too often overlooked, with fatal consequences. Also, the way in which a definition captures a concept, its formulation and its structure have a significant impact on the “flavor” of the derived theorems and their proofs. The principle of separation of concerns helps structuring definitions to avoid convoluted and error-prone formulations.

Most blunders discussed in this note could have been avoided by explicitly justifying all design decisions. Not only do explicit justifications make the conceptual issues clearer for the reader, but drafting them encourages more diligent and judicious thinking on the part of the writer. Unfortunately, such good habits are rare in mathematics. One noteworthy example is Naive Set Theory by Halmos [13] — hence required reading for all beginning students and perhaps for a fair number of textbook writers as well!

3 Examples A: relations and functions

3.1 A random sample of sources

In view of comparing basic definitions, most sources are textbooks. In the list below, P indicates texts of the ITP type, and O designates other texts from diverse areas of mathematics: algebra, analysis/calculus, discrete mathematics, logic, set theory etc. We shall see that P-texts typically contain fatal flaws whereas O-texts are mostly sound. Underscores indicate the rare exceptions to this pattern.

O 1 Apostol  P 12 Garnier  O 24 Kolmogorov  – 36 Shuard
O 3 Bartle  P 13 Gerstein  O 25 Krantz  O 37 Smith
P 4 Bloch  O 15 Goodaire  O 26 Larson  O 38 Sprecher
O 5 Bourbaki  P 16 Gries  O 28 Mendelson  O 39 Stewart
P 7 Chartrand  O 18 Halmos  O 29 Meyer  O 40 Suppes
P 8 Daepp I  O 19 Hammack  P 31 Roberts  O 41 Tarski
P 9 Daepp II  O 20 Herstein  O 33 Royden  P 42 Velleman
O 10 Dasgupta  O 21 ISO stdn.  O 34 Rudin  P 43 Wallis
O 11 Flett  O 22 Jech  P 35 Scheinerman  O 44 Zakon

Uncritical copying appears to be very common, so the reader must beware of the viruses!
3.2 Relations: two styles for defining the same concept

a. Soundness through simplicity

Definition 1 (Relation) A relation is a set of (ordered) pairs.

Or, in symbols, \[ \text{isrel} \equiv \forall z : R. \exists x. \exists y. (z = x, y) \] (1)

Clearly, \( R := \{(0, 2), (1, 3)\} \) is a relation, and will serve as a running example.

The following terminology is commonly used for classifying relations.

Definition 2 (Domain, range, relation from \( X \) to \( Y \))

i. The domain \( D_R \) of a relation \( R \) is the set of first members of the pairs in \( R \).

The range \( R_R \) of a relation \( R \) is the set of second members of the pairs in \( R \).

For our example, \( D_R = \{0, 1\} \) and \( R_R = \{2, 3\} \).

ii. A relation from \( X \) to \( Y \) is a relation whose domain is included in \( X \) and whose range is included in \( Y \).

Or, in symbols, \( \text{Xisrel} (X, Y) \equiv \text{Xisrel} \land D_R \subseteq X \land R_R \subseteq Y \) (2)

For instance, \( R \) is a relation from \( \{0, 1\} \) to \( \{2, 3\} \) and also from \( \{0, 1, 7\} \) to \( \{1, 2, 3\} \).

b. Inviting trouble by cramming

Whereas Definitions 1 and 2 keep the concepts relation and from \( X \) to \( Y \) cleanly apart (separation of concerns), this conceptual clarity is lost next.

Definition 3 (Crammed form) A relation from \( X \) to \( Y \) is a subset of \( X \times Y \).

Or, in symbols, \( \text{Xisrel} (X, Y) \equiv R \subseteq X \times Y \) (3)

For instance, \( R \subseteq \{0, 1\} \times \{2, 3\} \) and \( R \subseteq \{0, 1, 7\} \times \{1, 2, 3\} \). Hence, by Definition 3, \( R \) is a relation from \( \{0, 1\} \) to \( \{2, 3\} \) and also from \( \{0, 1, 7\} \) to \( \{1, 2, 3\} \).

The example suggests comparing the right-hand sides of (2) and (3), which yields

Theorem 1 (Equivalence of definitions) Definition 3 is equivalent to Definition 2 ii.

Moreover, Definition 3 seems nicely compact, so what can be the problem?

b. How cramming causes misconceptions leading to unsoundness

A popular criticism considers the “set of pairs” concept hard to grasp for beginners. However, this claim is always stated in a vague and rather speculative way, except by Shuard, who attributes the difficulty to the use of the Cartesian product.

Even so, one would not think that the Cartesian product is as obscure as suggested, were it not that Definition 3 has effectively caused widespread misconceptions, even unsoundness. Extensive correspondence indicates that some mathematicians endow the Cartesian product with illusory properties, which can be summarized as

Myth #1 Definition 3 makes \( X \) and \( Y \) attributes of \( R \). [and, taken further,]
One cannot define just a relation without adding “from \( X \) to \( Y \)”. 

4
This myth is debunked in principle by Theorem 1 and de facto by the numerous issue-free definitions in the literature [5, 22, 35, 40, 41, 44].

The misconceptions caused by Definition 3 are harmful in two different ways. In this analysis, assume $R$ is a relation from $X$ to $Y$ and $S$ is a relation from $U$ to $V$.

1. **Unsoundness** Question: when does the equality $S = R$ hold?
   No one would have failed to answer this question correctly 50 years ago: by the definitions considered, relations are sets, so $R = S$ is set equality and does not depend on $X$, $Y$, $U$, $V$. Most P-texts seem unaware of this, as they redefine $R = S$, which is a conceptual mistake. In some cases [31] this redefinition adds the requirement $X = U$ and $Y = V$. This evidently leads to contradictions, for instance, $R = R \Rightarrow \{0, 1\} = \{0, 1, 7\} \land \{2, 3\} = \{1, 2, 3\}$.

2. **Flawed judgement** Question: when is the composition $S \circ R$ defined?
   Again, this issue was clear 50 years ago: by definition, $z(S \circ R)x \equiv \exists y . z S y \land y R x$, regardless of $X$, $Y$, $U$, $V$. This generality comes entirely free of charge. By contrast, P-texts [31, 37, 42, 43] consider $R \circ S$ only for the special case $Y = U$. This restriction is totally unnecessary, and even unacceptable in practical applications.

Two comments. (Ad 1) For any mathematical object, the view on equality is the litmus test for understanding by an author or by a reader. (Ad 2) The undesirable restriction on composition will reappear for functions. It is absent in all analysis/calculus texts in our sample, regardless of their publication date — obviously in view of practical applicability.

For functions, misconceptions and unsoundness return in a more harmful form.

### 3.3 Functions: two styles for defining the same concept

**a. Soundness through simplicity**

**Definition 4 (Function)** (O [1, 5, 10, 11, 22, 28, 40, 41, 44], P [35])

A function is a set of ordered pairs no two of which have the same first member.

I.e. [5, 40], $f$ isfun $\equiv f$ isrel $\land \forall (x, y, z). (x, y) \in f \land (x, z) \in f \Rightarrow y = z$. (4)

As a tribute to simplicity and clarity, Definition 4 used Apostol’s wording [1].

The quantified expression in (4) states that no two pairs have the same first member. This property is called *functionality* [5, 29]. Hence Definition 4 can be transliterated as “a function is a functional relation”, which is Bourbaki’s wording [5].

Functionality is the justification for the following common convention: for any function $f$ and any $x$ in $D f$, we write $f(x)$ or $f x$ for the (unique) $y$ satisfying $(x, y) \in f$.

From Definition 4 Apostol derives the following practical equality criterion.

**Theorem 2 (Equality)** $f = g \equiv D f = D g \land \forall x : D f . f(x) = g(x)$.

The following terminology is commonly used for classifying functions.

**Definition 5** ($f : X \to Y$) (O [1, 10, 11, 22, 41], P [35], ISO standard [21])

A function from $X$ to $Y$ is a function with domain $X$ and range included in $Y$.

Notation: one writes $f : X \to Y$ to introduce a function $f$ from $X$ to $Y$.

---

1Qualifications of this kind should be assumed as a matter of fact: a statement can be checked meaningfully only against the definition in whose context it appears or against a logically equivalent definition. Hence such qualifications will be taken for granted without repeating them.
b. Inviting trouble by cramming  

Simplicity is not always popular, considering

Definition 6 (Function from $X$ to $Y$)  

i. A function $f$ from $X$ to $Y$ is a subset of $X \times Y$ containing exactly one pair of the form $(x, y)$ for each $x$ in $X$. One introduces $f$ by writing $f : X \rightarrow Y$.

ii. The set $X$ is called the domain of $f$.

A few remarks are in order here.

First, the references listed with Definition 6 contain no unsound additions. However, Definition 6 also appears in [4, 7, 8, 9, 12, 19, 31, 37, 14, 43], as discussed below.

Second, in the phrasing of Definition 6 we have corrected the widespread but unacceptable phrasing that starts with “A function $f$ from $X$ to $Y$, denoted by $f : A \rightarrow B$, is a relation $f \subseteq X \times Y$ such that . . .” or some similar preamble. This phrasing suggests that the function is denoted by $f : A \rightarrow B$ (in fact, it is denoted by just $f$), and that $f \subseteq X \times Y$ is a relation (in fact, it is a statement about the relation $f$).

Most importantly, Definition 6.ii entails a proof obligation: showing that $X$ is indeed determined by $f$. This task is easy, and yields $X = \mathcal{D} f$, in the notation from Definition 2.

All P-texts overlook this obligation as well as the following crucial facts.

Theorem 3 (Equivalence) Definitions 5 and 6 are equivalent.

Corollary (Theorem 2)  

$f = g \iff \mathcal{D} f = \mathcal{D} g \land \forall x : \mathcal{D} f. f(x) = g(x)$.

The misconceptions resulting from these oversights are similar to those discussed for relations, but even more insidious and more deeply ingrained.

c. How cramming causes misconceptions leading to unsoundness

(I) Not surprisingly, Myth #1 has a functional variant, apparently ineradicable.

Myth # 2 Definition 6 makes $X$ and $Y$ attributes of $f$. [and, taken further.] One cannot define just a function without adding “from $X$ to $Y”.

This myth is debunked in principle by Theorem 3 and de facto by the numerous issue-free definitions in the literature [1, 5, 10, 11, 22, 28, 35, 40, 41, 44].

(II) The misconceptions pave the way for a fatally unsound addition: the codomain. All P-texts using Definition 6 except [13], “embellish” it as follows.

Definition 7 (Annex to 6: codomain)  

iii. The set $Y$ is called the codomain of $f$.

The proof obligation (always overlooked) is showing that $Y$ is indeed determined by $f$.

This task is impossible, which reveals unsoundness. Indeed, given $f : X \rightarrow Y$, the codomain of $f$ is $Y$, by Definition 7. Let $Y'$ be any set that properly includes $Y$, that is: $Y \subseteq Y'$ but $Y \neq Y'$. Clearly, $f \subseteq X \times Y \subseteq X \times Y'$. Hence $f$ is a function from $X$ to $Y'$ and, again by Definition 7, the codomain of $f$ is $Y'$. Since $Y \neq Y'$, this is a contradiction.

(III) Other harmful consequences are reminiscent of those for relations. In this analysis, consider $f : X \rightarrow Y$ and $g : U \rightarrow V$.

• More unsoundness  Question: when does the equality $f = g$ hold?

No one would have failed to answer this question correctly 50 years ago: $f = g$ is set equality, which not depend on $X, Y, U, V$, but leads to Theorem 2. Among the P-texts, only [8, 13, 37] realized that Definition 6 implicitly defines function equality. Others redefine equality. This has resulted in various formulations.
Contradicting Definition 7 [8, 13, 31], sometimes intentionally [37].
Contradicting Definition 6 by requiring equal codomains [4, 12, 19].
Restrictively defining $f = g$ only in case $X = U$ and $Y = V$ [7, 9].

• Flawed judgement
  Question: when is the composition $g \circ f$ defined?
  Most analysis/calculus texts [1, 3, 26, 39], regardless of their publication date, exploit “full generality free of charge”, which amounts to
  \[
  D(g \circ f) = \{ x : D f \mid f x \in D g \} \quad \text{and} \quad (g \circ f) x = g(f x). \tag{5}
  \]
  By contrast, P-texts [31, 35, 37, 42, 43] consider $g \circ f$ only for the special case $Y = U$.
  A few [8, 9, 22, 25] relax this restriction to $R f \subseteq D g$.
Recall that restrictions on function composition are unacceptable for applications, and therefore are avoided in analysis/calculus texts.

d. Related definitional design issues and other considerations

(I) Onto-ness is a widely used notion for characterizing functions. Proper nomenclature reflects natural language in using “onto” as a preposition. This is the case for all textbooks using the formulation of Definition 8 (O [11, 22, 28, 41, 44], P [35]) and for many others using the equivalent formulation of Definition 9 (O [3, 18, 20, 24, 28, 41]).

Definition 8 (Onto Y) A function $f$ is onto $Y$, or surjective on $Y$, iff $R f = Y$.
Trouble arises from using “onto” as an adjective (P [4, 9, 19, 31, 32], O [16, 25]).

Definition 9 (“Onto”) A function $f : X \to Y$ is onto or surjective iff $R f = Y$.
Definition 9 entails the same proof obligation as Definition 6ii, exposing the same unsoundness: given a function $f : X \to Y$, we don’t know $Y$. For instance, $\{(0,2), (1,3)\}$ is a function from $\{0,1\}$ to $\{2,3\}$ and from $\{0,1,7\}$ to $\{1,2,3\}$. Is it onto?
Note that mostly the P-texts run into trouble, the only exceptions being [13, 37], although with a flawed account that hampers generality.

(II) Misconceptions and unsoundness are often downplayed by generalities such as

Myth # 3 Different function definitions just reflect different needs in various areas.

In fact, our sample is drawn from very diverse areas of mathematics, and all turn out using the same function concept. The differences reside only in the care devoted to the formulation. The instances of unsoundness found are genuine logical contradictions.

A related popular cover-up argument for unsound definitions is

Myth # 4 Definition 6 is informal and allows multiple views on the role of $Y$.

In fact, all cited sources for Definition 6 clearly write $f \subseteq X \times Y$, using unambiguous notions of subset and Cartesian product. The only exception is [4], where the definition starts with “a function $f$ from $A$ to $B$ is a subset $F \subseteq A \times B$”. The phrasing suggests that $f$ is a subset renamed $F$, which is bound to confuse students. Further clarification in [4] is reminiscent of the following experimental definition from Bourbaki.

(III) Consider the design option of associating a codomain with a function in a sound way. Since this is incompatible with Definition 6, it requires a different function concept.
Definition 10 A function $f$ is a triple $(F, X, Y)$ where $X$ and $Y$ are sets and $F$ is a functional set of pairs s.t. $F \subseteq X \times Y$ and $\mathcal{D}F = X$. (Bourbaki [5, p. 77])

Here, adding “the set $Y$ is called the codomain of $f$” is trivially sound. As an aside, note that $X$ is redundant in the triple since it equals $\mathcal{D}F$.

One should ask whether Definition 10 is a good design idea.

The only published literature evaluating the codomain as a function attribute seems to be a paper by Shuard [36]. This paper is not easy to find, but the reader can obtain a copy from raymond.boute@pandora.be In any case, the sole advantage mentioned by Shuard is making the use of “onto” as an adjective sound. Still, the selectivity in being able to say “$f$ is onto $Y$ but not onto $Z$” (onto as a preposition) seems preferable.

The disadvantages are more significant. Shuard notes that variants with codomains compare unfavorably with the simplicity of Definition 4 as used by Flett [11]. The complexity of associating a codomain has no redeeming value since, by Definition 5, $f : X \to Y$ already specifies $\mathcal{R} f \subseteq Y$, which suffices for all intents and purposes.

The codomain notion even destroys generality of useful algebraic concepts and properties, such as subfunction ($f \subseteq g$), compatibility ($f \odot g$), merge ($f \& g$), composition ($g \circ f$). For instance, $g \circ f$ requires $\mathcal{D}g = \mathcal{C}f$, a unacceptable restriction in practical domains.

Ironically, on the page following his Definition 10 Bourbaki announces using the term “function” for “functional set of pairs” [5, p. 78]. Halmos [17] finds such changes of mind amusing, but one might say, more appreciatively, “Come back, Bourbaki, all is forgiven!”.

The main conclusion is that making the codomain a function attribute is a useless and even harmful design idea.

### 3.4 Points of interest for improving ITP texts

Let’s briefly summarize some salient points regarding the definition of function.

The simplest definition is: “A function is a functional set of pairs” (Definition 4).

By the standards, $f : X \to Y$ specifies $f$ as a function from $X$ to $Y$, which means: a function satisfying $\mathcal{D}f = X$ and $\mathcal{R}f \subseteq Y$ (Definition 5). It is very important to realize that this amply suffices for all theoretical and practical purposes. Not fully appreciating this may well be the cause of all misconceptions. In particular, the idea of needing a different function concept or some “extra’s” (the infamous codomain) indicates an incomplete understanding of the ramifications of the available definitions. A modicum of symbolism is invaluable for clarifying perception.

Cramming “function” and “from $X$ to $Y$” into a single definition is unwise. Even if sound by itself, it obfuscates proof obligations. It also invites unsound “embellishments”, such as the codomain snafu. From an educational perspective, it is clear that the combined formulation has effectively confused many authors. What to say about their students?

Associating a codomain with a function can be made sound, but remains a useless and restrictive design idea.

McCabe has been quoted as saying “Any clod can have the facts, but having opinions is an art.”. Paraphrased into mathematical terms, “Any clod can have soundness, but having well-designed definitions is an art”, this quote is quite relevant to definitional design.
4 Examples B: variables, bindings, truth sets

Truth sets appear as examples in many ITP-type textbooks in a quite misleading way, whereas they are handled properly (with occasional minor lapses) in other texts [32]. The problems can be traced back to sloppy conventions or practices regarding variable bindings and to conceptual confusion similar to failing to distinguish between \( f \) and \( f(x) \) for functions.

4.1 Variables and bindings

a. Common mathematical practice Variables are usually handled according to various conventions that are rarely stated yet tacitly used in the large majority of textbooks. We make these conventions gradually more explicit in an informal way. The technicalities are well-explained in Barendregt’s classic text [2].

Semantically, a variable denotes an arbitrary mathematical object. Further elaboration is not essential. Syntactically, a variable occurs in an expression (in the trivial case, by itself). Such an occurrence can be either \textit{bound} or \textit{free}, as determined by bindings.

\textit{Binding} means introducing a variable in a context, using some appropriate convention. Common informal binding is textual, e.g., “Let \( x \) be a real number” or “For all real \( x \).” Problems are moderate here. By contrast, common practice for binding in symbolic form is quite flawed. Various \textit{ad hoc} notations are in use, often sloppy, ambiguous, and not helpful to understanding. This situation will be improved below by a simple convention.

In an expression, a variable is \textit{bound} in the context of its binding, and \textit{free} otherwise. For instance, in “For all real \( x, x^2 \geq y \), \( x \) is bound and \( y \) is free.

A distinguishing property is that bound variables may be renamed. For instance, “For all real \( z, z^2 \geq y \)” is equivalent to “For all real \( x, x^2 \geq y \).” The meaning of an expression depends on the free variables only. In the example just given, the free variable is \( y \), and the expression is true provided \( y \leq 0 \).

b. Clean conceptualization: distinction between statements and bindings A simple convention suffices to avoid some “unacceptable but generally accepted” notations.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Example</th>
<th>Prose equivalents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statement</td>
<td>( x \in X )</td>
<td>“( x ) is in ( X )” or “( x ) is a member of ( X )” etc.</td>
</tr>
<tr>
<td>Binding</td>
<td>( x : X )</td>
<td>“( x ) in ( X )” or “( x ) ranging over ( X )” etc.</td>
</tr>
</tbody>
</table>

A sanity check for proper usage is the \textit{prose test}: an expression with the correct binding symbols yields grammatically correct prose when read literally. For instance, the RHS of \( S \subseteq T \equiv \forall x : S. x \in T \) is read “for all \( x \) in \( S \), \( x \) is in \( T \)”, which is fine. However, the RHS of \( S \subseteq T \equiv \forall x \in S. x \in T \) is read “for all \( x \) is in \( S \), \( x \) is in \( T \)”. The presence of the statement “\( x \) is in \( S \)” where smooth prose expects a binding is clearly a violation.

In this manner, prose can improve symbolism by exposing junk notation.

Note that the choice of : as a replacement for the offending \( \in \) is unobtrusive, i.e., it will never hamper understanding by people used to sloppy conventions. Its origins are unclear, but the notation \( f : X \to Y \) fits the pattern if \( X \to Y \) is taken as denoting the set of functions from \( X \) to \( Y \) (sometimes written \( Y^X \)).

\footnote{In the common vernacular for bindings, saying that a variable is bound or free actually refers to specific occurrences in an expression. Similarly, the \textit{free variables in an expression} are those having free occurrences in that expression.}
c. Typical reward: taking set builder notation beyond sloppy syncopation

Apart from their role in the symbiosis between prose and symbolism, proper binding conventions offer some more concrete technical rewards.

For instance, nearly all introductory college mathematics books introduce the so-called set builder notation. Unfortunately, the notations are disparate and ad hoc, missing systematic rules. Still, the following common patterns clearly emerge.

Typical examples:

\[
E = \{n \in \mathbb{Z} \mid n \text{ mod } 2 = 0\} \quad E = \{2 \cdot m \mid m \in \mathbb{Z}\}
\]

Matching patterns:

\[
S = \{x \in X \mid p\} \quad T = \{e \mid x \in X\}
\]

(legend)

\text{p is a boolean expression} \quad \text{e is any expression}

The commonly intended reading for \(\{x \in X \mid p\}\) is “the set of all \(x\) in \(X\) satisfying \(p\)”, and for \(\{e \mid x \in X\}\) it is “the set of all \([values of]\ e\ as\ x\ ranges\ over\ X\)”. However, literal reading of \(x \in X\) reveals that both patterns fail the prose test. The offending notation also causes ambiguity, since \(U = \{x \in S \mid x \in T\}\) fits both patterns, as noted in [27].

As a result of such careless design, common usage of set builder notation remains restricted to introducing a set in compact form, as some form of syncopation, but does not extend to symbolic calculation.

Clear binding conventions support systematic symbolic rules.

Corrected patterns:

\[
S = \{x : X \mid p\} \quad T = \{e \mid x : X\}
\]

Symbolic rules:

\[
x \in S \equiv x \in X \land p \quad y \in T \equiv \exists x : X . y = e
\]

As expected, the prose test is successful:

\[
\{x : X \mid p\} \text{ is read as “the set of all } x \text{ in } X \text{ that satisfy } p\}
\]

\[
\{e \mid x : X\} \text{ is read as “the set of all } e \text{ as } x \text{ ranges over } X\}
\]

Moreover, the ambiguity is resolved:

\[
\{x : S \mid x \in T\} = S \cap T \text{ and}
\]

\[
\{x \in S \mid x : T\} \text{ is a set of truth values.}
\]

### 4.2 Predicates and truth sets

a. Clean conceptualization: statements versus predicates

Supposedly “everyone” knows that confusing \(f\) with \(f(x)\) is wrong. Writing \(\{x_i\}\) for a sequence \(x\) (which is a function) is similar, and precisely the notation that Halmos called “unacceptable but generally accepted”. Yet, many texts in signal processing still write \(x[i]\) for the sequence \(x\).

Also similar is confusing a statement (or boolean expression) \(P(x)\) with predicate \(P\). Proper distinction does not depend on whether truth values are accepted as part of the language of mathematics, or exiled to metamathematics. Indeed, it is a syntactic matter of selectivity in substitution, as explained next.

Consider a statement \(p\) that possibly contains some variables. If we write \(p\) as \(P(x)\) with the convention that \(P(e)\) stands for \(p\) with \(e\) substituted for every free occurrence of \(x\), then \(P\) is by definition a predicate. A simple notational convention for introducing a predicate linked to a given statement is writing \(P(x) \equiv p\).

For instance, given the statement \(x^2 > y\), writing \(P(x) \equiv x^2 > y\) introduces a predicate \(P\) such that \(P(z + 7)\) stands for \((z + 7)^2 > y\).
A formal explanation is derived from the syntactic conventions of the lambda calculus [2]. Given a statement \( p \) and a set \( D \), then \( \lambda x : D . p \) is a predicate with domain \( D \). Conversely, given a predicate \( P := \lambda x : D . p \) and \( x \) in \( D \), then \( P(x) \equiv p \).

In fact, “set \( D \)” is common mathematical vernacular for “expression \( D \) denoting a set”. Also, \( D \) may contain free variables, usually excluding \( x \) for simplicity. More importantly, if \( P \) is introduced by writing either “\( P := \lambda x : D . x^2 > y \)” or by writing “Consider a predicate \( P \) with domain \( D \) and \( P(x) \equiv x^2 > y \)”, then \( y \) occurs free in \( P \).

b. Truth sets  
Truth sets are properly defined for predicates, as in

**Definition 11 (Truth set)**  
Given a predicate \( P \) with domain \( D \), we define the truth set of \( P \) to be the set of elements \( x \) in \( D \) for which \( P(x) \) is true [32].

Formally:  
\[
T P = \{ x : D \mid P(x) \}
\]

Sound management of free and bound variables is demonstrated by the fact that the free variables in \( P \) (say, the expression \( \lambda x : D . p \)) and in \( T P \) (correspondingly, \( \{ x : D \mid p \} \)) are the same.

Interestingly, [32] is a discrete math book. By contrast, all ITP texts mentioning truth sets define them for statements by truth set of \( P(x) = \{ x : D \mid P(x) \} \) (or similar). But, in that case, what is truth set of \( x^2 > y \)?

The error is evident: in a proper definition for truth sets, it must be clear which variables are meant as the subjects of a given statement, and which variables are meant as parameters. The required selectivity is achieved syntactically by abstracting the subjects, resulting in a predicate. This is illustrated by the following examples.

<table>
<thead>
<tr>
<th>Predicate</th>
<th>Domain</th>
<th>Definition</th>
<th>Truth set</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P )</td>
<td>( \mathbb{R} )</td>
<td>( P(x) \equiv x^2 &gt; 9 )</td>
<td>( T P = { x : \mathbb{R} \mid x^2 &gt; 9 } )</td>
</tr>
<tr>
<td>( Q )</td>
<td>( \mathbb{R} )</td>
<td>( Q(x) \equiv y^2 &gt; 9 )</td>
<td>( T Q = { x : \mathbb{R} \mid y^2 &gt; 9 } )</td>
</tr>
<tr>
<td>( R )</td>
<td>( \mathbb{R}^2 )</td>
<td>( R(x, y) \equiv x &gt; y )</td>
<td>( T R = { (x, y) : \mathbb{R}^2 \mid x &gt; y } )</td>
</tr>
<tr>
<td>( S )</td>
<td>( \mathbb{R} )</td>
<td>( S(x) \equiv x &gt; y )</td>
<td>( T S = { x : \mathbb{R} \mid x &gt; y } )</td>
</tr>
</tbody>
</table>

It is instructive to analyze these four truth sets to appreciate the selective management of bindings and how it provides the expressivity that is essential for making truth sets more than curiosity restricted to single-variable statements.

5 Some concluding remarks

The mathematics reforms in past decades have played a rather ambivalent rôle, bringing about the current situation. Undoubtedly their most positive contribution was emphasizing the essence of mathematics as a *method*, rather than as a bag of facts. However, this is not really new, since good instructors have always been doing this. On the other hand, the results of the reforms fell far short of the promises. In my own country (Belgium), one observes even a marked decline, apparent in and perpetuated by the poor quality of the textbooks (of course, aggressively denied by the publishers).

Perhaps the most counterproductive element in the reforms is playing down the value of symbolism, even in the face of its spectacular results in analysis/calculus over many centuries. There is a valid excuse for the reticence in other areas of mathematics, especially

---

3The symbol \( := \) is read “defined as”, a global binding convention that satisfies the prose test.
those from which ITP texts draw their illustrative topics, such as set theory and logic. Indeed, symbolism in these domains is counterproductive in its current form. However, the shortcomings are not inherent in symbolism itself, but are due to carelessness in the design of symbolic notation and even more in its use. Clearly students should be encouraged to let symbolism fully assume its role as the most powerful mental tool for reasoning.

Technically, this is not difficult at all since the basis is available. Indeed, there always have been plenty of mathematicians who set the example by getting things “exactly right”. A little selectivity when adopting existing formulation and notation suffices. Yet, as the earlier quotes from Halmos and Quine indicate, willfulness and carelessness are serious impediments. In my own experience, they are more prevalent in mathematics than in any other branch of science. Might that be because mathematics has to work only on paper?

As for now, the main conclusion is that mathematics texts must be read more critically than half a century ago.

Appendix: symbolism as a tool for effective reasoning

Most likely, pictures were the earliest mental tools in “prehistorical” mathematics. As exemplified for algebra in [6], other tools evolved from the rhetorical phase (prose only) via syncopation (symbols as shorthands) to genuine symbolism (reliable notation and rules for symbolic reasoning).

The well-developed state of symbolism and its routine use in mature disciplines is illustrated in Figure 1, the point being that reasoning with symbols is considered commonplace in analysis/calculus and that no one would consider attempting to reproduce such calculations in prose.

From a technical viewpoint regarding the style, Figure 2 shows familiar textbook examples of reasoning by calculation, both in the general mixed style based on chaining equalities and inequalities, and in the purely equational style, chaining equalities only.

Logical results can be discovered and logical arguments can be presented using the same calculational style. The rules are even much simpler than for summation and integration since convergence, integrability and other complicated issues do not arise.

For instance, the statement of the theorem in Figure 3 appears in a typical ITP textbook [42]. Typically its proof is in prose, and the strategy consists in building a bridge between the given hypothesis and the given goal. The resulting argument is declared in [42] to be “the most complex proof we’ve done so far”.

Yet, using very simple calculation with quantifiers, the theorem can be derived symbolically by first expanding the formula expressing the hypothesis using the definitions for its various parts, and then reducing the formula using applicable rules, evidently without going backwards. Thus, the desired result is discovered and the converse is obtained as a free bonus.

Of course, for truly complex theorems, there is not just one path automatically leading to a neat result, and an equational chain (≡ only) may have to be weakened by occasional implication (⇒, which is ≤ for truth values), making the converse a separate issue.

Even so, for a proof in prose, the converse (if any) is always a separate issue, because natural language lacks the linguistic constructs for dealing with logical equality (equivalence) in addition to implication. For the same reason, proving set equality $S = T$ in prose always requires proving $S \subseteq T$ and $T \subseteq S$ separately, which is therefore often presented
\[ I'(k, m) := \int_0^1 B_2^*(t) \cos(m \pi t) \, dt \]
\[ = \int_0^1 (B_2(t) - B_2) \cos(m \pi t) \, dt \]
is equal to \( I(k, m) \), because \( \int_0^1 \cos(m \pi t) \, dt = 0 \) for \( m > 0 \). For fixed \( k \geq 1 \), summing (5) over \( m \) yields
\[ \frac{(-1)^{k+1}(2k)!}{2^{2k} \pi^{2k}} \zeta(2k) = \frac{(-1)^{k+1}(2k)!}{\pi^{2k}} \sum_{m=1}^{\infty} \frac{1}{(2m)^{2k}} \]
\[ = \sum_{m=1}^{\infty} I'(k, 2m) = \sum_{m=1}^{\infty} I'(k, m). \]

The telescoping trick. We will need the elementary trigonometric identity
\[ \cos(mx) = \frac{\sin \left( \frac{2m+1}{2} x \right)}{2 \sin \left( \frac{x}{2} \right)} - \frac{\sin \left( \frac{2m-1}{2} x \right)}{2 \sin \left( \frac{x}{2} \right)}. \] (6)

With the introduction of (6), we now have a telescoping series, yielding
\[ \frac{(-1)^{k+1}(2k)!}{2^{2k} \pi^{2k}} \zeta(2k) \]
\[ = \sum_{m=1}^{\infty} \int_0^1 B_2^*(t) \cos(m \pi t) \, dt \]
\[ = \lim_{N \to \infty} \sum_{n=1}^{N} \left( \int_0^1 B_2^*(t) \cos\left( \frac{2m+1}{2} \pi t \right) \, dt - \int_0^1 B_2^*(t) \cos\left( \frac{2m-1}{2} \pi t \right) \, dt \right) \]
\[ = \lim_{N \to \infty} \sum_{n=1}^{N} \left( \int_0^1 B_2^*(t) \frac{\sin \left( \frac{2m+1}{2} \pi t \right)}{2 \sin \left( \frac{x}{2} \right)} \, dt - \int_0^1 B_2^*(t) \frac{\sin \left( \frac{2m-1}{2} \pi t \right)}{2 \sin \left( \frac{x}{2} \right)} \, dt \right) \]
\[ = \lim_{N \to \infty} \sum_{n=1}^{N} \left( \int_0^1 B_2^*(t) \frac{\sin \left( \frac{2N+1}{2} \pi t \right)}{2 \sin \left( \frac{x}{2} \right)} \, dt - \int_0^1 B_2^*(t) \frac{\sin \left( \frac{2N-1}{2} \pi t \right)}{2 \sin \left( \frac{x}{2} \right)} \, dt \right) \]
\[ = \lim_{N \to \infty} \int_0^1 B_2^*(t) \frac{\sin \left( \frac{2N+1}{2} \pi t \right)}{2 \sin \left( \frac{x}{2} \right)} \, dt - \int_0^1 B_2^*(t) \frac{\sin \left( \frac{2N-1}{2} \pi t \right)}{2 \sin \left( \frac{x}{2} \right)} \, dt. \]

We observe that by (3), the value of the second term is
\[ \frac{1}{2} \int_0^1 B_2^*(t) \, dt = \frac{1}{2} \int_0^1 (B_2(t) - B_2) \, dt = \frac{B_3}{2} \]
Now we show that the limit in the first term is 0. Note that the function
\[ f(t) = \frac{B_2^*(t)}{2 \sin \left( \frac{x}{2} \right)} \]
for \( t \in (0, 1] \),

Ramanujan, in which the Bernoulli numbers appear. For positive \( \alpha, \beta \) with \( \alpha\beta = \pi^2 \) and \( k \) any nonzero integer, we have
\[ a^{-k} \left( \frac{1}{2} \zeta(2k + 1) + \sum_{m=1}^{\infty} \frac{n^{-2k+1}}{m^{2k-1}} \right) \]
\[ = (-\beta^{-k}) \left( \frac{1}{2} \zeta(2k + 1) + \sum_{m=1}^{\infty} \frac{n^{-2k+1}}{m^{2k-1}} \right) \]
\[ - 2k \sum_{a=0}^{k+1} (-1)^a \frac{B_{2a}}{(2a)!} (2k + 2 - 2a)! \zeta(a+1+\beta k), \]
see [5, Entry 23 (i) on page 275]; in the recent papers [9, 10], this formula has been analyzed from the standpoint of transcendence. The book [13, Section 4.2] contains a large collection of other formulas for \( \zeta(2k + 1) \).
Figure 2: Familiar styles of reasoning by calculation: mixed (left) and equational (right)

$$\frac{1}{n} \sum_{x = 1}^{n} p^n(x|\theta) q^n(x)$$

$$\leq \frac{1}{n} \sum_{x = 1}^{n} p^n(x|\theta) [1 - \log q^n(x)]$$

$$= \frac{1}{n} \left( L(p^n; q^n) + H_n(\theta) \right)$$

$$= \frac{1}{n} + \frac{1}{n} d(p^n, q^n) + H_n(\theta)$$

$$\leq \frac{2}{n} H_n(\theta)$$

$$F(s) = \int_{-\infty}^{+\infty} e^{-|x|} e^{-2\pi sx} dx$$

$$= 2 \int_{0}^{+\infty} e^{-x} \cos 2\pi sx \, dx$$

$$= 2 \frac{e^{-x}}{2\pi s} [1 - \pi s]$$

$$= 2 \frac{1}{4\pi^2 s^2 + 1}$$

Figure 3: Deriving a theorem using reasoning by calculation

**Theorem** Let $B$ be a set and $\mathcal{F}$ a set of sets. If $\bigcup \mathcal{F} \subseteq B$ then $\mathcal{F} \subseteq \mathcal{P} B$.

**Proof**

$\bigcup \mathcal{F} \subseteq B \equiv$ (Definition $\subseteq$) $\forall x . x \in \bigcup \mathcal{F} \Rightarrow x \in B$

$\equiv$ (Definition $\bigcup$) $\forall x . (\exists A \in \mathcal{F} . x \in A) \Rightarrow x \in B$

$\equiv$ (Rdistr. $\Rightarrow / \exists$) $\forall x . \forall A \in \mathcal{F} . x \in A \Rightarrow x \in B$

$\equiv$ (Swapping $\forall$) $\forall A \in \mathcal{F} . \forall x . x \in A \Rightarrow x \in B$

$\equiv$ (Definition $\subseteq$) $\forall A \in \mathcal{F} . A \subseteq B$

$\equiv$ (Definition $\mathcal{P}$) $\forall A \in \mathcal{F} . A \in \mathcal{P} B$

$\equiv$ (Definition $\subseteq$) $\mathcal{F} \subseteq \mathcal{P} B$

as the sole proof technique for set equality.

In any case, there is much to be gained by letting logic, set theory and related disciplines benefit from the mental tool of symbolism and helping students to become as fluent in quantifier calculus as in polynomial algebra.

This explains why Velleman emphasises that “using the notation and rules of logic can be very helpful when you are figuring out the strategy for a proof” [12]. Yet, he also mentions that [many?] mathematicians write their proofs in ordinary English as much as possible. Clearly, such self-imposed restrictions do not provide a constructive example.

In brief, there is not a single good reason for letting only algebra and calculus/analysis enjoy the advantages of symbolic reasoning — as they are doing de facto for centuries — and keep excluding all the rest of mathematics.

**References**


[20] Israel Herstein, Topics in Algebra. Xerox College Publishing (1964)


[40] Patrick Suppes, Axiomatic Set Theory. Dover (1972)


