Some ergonomics of mathematical notation

Roland C. Backhouse

30 September, 1991

Put your hand in your pocket and pull out a number of coins, preferably all of the same denomination. Show them to a friend and ask how many there are. If there are less than five your friend will be able to see instantly how many there are; if there are more than five he will be obliged to count them before giving a reliable response.

I have no doubt that there are many learned articles dealing with this and similar experiments in a proper, scientific way. For me, however, the experiment has a very simple and far-reaching — albeit subjective — significance. The experiment demonstrates to me just how *un*intelligent I and my fellow human beings are. We may try to convince ourselves of our supreme intelligence but the fact remains that we are quite incapable of assimilating or exploiting all but small amounts of information at any one time.

In spite of our inherent stupidity the human race has achieved a very great deal (achievement being quite different from intelligence). Recognition of one's limitations is the first step towards improving one's achievements.

The evolution of mathematical notation has been of fundamental importance to the development of science. Because it is both concise and precise, mathematical notation helps to simplify concepts to a level at which we can begin to understand them and to overcome our tendency to woolly and disorderly thinking. However, whether it is used well or badly can make all the difference between whether mathematical notation makes molehills out of mountains or mountains out of molehills. The ergonomics of mathematical notation is a little-discussed but vital aspect of creative mathematics.

A non-mathematical example may be the best introduction to the sort of points I want to make. If, when doing a crossword puzzle, I suspect that one of the answers is an anagram of some phrase then I write the letters of the

phrase in a circle. This notational trick is enormously helpful in enabling the eye to see different permutations of letters. Compactness of the notation is highly significant: a computer-generated listing of all permutations of a given set of letters may be a more reliable way of discovering all the anagrams but is decidedly less effective. Computer-generated lists are just not for human consumption!

Recognising human characteristics is important to the design of good notation. One of the first rules that one should learn about mathematical notation is that the precedence chosen for a binary operator should determine the size of the symbol used to denote that operator, the higher the precedence the smaller the symbol. This is because small symbols "pull" their neighbours together thus suggesting a grouping of the symbols. For example, in the expression

$$a + b.c + d$$

one naturally sees the sequence $b.c$ as a group because the variables b and c are close together.

Note that the size of a symbol should also include the amount of white space around it. Text produced by a typewriter illustrates this well. The expression

```
a+b.c+d
```

has been printed in teletype mode, i.e. in such a way that each symbol has exactly the same width. The intention may be that the dot has higher precedence than plus but one must work very hard in order to read the expression in that way.

The principle underlying the precedence rule is that mathematical notation should suggest relevant groupings of symbols, or at least not be biased to specific groupings. For example, if $\oplus$ is an associative operator then one should denote its application using infix notation; for then in an expression like

$$a \oplus b \oplus c \oplus d$$

one can choose at will whether to continue the calculation by manipulating $a \oplus b$, $b \oplus c$, or $c \oplus d$. In contrast, if Polish notation is used the expression above could be written in five different ways

$$\oplus(\oplus(\oplus(a, b), c), d)$$

$$\oplus(\oplus(a,b),\oplus(c,d))$$
$$\oplus(\oplus(a,\oplus(b,c)),d)$$
$$\oplus(a,\oplus(\oplus(b,c),d))$$
$$\oplus(a,\oplus(b,\oplus(c,d)))$$

each of which is biased to particular groupings of the arguments.

The advantages of infix notation for associative operators are not so striking because they are very familiar. A less familiar example is provided by so-called "abide" laws. Two binary operators $\otimes$ and $\oslash$ are said to abide with each other if for all u,v,w and x

$$(u \otimes v) \oslash (w \otimes x) = (u \oslash w) \otimes (v \oslash x)$$

Written as above the law seems hideously complex; a two-dimensional notation reveals the true nature of such laws. The name "abide" signifies that the operators can be written *ab*ove or bes*ide* each other as shown below

$$\begin{array}{ccc} u & \otimes & v \\ & \oslash & \\ w & \otimes & x \end{array} \quad = \quad \begin{array}{ccc} u & & v \\ \oslash & \otimes & \oslash \\ w & & x \end{array}$$

A standard example of an abide law is provided by multiplication and division in real arithmetic. (Replace "$\otimes$" by "$\times$" and "$\oslash$" by "/".) The validity of this law is the only justification I know for why the operands in a division are written one on top of the other. Take, for example,

$$\frac{u \cdot v}{w \cdot x}$$

Because the arguments are pulled together the eye is more readily encouraged to spot different groupings of the operands — $u \cdot v$, $\frac{u}{w}$, $w \cdot x$, $\frac{v}{x}$ and, since multiplication is commutative, $\frac{u}{x}$ or $\frac{v}{w}$.

Aside Abide laws abound in mathematics, sometimes being called interchange laws. However, they don't seem to be well known. One example occurs in boolean algebra: Suppose $p \ldots u$ are booleans and define

$$p\langle q\rangle r \quad \equiv \quad \texttt{if } q \texttt{ then } p \texttt{ else } r.$$

Then

$$\begin{array}{lll} p & \langle q\rangle & r \\ & \langle s\rangle & \\ t & \langle q\rangle & u \end{array} \quad = \quad \begin{array}{lll} p & & r \\ \langle s\rangle & \langle q\rangle & \langle s\rangle \\ t & & u \end{array}$$

End of Aside

(Readers of The Squigolist will know that the term "abide law" was coined by Richard Bird and that the above example of such a law is due to Tony Hoare.)

Subscripts and superscripts are probably the most abused elements of mathematical notation. Because they are smaller than the symbols around them they are easily overlooked. Just like the small print in legal documents this can be deliberately used to deceive the reader, or it can be used to suppress details that are only relevant in exceptional circumstances. Very occasionally deception can be beneficial! Suppose a given function distributes over a given binary operator. Denoting the function by C, the operator by $\times$ and function application by an infix dot, distributivity can be expressed syntactically by

$$C.(X \times Y) = C.X \times C.Y$$

for all X and Y. An alternative denotation is obtained by choosing c to denote the function and using superscripting to denote function application. We then obtain

$$(X \times Y)^c = X^c \times Y^c$$

for all X and Y. (To emphasise my point about the size of superscripts I have used capital letters for the dummies.) What is the essential difference between the two notations? Well, compare $C.X \times C.Y$ with $X^c \times Y^c$. In the former "X" and "Y" are relatively far apart, in the latter "X" and "Y" have been pulled together by the relative size of the dummies and the superscript. In the latter, therefore, the intention is that the eye is tricked into overlooking the superscript and grouping together X and Y. The notation avoids the need to consciously remember the distributivity law.

Now you know why the "Eindhoven School" insists on beautiful handwriting: clear handwriting, paying attention to the ergonomics of mathematical notation, pays dividends whereas bad handwriting can often deceive you into making mistakes. And those computer algebra systems that are currently all the rage? How anyone can begin to do creative mathematics with an input-output system that is hardly better than that of a teletype is beyond me!