What we seem to have learned.

The purpose of this note is to record our main experiences during the first three months we addressed ourselves to the question "how to tame the complexity of artefacts such as proofs and programs". At present we can only give an unsorted list: it is too early to assess their significance.

We have done most of our experiments with mathematical arguments; initially the presentation of the arguments was our main concern, but quickly we started to care about the arguments proper as well.

We started with the prejudice that brevity is a hallmark of mathematical elegance; our recent experiences have only strengthened it. In the beginning we committed in the name of brevity what we then identified as a sin, viz. brevity by omission. (It is an easy trap to fall into. Once alerted, we spotted many "sinful omissions" in the texts of others and even in our own first drafts.)

It took us some time to discover that giving the heuristics or buffering the shock of invention and giving the "clearest" presentation are not necessarily the same thing: the type of "clarity" we seem to be homing in on does not exclude surprises for the reader. (For the first author the ruthless separation of these two concerns came as something of a shock.)

It might be illuminating to point out that the type of "clarity" we are looking for is unthinkable (or unbelievable) for someone that identifies convenient with conventional. Our whole search for "clarity" is based on the conviction that some patterns of reasoning are objectively "simpler" than others in the same sense as in which doing arithmetic in Arabic numerals is objectively "simpler" than in Roman numerals.

To give a specific example, we know full well that the natural numbers are much, much older than the negative integers. We are inclined to regret that that aspect of history has been faithfully copied in the education of all of us. (If we are very puristic, we should even object to the name "natural numbers" since it suggests that the negative numbers are in some sense "unnatural".) It obscures the fact that the unrestricted integers have been invented to <u>simplify</u> matters. It seems that in order to become mathematically adult we have to learn to dissociate ourselves from our pasts: we have to rid ourselves of all sorts of common but confusing thinking habits. But we have to identify them first!

We knew — see EWD447 "On the rôle of scientific thought", August 1974 — that "separation of concerns" is essential. It now seems that such separation can and should be applied to a much greater extent than we anticipated three months ago.

Zero and One.

We knew that the mathematical community still hesitates whether zero is a natural number. The second author discovered this when he observed how the majority of the mathematicians he had asked proved

$$s(n+1) = s(n) + 1 - (p-1) \cdot m(n+1)$$

where $s(n)$ = the sum of the digits of the p-ary representation of n, and
$m(n)$ = the multiplicity of the factor p of n, i.e. the maximum value m such that p^m divides n.

The theorem follows from subtracting 1 from $n+1$, carrying out this simple computation in base p. The second author's observation was that most mathematicians introduced a case analysis depending on whether $n+1$ was divisible by p or not, unaware of the fact that the argument about the length of the train of trailing zeros of the p-ary representation of $n+1$ is also applicable when that train is empty.

We encountered several new instances. After having defined $P(n)$ as the number of unordered partitions of n, one author separately defined $P(0) = 1$. In another paper he derived the sequence $c_1, c_2, c_3, \ldots$ from the sequence $b_1, b_2, b_3, \ldots$ according to the formula

$$c_j = b_j - (j-1)\cdot m \quad .$$

Had he denoted the sequences by $b_0, b_1, b_2, \ldots$ and $c_0, c_1, c_2, \ldots$, the formula would have been

$$c_j = b_j - j\cdot m \quad .$$

Another mathematician – who knew perfectly well that the empty product should equal 1 – had severe problems in accepting that $(\underline{A}i: 0 \le i < n: B(i))$ should equal <u>true</u> for $n = 0$. Also the interpretation of the command "direct arrows towards your tabled neighbours" caused problems in the absence of tabled neighbours.

In classical Greece, one was not considered a number. As a result poor Euclid had to give two separate proofs for his algorithm for the greatest common divisor of two numbers, one for the case they had a common divisor and one for the case they had not. By now people should know better, but the other week we saw a (bright!) mathematician confused because he thought that the largest odd divisor of 15 was 5. Evidently, one is still a second-class citizen.

It can be argued that our Western languages with their singular and plural forms keep these distinctions alive. The circumstance offers a further advantage of formal over verbal arguments.

Postponing definitions and proofs.

This section is mainly about the arrangement of the material to be presented. The message is that most of our experiences with the presentation of programs seems to carry over to the presentation of mathematical arguments in general.

Programs are now usually presented in the so-called "top-to-bottom" or "top-down" fashion. There are good reasons for this. Programs have to bridge the gap between the problems and the tools; whereas the programming languages that provide their building blocks have been very stable — some would even argue: depressingly so — the general-purpose nature of the tool is faithfully reflected in the diversity of the problems to be solved. Top-down presentation presents first what is specific and unfamiliar, thereby providing the framework in which the later details fit. (This also explains why hardware designers, who over the last decades had the diversity at the side of their building blocks, tend to prefer the so-called "bottom-up" presentation.)

Since Euclid's Elements were written, mathematicians tend to present mathematical theories in a bottom-up fashion: axioms first! But presenting a solution based on known foundations is more akin to presenting a program and, accordingly, we found top-down presentations more appropriate. (This could be an example

of how "convenient" and "conventional" can diverge.) For instance, when demonstrating that two sets have the same cardinality it seems worth a sentence at the beginning, announcing whether a counting argument or a one-to-one correspondence will be developed.

A minor consequence of this principle is that we don't lump all definitions together at the beginning of the exposition, but postpone each definition until it is needed.

A major consequence is our tendency to use lemmata before we have given their proofs. Though unusual, this seems entirely correct. The statement of a lemma is a logical firewall between its usage and its proof; the use of a lemma is independent of how the lemma can be proved and, during study of its use, knowledge of its proof is therefore an unnecessary burden.

How naming can hurt.

One of the problems of presenting, annotating, or specifying a sizeable program is avoiding such naming conventions that the number of names needed explodes: not introducing avoidable names becomes vital. As an exercise we therefore tried to present a number of mathematical arguments with great precision and detail but using as few names as possible. We learned two things.

Firstly we discovered a few notational habits that create the need for new names. The "... convention" is one of them, e.g. Courant and Robbins assume

"$m = p_1 p_2 \cdots p_r = q_1 q_2 \cdots q_s$,

where the p's and the q's are primes." In the continuation of their text "p_r" and "q_s" both occur half a dozen times, though the actual values of r and s are totally irrelevant in the argument. The introduction of pictures in geometry – see AvG1/EWD780 – forces the introduction of nomenclature, viz. in order to refer in the text to components of the picture.

Secondly we learned that the introduction of names can complicate the argument – see EWD763/764 and EWD771/772, AvG0b – viz. when the need for case analyses is introduced because the naming has destroyed the symmetry. Both examples are very impressive but a caveat is in order: a 2-coloured grid of points and a 2-coloured complete graph are rather similar objects and the experience might not be transferable outside combinatorics.

On overspecification.

With the introduction of

"$m = p_1 p_2 \cdots p_r = q_1 q_2 \cdots q_s$" ,

Courant and Robbins have done more harm than just

introducing two superfluous subscripts: they have represented a bag of primes by a sequence of primes, though everyone knows that multiplication of integers is "commutative". (Multiplication is "commutative" means that you can "interchange" the factors without affecting the product: $a \cdot b = b \cdot a$. But "interchange" is only meaningful with respect to a notation that has destroyed the symmetry. The symmetry is the fundamental property; commutativity is a notational artefact.) The clumsiness culminates in Courant and Robbins's formulation of the conclusion:

> "Hence the prime decomposition of m' must be [read: "is", AvG/EWD] unique, aside from the order of the factors."

We had similar experiences in AvG4/EWD785 and in EWD787, which both dealt with so-called "unordered partitions", i.e. bags of positive integers with given sum. They were nicely formulated in terms of bags in contrast to the proofs of Post and Sylvester respectively; the partitions being "unordered", Post and Sylvester immediately took the liberty of ordering the parts — say in descending order — and carrying out the argument in terms of that order.

For Post, this had the consequence that he also had to introduce the "reverse" of a partition, i.e. the same partition written in ascending order.

For Sylvester it had the more dramatic consequence that he missed the generalization from 2 to k:

> "Consider for any natural number N the bags of positive integers whose sum is N. The number of those bags not containing a multiple of k ($k \geq 1$) equals the number of those bags containing no integer k or more times."

Using the term "unordered partitions" has two drawbacks. Firstly, starting from sequences in which the order does matter — $\langle 3,2 \rangle \neq \langle 2,3 \rangle$ — the concept is weakened, though adjectives and prefixes should strengthen a concept (instead of calling bags "multisets", we should call sets "unibags"). Secondly, the term "partition" hides the general concept of a bag: it groups the bags of positive integers quite arbitrarily by their sum.

21 May 1981

drs. A.J.M. van Gasteren
BP Venture Research Fellow
Dept. of Mathematics
University of Technology
5600 MB EINDHOVEN
The Netherlands

prof. dr. Edsger W. Dijkstra
Burroughs Research Fellow
Plataanstraat 5
5671 AL NUENEN
The Netherlands