Int. J. Metrol. Qual. Eng. **4**, 41–45 (2013)
© EDP Sciences 2013
DOI: 10.1051/ijmqe/2012036

Rounding, stipulation and notation issues in measurement

F. Pavese*

Torino, Italy

Received: 31 October 2012 / Accepted: 22 November 2012

Abstract. In the context of measurement and of the definition of measurement units, a problem well known in computing science, the inherent propagation and accumulation of rounding errors throughout the intermediate steps of numerical calculation, is discussed in this paper, with some issues in notation, namely of integer numbers.

Keywords: Notation; rounding; stipulation

1 Introduction

This paper intends to tackle, in the context of measurement, a problem well known in computing science [1], the inherent propagation and accumulation of rounding errors throughout the intermediate steps of numerical calculation, and some issues in notation, namely of integer numbers.

In this context, the use of the so-called "stipulated values", or "defined values", is intrinsic in definitions, namely those aiming to establish regulatory conditions of all kinds. Contrary to "consensus values", which are measured values with an associated uncertainty, stipulated values are rounded numbers – either real or integer – deemed exact by definition and have zero uncertainty. The propagation effect of rounding or truncation will occur when more than one stipulated value is combined in an algebraic expression. This may happen in measurement, e.g., when computing the values of multidimensional quantities and having to use more than one unit containing in its definition a stipulated value.

The issue deserves general attention of the experimentalist and of the metrologist, and, in particular, it places intriguing questions concerning the current debate on a more extensive use of stipulated values of "fundamental constants" in the definition of measurement units of the International System of Units (SI) [2–4], a field where missing a single digit of defined values can make the difference in the accuracy between using them and making them useless. The origin of the exact stipulated values is the measurement of those constants at their best accuracy at the moment of stipulation.

There will clearly be some degree of rounding error involved in such a procedure involving what are essentially truncated values, and some subsequent propagation problem.

* Correspondence: frpavese@gmail.com

2 Rounding and truncating

Let us start form the simplest example. Assume to have two rational numbers: $A = 5.6$ and $B = 4.6$. If rounded to integer numbers, they become $A_\mathrm{r} = 6$ and $B_\mathrm{r} = 5$, if truncated $A_\mathrm{t} = 5$ and $B_\mathrm{t} = 4$. The result of their sum is $R_\mathrm{S} = A + B = 10.2$ exactly, $R_\mathrm{Sr} = A_\mathrm{r} + B_\mathrm{r} = 11$, $R_\mathrm{St} = A_\mathrm{t} + B_\mathrm{t} = 9$. The result of their difference is $R_\mathrm{D} = A - B = 2.0$ exactly, $R_\mathrm{Dr} = A_\mathrm{r} - B_\mathrm{r} = 1$, $R_\mathrm{Dt} = A_\mathrm{t} - B_\mathrm{t} = 1$. The result of their product is $R_\mathrm{P} = AB = 25.76$ exactly, $R_\mathrm{Pr} = A_\mathrm{r}B_\mathrm{r} = 30$, $R_\mathrm{Pt} = A_\mathrm{t}B_\mathrm{t} = 20$. The result of their ratio is $R_\mathrm{R} = A/B = 1.2173\ldots$ (rational or real number), $R_\mathrm{Rr} = A_\mathrm{r}/B_\mathrm{r} = 1.2$, $R_\mathrm{Rt} = A_\mathrm{t}/B_\mathrm{t} = 1.25$.

Large errors may obviously occur and be propagated and expanded in the communication of results in rounded and truncated forms. If a long calculation can safely be rounded off to N decimals, it is not valid to round off intermediate steps to the same number of digits because round-off errors accumulate. A larger number of digits (say M) is required at intermediate steps and the difference $M - N$ are called the "guard digits".

In measurement, a first additional problem arises from the fact that an algebraic combination of stipulated values is said to be a stipulated value, requiring to also be exact by definition.

However, after stipulation, one might no longer take into account the fact that these numbers were originally in actuality estimates of real numbers, and affected by an experimental uncertainty. Therefore, one might not compute R from the originally imprecise numbers, and afterwards stipulate its value, either as R_r or R_t, in order to compensate for the rounding error. Nor could one take into account anymore the effects of the original uncertainty. It has been abolished by definition, so that, in general, "guard digits" are not admitted in stipulation.

Let us take an example in measurement, concerning the molar gas constant $R = k_\mathrm{B}N_\mathrm{A}$, where k_B is the Boltzmann constant, $k_\mathrm{B} = 1.3806488(13) \times 10^{-23}$ J K^{-1}

Article published by EDP Sciences

https://doi.org/10.1051/ijmqe/2012036 Published online by Cambridge University Press

(CODATA 2010 [5])[1], and N_A the Avogadro number, $N_A = 6.02214129(27) \times 10^{23}$ mol^{-1} (CODATA 2010 [5], see later Section 4 for a distinct problem for N_A). Should they be stipulated (exact) numbers, for the definition of the measurement units kelvin and mole, respectively, but R not be stipulated, the results of the product of the two rational numbers would be a rational number with a larger number of decimal digits (or even a real number in other circumstances).

R has also been measured directly: its CODATA 2010 value is 8.3144621(75) J mol^{-1} K^{-1}, to be compared with the results of the above product: 8.31446214546895 J mol^{-1} K^{-1} exactly.

However, to which digit should be truncated the latter, certainly having more digits than the significant ones? To the CODATA digits corresponding to the uncertain R? It does not seem correct.

The above latter value of R is obviously consistent with the former, because all digits reported for $k_B N_A$, have been used. However, being the uncertanties of k_B and N_A reported with two digits, the second one is obviously a "guard digit" that should not be used in stipulation. See Section 3.1 for a consequence of this fact.

Two more problems arise in measurement. First, let us modify the initial example by adding a digit to the rational numbers: $A = 5.66$ and $B = 4.66$. If rounded in the usual way one obtains $A_r = 5.7$ and $B_r = 4.7$, now also rational numbers; if truncated, they become $A_t = 5.6$ and $B_t = 4.6$. The result of their ratio is now $R_R = A/B = 1.21459\ldots$, $R_{Rr} = A_r/B_r = 1.21276\ldots$, $R_{Rt} = A_t/B_t = 1.21739\ldots$: in general, they all are real numbers now. Thus, one might not expect that the result of a ratio operation is still a rounded number with a manageable number of digits, but this is in fact not generally true. A common case is when $R = 1/A$.

Secondly, one is not always dealing originally with real numbers. In the case of an integer number (typically, the result of a counting), is rounding (stipulation) admitted, being rounding a concept usually linked to real numbers? A corollary of this problem is: which is the correct notation for an integer value of a discrete quantity of which not all digits (either some of the most significant or some of the least significant) are known? See Sections 3.2 and 4: this problem among others was initially discussed in [4].

3 An application to measurement: stipulation of measurement units

The consequences of the previous considerations can be applied to the case of an extensive use of stipulated values in the definition of SI units, as is currently being proposed. They are significant also in the context of the documented conflict between the SI and the requirements of many data systems and informatics particularly evident in sensor and instrumentation technologies [6].

3.1 More than one value stipulated

If the value of more than one "fundamental constant" is stipulated, should the values of other constants that are algebraic expressions of them be computed as a combination of the stipulated values, or of the original values?

For example, the Stefan-Boltzmann constant $\sigma = 2\pi^5 k_B^4/15h^3c_0^2$ is given the value $5.670373(21) \times 10^{-8}$ W m^{-2} K^{-4} [5]. This value is computed from the values [5] of the three constants $k_B = 1.3806488 \times 10^{-23}$ J K^{-1}, $h = 6.62606957(29) \times 10^{-34}$ J s and $c_0 = 299\ 792\ 458$ m s^{-1} (the latter already a stipulated value), using all the reported digits, including the uncertain ones – $\sigma = 5.670372623 \times 10^{-8}$ W m^{-2} K^{-4} before rounding. Using instead the stipulated values for all three constants, rounded by excluding both the uncertain digits ($k_B = 1.380\ 65 \times 10^{-23}$ J K^{-1}, $h = 6.626069 \times 10^{-34}$ J s), one obtains $\sigma = 5.67039380 \times 10^{-8}$ W m^{-2} K^{-4}. An identical result is obtained in this example by rounding to the first uncertain digit ($k_B = 1.380649 \times 10^{-23}$ J K^{-1}, $h = 6.6260696 \times 10^{-34}$ J s). An obvious rounding error occurs.

Similarly, in the case of R in Section 2 the following stipulated (rounded) values should be used limited to the first uncertain digit: $k_B = 1.380649 \times 10^{-23}$ J K^{-1} and $N_A = 6.0221413 \times 10^{23}$ mol^{-1}. Consequently, $R = 8.31446336370370$ J mol^{-1} K^{-1}, not compatible with the CODATA value for R.

When using values already having been stipulated, no uncertainty can be associated to the value of σ, a real number, nor to R, a rational number: in fact, in [2] the fundamental constants obtained from algebraic operations using stipulated constants are said to have zero associated uncertainty. However, the questions already placed in Section 2, still arise. In addition, the use for the stipulation of all uncertain digits, typically two, looks inconsistent with the very concept of stipulation: the less significant digit is generally allowed in the notation of uncertainty only to act as a "guard digit", while it would be meaningless to upgrade its meaning to a meaningful digit of an experimental value and in stipulation to an exact digit.

Thus, are the derived constants to also be considered as stipulated – i.e. exact? To which digit stipulation of σ in the above example should stop? To the same used for expressing the uncertain constant – $\sigma = 5.67039410^{-8}$ W m^{-2} K^{-4} – or to only the digits exempt from rounding error – $\sigma = 5.670 \times 4 \times 10^{-8}$ W m^{-2} K^{-4}? Could "guard digits" be admitted in stipulation?

[1] The CODATA values are used here. However, note that the CODATA values have been elaborated using a "Least Squares Adjustment" procedure that *alters* the values of the constants, in the meantime that obtains the best *consistency* and lower uncertainties of those values for all constants considered. They are *not* the simple mean of the *measured* values, and the obtained uncertainty is in general *better* than can be obtained experimentally, and *should not be confused* with the latter.

https://doi.org/10.1051/ijmqe/2012036 Published online by Cambridge University Press

3.2 Inverse of a stipulated value

Are inverse numbers to be directly considered stipulated values too – with no associated uncertainty? How can their rounding correctly be performed?

Often the definitions of units use the inverse of a stipulated value – e.g., metre $(1/c_0)$, second $(1/\Delta(^{133}\mathrm{Cs})_{\mathrm{hfs}})$, mole $(1/N_{\mathrm{A}})$. However, the inverse of an integer number is normally a real number, so that any stipulation applied to a number, resulting in a rounded number, does not directly apply to its inverse.

Let us take for example $1/\alpha$ for which also a separate CODATA value exists. Let us stipulate $\alpha = 7.2973525698(24) \times 10^{-3}$ as $\alpha^* = 0.0072973525698$, thus $(1/\alpha)^* = 137.035999074306\ldots$ (limited to 15 digits using MS Excel): the number is a real one, with no obvious rounding or truncation. For CODATA 2010, the adjusted value is $1/\alpha = 137.035999074(44)$, but the stipulation to $(1/\alpha = 137.035999074$ can only be a *distinct* stipulation. See also Section 2.

3.3 Propagation of stipulations

If one uses a stipulated value of a multidimensional constant in the definition of a specific unit, should the other corresponding units also become part of that definition?

For example, consider a present definition: "The metre is the length of path travelled by light in vacuum during a time interval of 1/299 792 458 of a second", obviously making use of a defined value of c_0, whose dimensions are [m s^{-1}]. Considering that the present definition of the second makes use of a stipulated number of "periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium 133 atom", $\Delta(^{133}\mathrm{Cs})_{\mathrm{hfs}} = 9\ 192\ 631\ 770\ \mathrm{s}^{-1}$, one should, in principle, complete the metre definition by adding: "…, for the magnitude of the second being set 'by fixing the numerical value of the ground state hyperfine splitting frequency of the caesium 133 atom, at rest and at a temperature of 0 K, to be equal to exactly 9 192 631 770 when it is expressed in the SI unit s^{-1}' " – where the internal hyphened definition is that proposed one by the BIPM Consultative Committee of Units (CCU) [3]. Obviously, should the stipulated value in the definition of the second be changed, also the stipulated value of c_0 would change.

Incidentally, the "metrological compatibility", according to its definition in the International Vocabulary of Metrology (VIM [7]), can be ensured only within a stated uncertainty. Therefore, the values of c_0 and $\Delta(^{133}\mathrm{Cs})_{\mathrm{hfs}}$ can be compatible with each other only within the uncertainty associated to them before stipulation, $c_0 = 299\ 792\ 458.0(1.3)$ m s^{-1} [8] and $\Delta(^{133}\mathrm{Cs})_{\mathrm{hfs}} = 9\ 192\ 631\ 770\ (20)$ Hz [9]. This means that not necessarily the value 299 792 458 m s^{-1} corresponds exactly to the value 9 192 631 770 Hz, since these are only the first moments (expected values) of two probability distributions, whose second moments (standard uncertainties) are reported in parentheses: not necessarily the stipulated values using the expected values are consistent with each other to their last digit.

3.4 Are algebraic expressions of stipulated values in turn stipulated values?

In reference [10] (table 2), the algebraic expressions of stipulated values, for example $R = k_{\mathrm{B}}N_{\mathrm{A}}$ or $F = N_{\mathrm{A}}e$ or $K_{\mathrm{J}} = 2e/h$ or $R_{\mathrm{K}} = h/e^2$, are given for granted to become in turn stipulated values, i.e. also exact. This opinion is not a direct consequence of the issues illustrated in the previous sections.

In fact, the stipulation of c_0, and the proposed one for h, e, k_{B} and N_{A} is only a consequence of the fact that an uncertain value cannot be used in the definition of a unit (like now, e.g., for c_0, or for the triple point of water, 273.16 K exactly, in the present definition of the kelvin).

Therefore, the exactness only applies for the purpose of the definition of those units. It does not suppress the uncertainty of the values of the constants for other purposes, like namely are the calculation of the value of another constant depending on one or more of the stipulated constants. Each of those constants should be specifically stipulated too, if needed; otherwise, they would retain the original uncertainty resulting from the uncertainties associated with the original experimental values of the involved constants, irrespective to the fact that their values have been stipulated in the definition of a measurement unit.

In other words, the stipulation of the value of a constant has not the general purpose of suppressing the uncertainty of the value existing before stipulation. Misunderstanding this issue might give rise to the very dangerous misunderstanding, that the value is actually exact.

In this respect, when a constant is an algebraic expression of other constants, the stipulation sequence is not irrelevant. For example, for $k_{\mathrm{B}} = R/N_{\mathrm{A}}$, starting from the experimental values, $k_{\mathrm{B}} = 6.02214082(18) \times 10^{23}/8.314463(30)$, the current proposal on the floor [3] requires stipulating k_{B} and N_{A}, not R.

Should the stipulation of k_{B} be done at the same time of that of N_{A}, i.e. starting for both from the uncertain values, one gets $k_{\mathrm{B}} = 1.380649(30) \times 10^{-23}$ J K^{-1} – the uncertainty is arising from R. By rounding the first uncertain digit, the stipulated number becomes $k_{\mathrm{B}} = 1.38065 \times 10^{-23}$ J K^{-1}. This coincides with the stipulated value arising directly from the experimental values: $k_{\mathrm{B}} = 1.38065 \times 10^{-23}$ J K^{-1}. For shake of comparison, with the use of the CODATA 2010 adjusted value, $1.3806488(13) \times 10^{-23}$ J K^{-1}, one would get $k_{\mathrm{B}} = 1.380649 \times 10^{-23}$ J K^{-1}: note that the last digit is not experimentally justified.

Should instead the stipulation of k_{B} be performed after N_{A} having been stipulated, the stipulation of N_{A} would be 8.3145 mol^{-1} (the use of uncertain digits is not justified), whence $k_{\mathrm{B}} = 1.380655(30) \times 10^{-23}$ J K^{-1}. By rounding again at the first uncertain digit, the stipulated number becomes 1.38066×10^{-23} J K^{-1}, different from the previous stipulated value.

https://doi.org/10.1051/ijmqe/2012036 Published online by Cambridge University Press

In general, stipulation *does not have a transitive property*. Each of those constants should subsequently be specifically stipulated too, if needed; a specific, not arbitrary, reason should apply. Otherwise, they would retain the original uncertainty resulting from the uncertainties associated with the original experimental values of the involved constants, irrespective to the fact that their values might have been stipulated in the definition of a measurement unit.

4 Notation problems in rounded integer numbers

This section expands the contents of [4]. An example of intrinsic integer number in measurement is the result of a count. An example of a count is the Avogadro number, whose known value is usually expressed as

$$6.02214129 \times 10^{23}. \quad (1)$$

It is not evident from this expression that the number is an integer number. Is it correct? I do not think so.

The count is an integer with 24 digits, as shown below

□□□□□□□□□□□□□□□□□□□□□□□□
6 0 2 2 1 4 1 2 9

where this expression pictures the fact that only the values of the upper digits of the count are assigned (not typical for a count).

The same expression would have a different meaning for a real number. The known value of the gas constant is

■□□□□□□□□□•••
8. 3 1 4 4 6 2 1

First of all, notice the existence of the decimal point: the infinite digits on the right of it are decreasingly significant by power of 10 with respect to the value 8. On the contrary, in a count, the finite number of digits on the right of the first increases the count indicated by the first digit on the left (like in the integer part of a real number $>10n$, when $n > 1$).

Secondly, the number of digits on the right of the decimal point is infinite. This is the normal situation when truncation or rounding is justified.

In the case of real numbers, the representation is correct and not deceiving, but meaning, for a truncated number, $6.022141290 \times 10^{23}$, where $\underline{0}$ means here a periodic zero.

On the contrary it is deceiving for an integer number when only the upper digits are known: the remaining digits are not zeros, but unknown values. One cannot truncate an integer number.

How to express the result of a count correctly, namely if one wants to provide a value exact be definition when having less than 24 values available for the 24 digits? The situation is the following:

Expressed in:	Counts:	
10^{23}	6	exactly
10^{22}	60	“
10^{21}	602	“
	...	
10^{17}	6 022 141	“
10^{16}	60 221 412	“ (stipulated)*
10^{15}	602 214 129	“ (stipulated)*
10^{14}	6 022 141 290	incorrect
10^{13}	60 221 412 900	incorrect
	...	

* If the last or the last two digits, respectively, are originally affected by an uncertainty.

In fact, one can express correctly only sub-multiples. This corresponds to performing the count in “packets” (low-resolution count): in the case above in packets of 10^{17} counts (for an exact value). In the case of measurements, one should use the corresponding prefix of the unit for the relevant multiple. However, large numbers are defined only in 10^3 steps (e.g., prefix “mega”), nor the use more than one prefix is allowed.

In general, the problem arises that a specific notation does not exist, as far as I know, for expressing an integer number whose upper digits only are known.

For this purpose we need a specific new format. If it does not exist yet, I propose the following, using symbols existing in currently existing fonts. For a generic integer I of N total digits, with only the upper M digits known:

$$I_M| \rightarrow N \quad (2)$$

so that expression (1) should become better written:

$$602\,214\,179| \rightarrow 24,$$

or, for the value of a measurement affected by uncertainty:

$$602\,214\,179(30)| \rightarrow 24.$$

In the case the value of I is stipulated, a double vertical bar || should replace the single bar |:

$$602\,214\,179(30)|| \rightarrow 24.$$

The normal arithmetic rules apply. In case the number is the value of a quantity, so it is followed by the used unit, N only will change as in the notation of real numbers, according to the used multiple of the unit.

So, how to express an integer number with less digits than are known? It may happen for several reasons, the simplest being when one is making an approximate statement. In these cases, I propose to use expression (2) with truncated (or rounded) I_{M} and using the mandatory indication “approximately” or the corresponding symbol (usually ≈, as recommended in [11]) in front of I_{M}:

$$\approx 602\,214| \rightarrow 24.$$

https://doi.org/10.1051/ijmqe/2012036 Published online by Cambridge University Press

5 Conclusions

In conclusion, apparently lexical-only or notation-only issues may bring to basic conceptual and practical dilemmas in many circumstances, particularly important in a regulatory field like that of the definitions of the measurement units.

References

1. IEEE 754-2008: Standard for Floating-Point Arithmetic, http://ieeexplore.ieee.org/servlet/opac?punumber=4610933
2. I.M. Mills, P.J. Mohr, T.J. Quinn, B.N. Taylor, E.R. Williams, Adapting the International System of Units to the twenty-first century, Phil. Trans. R. Soc. A **369**, 3907–3924 (2011)
3. CGPM, On the possible future revision of the International System of Units, the SI, in *Comptes Rendus des Séances de la XXIV Conférence Générale, Resolution 1* (Bureau International des Poids et Mesures, Sèvres, 2011), http://www.bipm.org/en/convention/cgpm/resolutions.html
4. F. Pavese, Some reflections on the proposed redefinition of the unit for the amount of substance and of other SI units, Accred. Qual. Assur. **16**, 161–165 (2011)
5. CODATA 2010 values available at http://physics.nist.gov/cuu/Constants/index.html
6. M. Foster, The next 50 years of the SI: a review of the opportunities for the e-Science age, Metrologia **47**, R41–R51 (2010)
7. BIPM, International Vocabulary of Metrology – Basic Metrology – Basic and General Concepts and Associated Terms (VIM), 3rd edn. (BIPM/ISO, Sèvres, 2008). Available at http://www.bipm.org/en/publications/guides/vim
8. BIPM, Comptes Rendus de la 15^{e} CGPM, Metrologia **11**, 179–180 (1975), http://www.bipm.org/en/CGPM/db/15/2/
9. W. Markowitz, R. Glenn Hall, L. Essen, J.V.L. Parry, Frequency of cesium in terms of ephemeris time, Phys. Rev. Lett. **1**, 105–107 (1958)
10. I.M. Mills, Closing comments, Chem. Int. **32**, 10–11 (2010)
11. Commission IUPAC I-1 (E. Richard Cohen, Tomislav Cvitas, Jeremy G. Frey, Bertil Holmström, Kozo Kuchitsum Roberto Marquardt, Ian Mills, Franco Pavese, Martin Quack, Jürgen Stohner, Herbert L. Strauss, Michio Takami, Anders J. Thor), *Quantities, Units and Symbols in Physical Chemistry*, 3th edn. (Monograph, RSC, London, 2009)

https://doi.org/10.1051/ijmqe/2012036 Published online by Cambridge University Press