Centre for Evidence Based Medicine, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK

Twitter @JKAronson

Cite this as: *BMJ* 2023;380:p529
http://dx.doi.org/10.1136/bmj.p529
Published: 3 March 2023

When I use a word . . . . Data—certainly plural, rarely singular

The UK newspaper *The Financial Times* recently instructed its contributors that its style guide has been changed. The word "data" is henceforth to be considered singular. This diktat mandates "data is," not "data are," "this data," not "these data," and so on. They can't have been studying the evidence—the etymology of the word, grammatical considerations, and above all usage. Etymologically, the word "data" is the plural form of the Latin word "datum," something that is given or is due to be given—a present, a debit, or a debt. And that should be the end of it. "Data" is plural. Grammatical analysis supports this usage. However, the etymological fallacy tells us not to assume that the etymology of a word rigidly determines what it means or how it is to be used in English. In computer science "data" implies information, in a technical sense, carried by an electrical current, and "data" can therefore be regarded as singular. However, in common parlance it refers to pieces of information, in a non-technical sense, such as personal characteristics, and is therefore plural. Like a subatomic particle, functioning as either a particle or a wave, "data" can function as either singular or plural, depending on the language game in which it is being used. Since most uses, such as matters discussed in newspapers, involve common parlance, "data" should continue to be regarded as plural, except on the rare occasions when computer science is the specific subject matter.

Jeffrey K Aronson

Data—singular or plural?

The UK newspaper *The Financial Times* recently instructed its contributors that its style guide has been changed. The word "data" is henceforth to be considered singular. One should therefore write "data is," not "data are," "this data," not "these data," and so on. Before this, and for a few years, the use of the singular or plural was optional. They can't have been studying the evidence—the etymology of the word, grammatical considerations, and above all usage.

The roots of giving

When investigating the origins of words one is often led back to what one might call the mother tongues, a range of proto-languages, such as Proto-IndoEuropean and Proto-Semitic. These languages are themselves not known from records but have been deduced from the patterns of living languages, tracing words back to their hypothesised originals. English words can most often be traced back to Proto-IndoEuropean roots.

Take, for example, the versatile IndoEuropean root DHE, the so-called e-grade form of the root, meaning to set or put down, to make or shape. Because vowels change readily when words develop, the e-grade form can become an o-grade form, DHO, or a zero-grade form, DHə, in which the final vowel is replaced by a neutral vowel sound called a schwa, after the Hebrew vowel of that name. The schwa, represented by an inverted e (ə or ə) typically occurs in weakly stressed syllables, like the final a in "data" (/ˈdɑːtə/). These various forms can also have prefixes and suffixes and may be doubled (technically known as reduplication), giving rise to a myriad of words from a single root. DHE, for example, gives deed and misdeed, DHO gives do, doing, and done, DHEM gives deem and theme, and DHOM gives doom and words ending -dom, like kingdom and leechdom. There are many more.

Now take the IndoEuropean o-grade root DO and its zero-grade form Də, which means to give. In Sanskrit this gave rise to dadāmi and in Greek δίδωμι, both meaning I give. Note the reduplication of the root in both cases. This typically happens in verbs when the action is repeated—giving is supposed to be habitual. The Latin verb to give is dārē (in which both vowels are pronounced separately, as marked), which also reduplicates in the perfect tense as dedi, mimicking the repetition of a past action.

The past participle of the Latin verb dare is datum, meaning "given," which then becomes a noun of neuter gender, meaning something that is given or is due to be given—a present, a debit, or a debt. The plural of "datum" is "data." And that should be the end of it. "Data" is plural.

The etymological fallacy

However, one should not be seduced by the etymology of a word (the etymological fallacy)—what it once meant does not necessarily tell you what it means now. English is not Latin and words mature with time. Consider, as an example, "agenda." The Latin verb agere means to do, and its gerund, agendum, means something that needs to be or must be done. So agenda, the plural form, means things that need to be done. When the word was first used in English it implied a list of things to be done, a list of agenda, but with time the list of plural things just became a list called the agenda. Nowadays it can also mean a plan of some kind (as in a hidden agenda). Similarly, "stamina," now a singular noun, arose from the plural of "stamen," the thread of life spun by the Fates; the longer the thread the more stamina you had.

These two examples, apparently comparable, are not. That is because one is a count noun and the other a non-count noun:

“agenda” is a singular count noun (plural “agendas”);

“stamina” is a singular non-count noun (no plural).

“Data” is clearly not a singular count noun, as agenda is; not even the staunchest proponents of its singularity suggest that there is a word “datas.” But is it the plural form of the singular count noun “datum” or a singular non-count noun, like stamina?

Grammar

The grammatical problem in considering whether “data” is singular or plural arises from the fact that the singular form, datum, is generally used only in a technical sense to mean a baseline, benchmark, or reference point (as in datum level, datum line, datum mark, datum point). Although it can be used to mean a single piece of information, such usage is rare. On the other hand, “data” is used to mean either a whole lot of pieces of information (technically the plural of a count noun—one datum, many data) or a collection of such pieces (technically a non-count or mass noun—much data).

This is similar to the use of collective nouns, such as “board,” “cabinet,” or “government,” which are singular when they refer to a group but plural when they refer to the individual members of the group. Thus, when the late Queen referred to “My government” she used the singular. Here is an example from her speech to parliament in March 2018: “My government is committed to peace in Northern Ireland.” However, the plural would be appropriate in a sentence such as “The government are at loggerheads over the question of Brexit.” If we regard “data” as a word of this type, it should be plural when we have in mind some or all of the individual pieces of data (e.g. “some/all of the data suggest …”) and singular when referring to the agglomeration (“en masse, the data suggests …”). Even in the latter case, however, the plural use would not be amiss, and just as appropriate.

You can test whether you want to use the plural or singular by qualifying “data” with words such as “all” (“all the data are”), “many” (there are many data), and “much” (“much data supports”). Doing that will help you to decide whether you are thinking of the individual pieces of information or the whole collection or a discrete part of it. However, if you want to think of it as a conglomeration of pieces of information, it would be better to use a singular term such as “database,” avoiding the possibility of ambiguity, not to mention calumny.

Computing science

As I have previously pointed out, the members of a particular scientific constituency may define a word differently from the way in which it is defined by another constituency, scientific or otherwise; in that case the different constituencies are playing different language games.[1] An example that I have discussed before is “information,” which sometimes means something different to computer scientists than it does in colloquial parlance.[2]

On one occasion I was verbally accosted (the word is not too strong) by a professor of computing science who demanded to know whether in my view the word “data” was singular or plural. When I suggested the latter he asserted otherwise so forcefully that contradiction seemed unwise.

To oversimplify, to the ordinary user, “data” refers to a collection of facts or pieces of information (in the colloquial sense), while to the computer scientist it represents a single bundle of stuff, electronically represented as information, in the computer science sense.

This is illustrated in a short passage from Claude Shannon’s essay *The Mathematical Theory of Communication*, in a section titled “Equivocation and channel capacity”: “We consider a communication system and an observer (or auxiliary device) who can see both what is sent and what is recovered (with errors due to noise). This observer notes the errors in the recovered message and transmits data to the receiving point over a ‘correction channel’ to enable the receiver to correct the errors. … If the correction channel has a capacity equal to Hy(x) it is possible to so encode the correction data as to send it over this channel and correct all but an arbitrarily small fraction ε of the errors.” [3] An accompanying figure showed what Shannon called “correction data” being transmitted as an electrical signal, making it clear, even if he had not said “to send it” that he regarded data in this sense as a single, or singular, entity. Although, admittedly, if that were so, one would have expected him to have written “a data,” as he would undoubtedly have written “a signal.” It would have been much better had he chosen words other than “information” and “data” to describe his analyses in the first place. But specialists often endow ordinary words with extraordinary meanings.[4]

My assertive interlocutor might have pointed all this out, invoking his own language game, but instead he used the argument that “data” is obviously, as he put it, singular in compound nouns such as “database” and “databank.” That argument, however, was flawed. There is a technical term for nouns that are formed by joining two nouns together; it is “tatpurusha.” The word is Sanskrit and literally means “his servant,” referring to the fact that the meaning of one part is subservient to the meaning of the other. A boathouse is a building in which boats are kept and a houseboat is a boat that functions as a dwelling; the subservience varies with the order of the words.

In tatpurusha the first element can be singular or plural, and whichever it is the whole can refer to one or more objects. A boathouse can contain one boat or more than one; a clotheshorse is a frame on which one or several pieces of clothing can be hung. Some words are singular but have plural forms, such as trousers, but the quasi-singular form is often used in tatpurusha, as in trouser-press and trouser suit. When the first element has the same forms in both singular and plural you can’t tell whether it’s singular or plural, but it generally refers to a plurality of objects. For example, you don’t expect to see just one sheep in a sheep-pen or one deer in a deer park, although you might. So you can’t, strictly speaking, tell whether the occurrence of “data” in “database” or “databank” is singular or plural. However, each is clearly a collection, a plurality, of individual pieces of data—they wouldn’t be databases or databanks if they contained only one piece of information.

Synthesis

My conclusion from all of this is that “data,” which is etymologically plural, can be regarded as either singular or plural, depending on the language game in which it is used. If it is used to refer to an electrical signal in computer science it could be regarded as singular, although the absence of a plural form, “datas,” and of collocation with the indefinite article, i.e. “a data,” argues against that. However, in common parlance it is better regarded as plural. After all if you are concerned that by posting information on Facebook, or some other form of social media, you are giving others access to your personal data, those data will include many different types of information (in the ordinary sense): name, age, height and weight, sex, gender, and sexual preferences, and much else besides.

One further question remains, however. A major determinant of what a word means or how it is defined in a language game depends

on how people use it in that game. This is inherent in Ludwig Wittgenstein's comment that "For a *large* class of cases—though not for all—in which we employ the word 'meaning' it can be defined thus: the meaning of a word is its use in the language."[5] Or, as more simply expressed by Lewis Carroll's Humpty Dumpty, "When I use a word, 'it means just what I use it to mean—neither more nor less."[1]

How people generally *do* use the word "data"—as singular or plural—is for exploration in another "When I Use a Word" article.

Competing interests: None declared

Provenance and peer review: Not commissioned; not externally peer reviewed.

1 Aronson JK. When I use a word. . . . Language games. *BMJ* 2023;380:. doi: 10.1136/bmj.p403. pmid: 36807029

2 Aronson JK. When I use a word. . . . Information. *BMJ* 2023;380:. doi: 10.1136/bmj.p465. pmid: 36828550

3 Shannon CE, Weaver W. *The Mathematical Theory of Communication.* University of Illinois Press, 1949: -9.

4 Aronson J. When I use a word ... Ordinary words with extraordinary meanings. *QJM* 2008;101:-4. doi: 10.1093/qjmed/hcn127. pmid: 18820314

5 Wittgenstein L. *Aphorism 43. Philosophical Investigations. The German Text with a Revised Translation by Gertrude Elizabeth Margaret Anscombe.* 3rd ed. Blackwell Publications, 2001: .