“Data are” or “data is”? A pedant writes

Kieron O’Hara

Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, United Kingdom, kmoh@soton.ac.uk

October 2020, https://www.webscience.org/2020/10/24/data-are-or-data-is-a-pedant-writes/

It is one of the divisive questions of our times. Is the word ‘data’ singular or plural? Some say “this data is …”, “the data doesn’t tell us …”, “we have less data than …”; others “these data are …”, “the data don’t tell us …”, “we have fewer data than …”.

The singular use, often heard in computer science departments and probably more commonly in popular speech, treats ‘data’ as an uncountable noun or singular mass noun, like ‘water’ or ‘education’. It therefore has no plural – we don’t speak of ‘the datas’, any more than we speak of ‘the educations’ or ‘the waters’ (actually, ‘the waters’ is a usable term, but specifically refers to a source of spring water: ‘I came to Casablanca for the waters’). Such nouns only take plurals when combined with a specific unit of measurement.

The plural use, more often heard in social science and statistics departments, says that ‘data’ is the plural of ‘datum’. A datum is something like ‘x = 40’, and if we create a file containing the datum ‘x = 40’ and the datum ‘y = 50’, we have data.

There are two common views of the rights and wrongs of this. One is that either is OK. As long as use is consistent and you make yourself understood, it doesn’t really matter. The other is that the plural use is correct, because ‘data’ is a Latin word, the plural of ‘datum’, which means ‘the given’. You show your ignorance of the classical heritage of English if you misuse the word.

I used to hold the first of these views. Clearly, the roof won’t fall in if we preserve both uses, because they are both conventional, more or less. It’s bad practice to have both uses in the same piece of work, so having chosen one convention, stick with it for good style, but don’t angst about it. Personally, I tended to use ‘data’ as a singular mass noun, but I didn’t hold it against others who didn’t.

However, having been ticked off often enough by holders of the second of these views, I reflected upon it, and I now think the weight of argument is in favour of a third view: that ‘data’ *is* a singular mass noun, and that the plural use, even if verified by Cicero himself, is simply incorrect.

We can take into account four considerations. None of them in itself is decisive, but cumulatively I believe that at a minimum they put the burden of proof on the pluralists. If you are either a pluralist or an agnostic, you need counterarguments.

1. You are not speaking Latin.

‘Data’ is a word of English. It happens to have a homograph in Latin, because we borrowed the word. There are many other English words with this property, from ‘abdomen’ to ‘vomit’, and we don’t worry about their Latin grammar.

When we use words from other languages in English sentences, we often write them in italics, and then we do worry about their grammar. After all, if you are showing off, you had better show off correctly. For instance, *hoi polloi* is a Greek phrase meaning ‘the people’, or ‘the many’, and figuratively ‘the rabble’. ‘*Hoi*’ is a definite article, so it is incorrect to say ‘the *hoi polloi*’, because that means ‘the the rabble.’ There are lots of other phrases we borrow, and we have to get them right –

for instance, the correct plural of the French phrase '*fait accompli*' is '*faits accomplis*', even if we use it in an English sentence.

'Data' is not like this. It is a word of English that should behave like an English word. That does not tell us, of course, whether it is a singular mass noun or a plural, but it does tell us that the Latin rules are irrelevant.

2. You are being inconsistent

There are other singular words in English which began as Latin plurals. 'Agenda' is one of them – it is the Latin plural of 'agendum', that which is to be done. Yet no-one in their right mind uses 'agenda' in English as a plural. No-one says 'The agenda are on a slide, and I'm projecting them onto the screen.' Absolutely everyone, even a data-pluralist, says 'The agenda is on a slide, and I'm projecting it onto the screen.' But if it is acceptable to treat 'data' as an English plural, why not 'agenda'?

3. What do the French do?

Ah, the pluralist might say, others treat their words for 'data' as plurals. The French say 'les données', while the Dutch 'de gegevens'. We should take a lesson from them.

Ah, I reply, but they have not absorbed the Latin word. Rather, they have both translated it as 'the given'. Which is fine – we certainly do not want to dictate to the French and Dutch how they should communicate. But note that when we *translate* Latin 'data' into English, we get a singular term, 'the given'. They get a plural term, and hence their words for 'data' are plural. Had we followed their strategy, we would have got an unambiguously singular term, and we would talk about 'the given', 'givenbases', 'big given', 'metagiven' and so on. No-one would ever say 'the givens.'

4. What do the English do?

This last point trades on something about the way that English treats abstractions such as 'the given' or 'the data'. The singular tends to be used. And we can see this if we compare 'data' with words of similar function in English, such as 'information', 'knowledge' and 'wisdom'. These are also singular mass nouns – if we add more knowledge to our (singular) knowledge, we end up with (singular) knowledge. This kind of abstraction over semantic/epistemological concepts requires, in English, if not in Latin, French or Dutch, a singular grammar.

Not only should we not be surprised to find that 'data' acts in the same way, it will actually lead to greater conceptual clarity – in English – if we reject the pluralist view and the agnostic view, and accept the singular view.

It should be said that the word 'datum' is also useful – we do sometimes want to refer to a single item, and 'datum' will do for that, as well as alternatives like 'piece of data' and 'datapoint'. In the same way, we talk of a 'piece of information', an 'item of knowledge', a 'nugget of wisdom'. I don't rule out the use of 'datum', only the mistaken view that in English it is the singular of 'data'.

One of the penalties of being a pedant is that, despite your being right, no-one cares, and I don't suppose you do. But anyway, it is off my chest now, and I can go on with the rest of my life.