

# Data is typically not a plural

Ben Klemens

10 June 2009

When we learned all those darn grammatical exceptions, we were usually told that they came about in some distant past, due to some arcane relic of old Dutch or something. But here in the new millennium, we have the chance to witness the development of a new grammatical exception.

If this sounds boring, bear with me: by the end of the column, about 360,000 people will die over this corner of grammar.

See, English has the concept of a collective singular, wherein a group of elements is treated as a unit: e.g., *that clump of birds is moving pretty fast*. The new exception is that this concept can apply to any group of anything *except data*. *The data shows a steep slope* is considered incorrect by some, who prefer *the data show a steep slope*.

If you are one of the people who think that *the data is* is wrong, please stop.

**Some examples** First, let us imagine a world where English grammar would require all groups to remain plural:

1. The agenda are on the table.
2. The trivia in this book are silly.
3. Steely Dan are playing at the pavillion.
4. The NIH owe me \$12,000.
5. The U.S.A. are in a recession.

Notes:

1. Agendum/agenda has the same Latin-based form as datum/data. Yet I have never heard a person who uses *the data are* use *the agenda are*.

2. Sentence #2 is the only one that is actually incorrect, due to the odd history of trivia. Here's the definition of *trivium* from the OED: in the Middle Ages, the lower division of the seven liberal arts, comprising grammar, rhetoric, and logic. That is, trivium was itself once a collective singular. The meaning evolved, and we can now group together a collective unit of facts about the trivium into bundles that are collectively a unit: trivia. In the present day, *trivia* is always a singular, because *trivium* refers not to individual facts but to the above fields of study. The singular of *trivia* is basically lost. [And since I know you're gnawing to know, the other part of the seven liberal arts is the quadrivium: "the four mathematical sciences, arithmetic, geometry, astronomy, and music"].

3. Bands and orchestras are a great example of the whole being more than its parts.

4. The acronym in number 4 expands to *National Institutes of Health*, and they do continue to "lose" my invoices as quickly as I can send them. Acronyms are a great way to cohere a plural into a singular.

5. The 360,000 casualties mentioned above come from #5: the question of whether *the U.S.A. are* or *the U.S.A. is* is the difference between a Confederacy and a Federation, and was basically resolved by a civil war. People fought and died over the question of whether a set of elements should be taken as separate elements or a unit, just a box of parts or a coherent whole.

More mundane examples still reveal different points of view. Both *the flock of birds are flying* and *the flock of birds is flying* are correct, but one or the other probably sounds off to you. Maybe you flinched when I wrote *agendum/agenda has* at bullet point one above. Here, grammar is a window to the soul. I think that some people generally lean toward seeing the parts and some generally lean toward seeing the whole. [Linguist readers are welcome to leave citations regarding my claim in the comments.] In one case this difference in thinking led to a war, but in most cases it seems to just lead to people correcting other folks' grammar when the grammar really just reflects a difference in perception.

[Oh, and *hair* is an interesting case: there's a form *your hairs* for a set of items that is not to be taken as a whole, and *your hair* referring to the whole mop on your head. It'd be great if we'd evolved more pairs like that, like maybe *datums* and *data*.]

**The math section** Let's get back to *data*, which is in the mathematical realm. Precision matters in math, and grammar needs to follow along. The sentence *that set of numbers is prime* is incoherent: only the individual numbers can be prime; a set can't be prime. The sentence *that set of numbers are dense*<sup>1</sup> is incoherent: only the set as a whole can be dense; individual numbers are not dense. We need both *the set is* and *the set are* in our grammar.

Similarly with data: sometimes we are looking at the gestalt, such as statistic like the estimates of a regression parameter; sometimes we are looking at the individual elements, such as when we point out that all the numbers are positive. *The data are a matrix* is incoherent: on the left-hand side of the *are*, we refer to a plural, while on the right-hand side, we're stating a singular; the sentence reduces to *a plural = a singular*. It's a perfect demonstration that the left-hand side is meant to be taken as a collective singular, as expressed perfectly by *the data is a matrix*.

Efforts have been made to base the entirety of mathematics on sets of objects; in a world where collections are central, we desperately need both *the set of items is* and *the set of items are* to function; *the data is/the data are* is just a synonym.

**Why the new exception?** [Disclaimer to Ms. LDWH of Princeton, PA: the following paragraph does not apply to you. I know you're just following the darn style guide.]

So why are *the agenda is* and *the set of elements is* OK, while *the data is* is now considered to be wrong? I can't put this politely, but I get the vibe that the people who correct *the data is* are just trying to indicate smartness—and failing. The process is perfect for the person working too hard at smart: (1) Identify trivia: data is actually a plural, and has a Latin-sounding singular. (2) Payoff: feel smarter for knowing trivia. (3) Find somebody who seems to not seem to know your fact. (4) Big payoff: correct them!

---

<sup>1</sup>*Dense*: between any two elements of a set, there is another element of a set. E.g., between the real number 1.1 and the real number 1.2, there is 1.15.

[Another of my pet peeves, which I've mentioned before, fits the same form: the use of *methodology* (the study of methods) as a synonym for *method*. Look at me! I used a five-syllable word! I think it's a synonym for a two syllable word, but I chose to use the longer word anyway!]

But, as above, there are times when *data* is a pile of parts, and times when it has meaning only as a whole. In all sorts of situations, our brains are wired to sometimes see the parts and sometimes the whole, and there's no point starting wars with people who see things differently.