**Editor Loeske Kruuk**:

We can see the potential relevance of the topic for a biological journal, but it needs to be presented much more clearly. For example, your cover letter notes that the paper is "of broad appeal, bridging between linguistics, cognitive psychology, and semiotics", without mentioning biology - so I wonder if you are aiming for a different audience? I am therefore not clear as to whether it fits with Proc B's remit of addressing fundamental biological principles.

**Associate Editor Katie Slocombe**:

My main concern is whether this article addresses a fundamental biological principle, and is therefore a good fit for the journal. R2 shares some concerns about the suitability of the paper for the readership of the journal, so in your revisions you need to consider how you can make the relevance of your findings to understanding fundamental biological principles clear.  
  
**Referee 1:**  
*1. In the introduction the authors write (80-81): does the imitation of a particular water-splashing sound become, over generations of repeated  imitation, a better label for the more general category of water-splashing sounds? What does this mean in linguistic terms? Do the authors refer in any way to the iconicity of language?*

*2. In “Collecting vocal imitations”, 111-112, “Participants were allowed to listen  
to each target sound multiple times” How many times were allowed to listen to each target? Was there a maximum number of repetitions allowed to each participant?*

There was no maximum number of repetitions allowed to each participant. The quoted text now reads:

Participants were allowed to listen to each target sound as many times as needed, but were only allowed a single recording in response.

*3.  Figures in the text, please, the graphs must include the headings on the x and y axes*

We’re sorry for this. There was an error in the conversion of our document into the online manuscript management system. A support request was submitted during our initial submission but was not addressed before the manuscript was sent out. Our submission now appears correctly online as it does offline.  
  
**Referee 2:**  
*1.My first general issue concerns what these findings tell us. The paper shows that modern English-speaking humans show biases, that are present in their language, in the specific task. Namely, they can imitate sounds (and we have plenty examples of onomatopoaies in English as well across languages), and their imitations become more word-like (and there are plenty indications from previous studies using iterated learning - see point 2 below) that through generations language-like properties emerge in this paradigm. Thus, the findings are not surprising.*  
  
*Still, unsurprising findings can be informative about language evolution. Here, however, I also have a problem. In addition to the usual difficulty in making any sort of claim concerning language origin and evolution starting from modern humans (who already know a language), the present findings do not allow us to claim that imitation in vocalizations is a sine qua non, nor that they have been the first step. They do not falsify the possibility that gesture came first, nor that a combination of gesture and vocalization was key. They simply show that imitation in vocalization might have played a part. Thus, I find the argument proposed in the paper not to be fully justified on the basis of the findings.*

We take the Reviewer’s point that our hypothesis about the formation of conventional words from vocal imitations is not new (although we do believe we are the first to attempt to document the transition from vocal imitations to conventional words). We also agree that these findings do not directly falsify other hypotheses about language evolution, namely the role of gesture. In light of these points, we have qualified our main argument not to be the final story about how language evolved, but a specific test of the minimal conditions under which vocal imitations might give rise to conventional words. In the introduction, we now state our research question as follows:

Thus, converging evidence suggests that people can use vocal imitation as an effective means of communication. At the same time, vocal imitations are not words. If vocal imitation played a role in the origin of some spoken words, then it is necessary to identify the minimal conditions under which vocal imitations can give rise to words that can be integrated into the vocabulary of a language. In this research we ask whether the intention to communicate is necessary for establishing linguistic convention, or whether vocal imitations might transition to more word-like forms through sheer repetition --- without an explicit communicative goal. To answer this question, we recruited participants to play an online version of the children's game of "Telephone". …

We also included a paragraph discussing the implication of our results for theories of language evolution that involve gesture.

*2. There is by now an important tradition within language studies in using the method presented here in Experiment 1 (and multiple variations on this methods). This method is referred as iterated learning has been introduced by Simon Kirby and colleagues (Kirby et al., 2008). It is rather puzzling that the present paper does not make any reference to the previous papers, to the name of the paradigm and, importantly, to the learning mechanisms that have abundantly discussed in the literature as underscoring the processes simulated with the use of this paradigm. This, in my mind, is especially important. On line 419, they write “Our results show that through simple repetition...”, which seems to me to trivialise what is going on in the study. In fact, they don’t really address what is going on – what the learning mechanisms may be that give rise to rendering imitations more word-like through generations. It seems strange to me to use a very particular and well-discussed experimental paradigm and then not discuss at all why you used it or what the effects of it are.*

which is a different question then whether linguistic structure can emerge from unstructured units.

*3. The authors call ‘imitative’ what, in the literature is more commonly referred as ‘iconic”. They do not define what they mean by ‘imitative’ nor whether ‘imitative’ is still a correct label to use for the word-like productions that they observe in later generations (where the vocalizations are indeed less imitative). This is an important point if we want to understand the mechanisms that underscore the effects found.*  
  
*4. I am unsure about the reasoning behind experiment 3. They argue that if vocalizations are more word-like, then they should more likely refer to categories rather than specific exemplars. I am always uneasy when talking about categorization levels as we do not know, really, whether the specific sounds within each category might already be categories standing for various tokens of the same type. That having been said, I know that some people have argued that iconicity is at the level of categories, not exemplars (e.g., Monagham, 2012). However, we also know that speakers may have more or fewer specific labels within a category (e.g., Eskimos have 10 different words for snow, English speakers only have one) depending upon whether it is important to differentiate or not. Hence one can argue that there is a bias toward category-level labels just because making finer distinctions is not so relevant to the task.*  
  
*5. The different results for spoken and written labels that they got in Experiment 3 may be task related: in the spoken version, there were three questions, only 2 in the written version. This alternative account needs to be addressed.*  
  
*6. Related to 5, the issue of differences in difficulty between questions in Experiment 3, I think, is not sufficiently addressed.*