Thank you for the opportunity to address Reviewer 2’s remaining concerns with our manuscript. Here are the specific changes we made. We focus in this letter on the specific issues raised in your letter, but we are happy to provide more detailed responses to Reviewer 2, if needed.

*1. Acknowledge that imitation alone is not sufficient for sounds to be used as words - other cognitive and intentional processes are also required for this.*

We have added a *Limitations* section in the General Discussion to acknowledge this point. The text now reads:

Although our results show that repeated imitations lead to increases in stability of spoken (as well as transcribed) forms, we recognize that there are additional requirements for the vocalizations to be incorporated into a linguistic system. One of these may be familiarity with the referents that are being imitated. The extent to which our results depend on prior familiarity with the referents can be measured by extending our procedure to less familiar referential domains. Another design limitation is the use of auditory referents that can be imitated (environmental sounds). But although vocal imitation may seem to be restricted to auditory referents, prior results indicate that people show considerable agreement on how to vocally “imitate” non-auditory and even somewhat abstract meanings [54,33].

*2. You may wish to consider revising the title so you are not claiming that these vocalisations are words, but rather more word-like (e.g. Imitation makes human vocalisations more word-like)*

We updated the title of the manuscript following your suggestion. We included “Repeated” to emphasize the iterative nature of our paradigm.

*3. You could discuss in the general discussion that imitation in modern humans is shaped by the biases and constraints of a modern, language competent brain and thus, prior to language evolving, imitation may not have had the same effects as you found here (so be more tentative about how clearly these findings relate to early word emergence in our ancestors).*

This is a fair point and applies to nearly all experimental investigations of language evolution. We have addressed this concern in the *Limitations* section of the General Discussion:

In addition, the present studies—like nearly all experimental investigations of the evolution of language—are limited in their inferential power by the use of participants who already speak at least one language. It may turn out that the ability to repeat vocal imitations and converge on more word-like forms only arises in people who already know and use a full linguistic system, which would limit the relevance of our findings for the origins of spoken words.

*4. You could clarify the number of iterations in your chains needed to see the stabilisation of sounds.*

We included the following paragraph in the results of Experiment 1 to clarify the length of the chains and the role that chain length played in our analyses:

Although in some chains, imitations were repeated up to 8 times, increasing similarity between generations *n* and *n+1* could be detected after fewer repetitions, in 5 generations or fewer. Imitations from chains that did not reach 5 generations due to experimental constraints (see Fig. 1) were included in all analyses, which included appropriate random effects to assure that these shorter chains were not treated equally to longer chains. However, chains with fewer than 5 generations were excluded from analyses involving transcriptions of the first and last imitation in each chain because these analyses collapse across generation.

Reviewer(s)' Comments to Author:  
  
Referee: 1  
  
Comments to the Author(s).  
Edminston et al, present an interesting and very innovative work aimed at studying the role of vocal imitation in the evolution of language and in the generation of vocabulary.  In this resubmitted manuscript the authors have  appropriately discussed the issues I raised in the first review.  
  
Referee: 2  
  
Comments to the Author(s).  
Review of “The emergence of words from vocal imitations” by Edmiston et al  
  
The manuscript is a resubmission of a previous version I reviewed. I think the authors did a good job in responding to many of the concerns myself and the other reviewers raised to the previous version (e.g., comparing the methods they used in Exp 1 with iterated learning, explain better the methodology used in Exp 3). The current version is very well written and clearer than the previous one.  
  
Perhaps it is because this version makes the basic assumptions and claims clearer, I am afraid that I still have important reservations concerning the paper. In particular:  
  
1.      The assumption that imitation is the only key to the origin of language. It is argued that because many signs are transparently linked to their referents (they are iconic), thus their origin must be rooted in imitation. The argument then continues that, if this is the case for sign languages, it could also be the case for vocal imitation. I do question the plausibility of the argument. My main point is that the fact that imitation is possible and can be seen in some modern signs (as well as in onomatopoeias) is not sufficient to argue that it was the key to language origin. This neglect many other aspects that have been discussed, perhaps to me most important is that in order to have a communicative system, the signs (being imitative or not) have to stand for their referents. In other words, there must be a mental representation of an object in the environment that can be evoked by an imitative/iconic form in order for such form to be a word (see discussions in Bickerton D. 2009 Adam’s tongue: how humans made language, how language made humans. New York, NY: Hill & Wang.  
Kendon A. 1991 Some considerations for a theory of language origins. Man 26, 199 – 221. (doi:10.2307/ 2803829). Imitation can help evoking the referent but there needs to be a mental representation to link the imitation to the referent in the world.  Note in this respect that a fundamental difference between non-human calls and human language is what Bickerton (2009, see also Perniss and Vigliocco, 2014) call functional vs conceptual reference, not (or not only) in the mechanisms of imitation.

This is a good point. When we were designing the experiment, we considered adding pictures of objects to present to the participants during the imitation procedure to help them form the correct associations. We decided it was better to start with a baseline condition, with no referent objects, just to gauge the fidelity of vocal imitation, without even the simple goal of learning the association between a sound and a referent. We were surprised to see the vocalizations converge rather rapidly on novel word forms, even without a referent object. In retrospect, we believe that by using everyday environmental sounds that participants were familiar with allowed them to invoke enough of a mental representation of the likely sound source to complete an iconic referent. We would predict that adding context would only increase the rate of convergence, although there are a number of different ways this could be tested. Our argument is that vocal imitation is an independent contributor to some linguistic convention, but not that vocal imitation is the sole contributor to language evolution.

2.      The expectation that in Experiment 1 vocalizations become more word-like (vocalizations are also referred as words in the text – e.g., title of paper -- which I think is misleading). That is, when modern subjects are asked to imitate through generations environmental sounds, these imitations become less similar to the original seed sound but begin to incorporate features of human language such as they are easier to copy, they are easier to write and they tend to become more categorical. My main problem here is that there is a straightforward alternative account for this. Imitations reflect biological constraints on what is easier to hear and to articulate as well as cognitive biases related to the fact that modern humans have language and, arguably, speech is the most common type of vocalization we produce and hear. Through generations, the biological and cognitive biases become more visible. However, no such biases and constraints might have been at work during language evolution as the biological and cognitive constraints could be the result of evolution, not the mechanism underscoring it.

There are two ways to falsify our claims, the first would be to show that convergence and stabilization only occurs when imitators speak the same language, which can be tested by comparing the rates of convergence among imitators who do not speak a common language. The second would be to demonstrate that people without any language, engaging in repeated imitation, would be unable to develop more conventional forms. Although we are of course unable to do the ideal experiment, we do acknowledge how our claims may need to be updated given further evidence.

3.      A more minor issue relates to how was the number of generations (max 8) decided? I understand that the vocalization at the end of the chains were more similar to one another, but I would like to see that vocalizations stabilize and that this is used to decide how many generations to have.  Related, roughly half of the chains do not have more than 4-5. How is this taken into account? How many generations are necessary to see the biases in action?

The max of 8 generations (more would have been better) was decided largely based on limitations of experimental resources. With 16 seed sounds and up to four chains off of each seed sound, 8 generations of repetition amounted to 512 imitations. Unfortunately, mostly due to issues having to do with running an experiment online, some of these had to be discarded. This is the primary reason the chains are not all equal length. If we would have designed the experiment differently, we could have rerouted future participants to complete each chain. But we chose instead to continue with the original method which involved random assignment and account for the length of the chains in the statistical models wherever appropriate. For the hierarchical regression models, we include random slopes and intercepts for each chain, meaning the model estimates take into account the length of the particular chain.  
  
In sum, I think the work is rigorous and well presented. However, I do not think it really advances our understanding of  language origin. It may be that a revision that more clearly articulates the assumptions and how these are necessary to explain the emergence of language could be sufficient. I am doubtful, however, that the methods used are the best way to address the question.