Reviewer #1:

The authors have done much to respond to the first set of reviews, most of it quite effective and welcome. They also continue to state a conclusion (not so forcefully in the revision as in the previous version) that the findings in their Study 2 support a language-as-causal theory of observed correlations between use of gender-distinct nouns for occupations and implicit gender stereotypes. I looked for the reasoning that might support this causal conclusion, but I could not find it. What I did find were assertions that the language-as-causal theory had been confirmed—not the underlying reasoning to support those assertions.

This is what I found (underlines explained below): "If language is causally related to implicit associations, then differences in the structural aspects of language that act to exaggerate linguistic gender association should predict greater implicit association. This relationship is not predicted by the language-as-reflection hypothesis." (lines 325-329). "the positive association between prevalence of gender-specific terms and implicit association (Study 2) is most parsimoniously explained by language statistics (partially) causing the observed differences in implicit association. It is implausible that non-linguistically acquired gender associations could have changed the lexical inventory of the language rapidly enough to explain the differences in IAT performance that we observed." (lines 464–470)

The authors did not state (and I could not guess) the logic suggested by the "if ... then" structure of the first sentence above, "not predicted" in the second, "most parsimoniously" in the third, and (perhaps most important) "implausible" in the fourth.

A reverse causal hypothesis seems more plausible than the language-as-causal hypothesis. The reverse hypothesis is suggested to me by multiple communications I have received advising me to change my use of gender pronouns to align with society's changing understanding of gender, which is shifting toward regarding gender as non-binary. Relatedly, I was aware in preceding years of organized efforts to replace use of gender-differentiated occupation terms with single terms (e.g., actor, waitperson). These observations (which I am sure are not unique to me) tell me that societally changed (explicit) gender cognitions have caused language change in fairly short time periods (this is the reverse causal direction). Conceivably, both causal directions can operate. Can the authors point to similarly familiar evidence showing that recently changed language structures have caused changes in (implicit) gender attitudes or stereotypes? I can believe that language changes, including ones resulting from shifts in society's cognitive understanding of gender, might eventually influence implicit gender cognitions. But I haven't yet seen the evidence for that. There is evidence that implicit attitudes toward sexual orientation have changed in the last decade or so (see the article by Charlesworth & Banaji, Psychological Science, 2019, 30(2), 174–192). Were these changes preceded by changes in language structures that reference sexual orientation? I find it more plausible that prior societal changes in explicit attitudes preceded changes

in media presentations toward more approving treatments of same-sex relationships, which in turn might be responsible for the shifts in implicit attitudes. My suggestion, then, is that explicit cognition (perhaps influence by language via interpersonal persuasion—not the same as change in associative language structure) can cause usage changes, a causal sequence that now seems more evident than are effects of changes in associative structure of language on implicit attitudes or stereotypes.

Bottom line: The authors' revisions are mostly fine, but I sincerely believe they would be wiser to limit their present advocacy of the language-as-causal theory to their observations about the types of (not-yet-obtained) findings that could support that theory.

Thank you for this feedback. We agree that our data do not provide strong evidence for the language-as-causal hypothesis. We have broadly revised the introduction and discussion to more clearly make this point. For example, we now state in the introduction:

"The correlational approach of the present work does not allow us to distinguish between these possibilities; our goal is to establish whether there is in fact a correspondence between psychological and linguistic gender associations. Establishing whether such a correspondence exists is a prerequisite to understanding the underlying causal model."

In addition, we have revised the text cited above to clarify that a causal interpretation is not warranted. Specifically, we have removed lines 464–470 cited above in the discussion, and clarified the logic behind lines 325-329 above with the sentence below:

"This relationship is difficult to explain if language merely reflects cultural stereotypes, since structural aspects of language are relatively fixed."

In other words, finding that structural aspects of language correspond to psychological stereotypes would provide some evidence that language is not merely a reflection of psychological stereotypes, since structural aspects of language do not change very quickly. Nevertheless, given that we only find this correspondence at the level of words (i.e. gender distinctions for occupation terms) and not grammatical gender, we acknowledge that this is only weak evidence that language plays a causal role in shaping stereotypes. The text no longer frames this finding in causal terms.

We also agree that it is certainly possible that both causal pathways are at play (language influences stereotypes; stereotypes influence language). In the previous revision we added a discussion of the kinds of additional evidence that could be used to more directly test the causal hypothesis, including the kind of data used by Charlesworth and Banaji (2019). Thank you for pointing us to this work -- we now cite this paper.

Reviewer #2:

Remarks to the Author:
I served as one of the previous reviewers, and I continue to be impressed by the scale and novel findings of this research. Many aspects of the measures and claims have been clarified in the revision; yet some concerns linger.

The claims for causality continue to be overstated, although I appreciate that this version includes calls for various designs that would build a stronger case for the causal relationship. It is still the case that a different third variable could be the root cause of both gendered grammatical structure and the implicit associations. The framing of the question can thus be more precise. The introduction poses questions that correlational methods cannot answer. For example, the authors review experimental work showing causal effects of language on beliefs. Then, they could pose their question as examining whether we see parallel effects in widespread, transnational data use. That in and of itself is important. But instead the authors describe the previous experimental work and their current research in this way: "Such work shows that in certain experimental settings, language can influence stereotype formation. We were interested in whether it actually does, and by what means." But these are not questions that can be answered by correlational methods because that hidden third variable might be causing both of the variables measured here. A society that strongly demarcates male and female roles varies in a host of ways – communicates and supports that role structure in a host of ways – that exist alongside or can contribute to gendered language.

Thank you for this point. We have broadly revised the manuscript to ensure that we do not make causal claims about these data. Among other changes, we have revised the sentence quoted above in the introduction to reflect the reviewers suggested framing:

"Such work shows that in certain experimental settings, language can influence stereotype formation. We ask whether a similar correspondence between language associations and stereotypes exists in a large corpus of naturalistic text and among an international sample of participants."

I want to be clear that I do not think these authors need causal evidence for these data to be worthy of publication. But I do think it is necessary for the manuscript to be clear in what it can and cannot show. It can rule out some causal relationships (if no association exists) but it cannot provide positive proof of causality. The results presented here are fascinating regardless, but the authors need to own what the data can and cannot say in their framing and discussion of their findings.

Thank you. We agree. We have revised the introduction to explicitly make this point, by including the following paragraph:

“Discovering that gender associations in language are correlated with people’s implicit and explicit gender associations can be interpreted in several ways (22, 25). One possibility is that some cultures have stronger gender stereotypes and these are reflected in what people talk about. Language, on this view, simply reflects pre-existing associations. However, language may not only reflect pre-existing stereotypes, but may also provide a distinct source of information for learning about them, thereby constituting a causal influence on the associations people learn (26). Another possibility is that a third variable influences both language and psychological associations. The correlational approach of the present work does not allow us to distinguish between these possibilities; our goal is to establish whether there is in fact a correspondence between psychological and linguistic gender associations. Establishing whether such a correspondence exists is a prerequisite to understanding the underlying causal model.”

> *My prior review noted reasons to use “association” rather than “bias,” and the response letter claims that this change is made. Yet “bias” continues to be used throughout the manuscript, and I think it brings baggage that is unnecessary. I understand different disciplines use “bias” to mean “tendency.” But in the intergroup literatures, “bias” denotes prejudice, which is not something that these associations can show. I continue to recommend using “association” rather than “bias” in describing the findings of the Implicit Association Test.*

We appreciate this feedback and have revised the manuscript to avoid the term “bias”.

> *I find this point fascinating and an excellent example that dissociates the propositional information of language from the structure:*
> *For example, a Finnish document about nursing being unsuitable for men would still use a gender-neutral form of “nurse” while a Spanish document promoting nursing careers to men would be committed to using gender-marked forms.*
> *Can any data in hand speak to this specific point – to effects of structure above and beyond propositional content?*

Excellent question. We have a pending grant to investigate this idea in more depth. One systematic way of dissociating propositional content from the structure is to augment the training corpora. For example, we can augment the Spanish training corpus by replacing “enfermera” (nurse: fem) and “enfermero” (nurse: masc) with a (non-existent) neutral form (e.g., “enfermerx”), thereby simulating what the gender association would be if Spanish did not distinguish between male and female versions of “nurse”. Importantly, this augmentation does not change the propositional content at all. In the analysis presented in Study 2, nurse is strongly female biased for Spanish (gender bias = .12). This disparity likely reflects the predominance of women in the occupation (e.g., in the untranslated corpus of spanish Wikipedia text, the feminine form of nurse (“enfermera”) is used 78% of the time and the masculine (“enfermero”) 22% of the time). The key prediction the proposed analysis would test is whether the gender bias for nurse is lower when the model is trained on the augmented corpus with only gender-neural forms for nurse, compared to the unaugmented corpus.

As noted earlier, I appreciate that the authors They note the need for different forms of causal studies in the discussion. However, they focus on experimental manipulations of "language statistics," whereas a much broader point can be made. The core claim here is that immersion in more or less gendered languages causes gender-stereotypic beliefs. The implication is that even momentary immersion could have these effects; for example, bilingual speakers might show larger gendered associations after having conversed in a more gendered language than a less gendered language.

The authors note less concern about issues of reliability in the IAT given the focus on group-level differences. I strongly suggest citing the Payne, Vuletich, & Lundberg (2017) review that present the case that the IAT is stronger in terms of reliability and validity when aggregated to the group level to reflect context than when used at the individual level. They provide a strong case for this point.

Thank you for pointing us to this very relevant paper. We now cite it in the Discussion. We've also added a few sentences in the discussion that highlight the need to understand the time-course of these effects: if there is indeed a causal link, how much exposure to the relevant language statistics is needed to influence people's gender-stereotypes and what might be the effect of exposure to multiple languages.