We are very grateful to the three reviewers for their careful reading of our paper, and for their insightful comments and suggestions. Our responses are below with the reviewers' comments in italics.

- **Reviewer 1** speaks very positively about our paper tackling an *"important topic under a new perspective"*, and we too anticipate that it will *"open up space and motivate new research"* in this area. In response to specific points:
  - *"The paper needs more details on the experiments..."*Although we do give full details (of embedding models, training data, embedding dimension, etc) in the Figure captions we appreciate that these might have gotten somewhat lost; we will revise to clarify, and expand explanation in the main text.
  - *"Give a clearer motivation for how embeddings were constrained in §4.1...the authors do not show why this reduced set of solutions would be better than a random choice in the full set [nor give results for it, which] could be presented in Table 1...Why not expand Table 1 to [include LSA, Word2Vec]".* As pointed out in Remark 3, the constraints in §4.1 are related to the widely used Gram-Schmidt process, and are natural under the group-theoretic formalism. We make no claim that this particular solution subset is optimal (or even at all good). We will amend the text to make this clear, and also follow the excellent suggestion to add results corresponding to §4.1, as well as Word2Vec, to Table 1.
  - *"Do they optimize [Λ] on the same data D used later to get the scores?"* We did not use cross-validation for the results in Table 1, because the purpose of these results is to understand how much g can be increased without changing f, as we say in the caption to indicate the "substantial scope for improving performance scores via an appropriate choice of Λ." We will amend presentation to further emphasise this, and also expand our caution of overfitting (line 250) to discuss specifically the need for cross-validation when producing predictive models.
  - *"Fig 2 (a,b), the red lines do not match when $\alpha = 0$...".* Thank you for pointing this out (and conjecturing correctly the reason!). We will amend.
  - *"Fig 3, results for upper triangular...have better performance than diagonal ones...[Intuition?]...What are red lines...original V results?"* Our intuition here is that the highest performing instances do (slightly) better for the upper triangular case than the diagonal case since the extra degrees of freedom give greater scope for a solution with large g; and that the left tail is long because with |N(0,1)| random diagonal elements there is a substantial chance of small elements that essentially wipe out whole rows of V. We will add this discussion in the revision. The red lines are indeed the results for the original V^*, which we will now state in the caption.
  - *Various typos:* Thank you for pointing these out, all of which we will fix.
- We share **Reviewer 2**'s craving in answering the "meta question" of *why* embeddings work. Given the impressive success of numerous embedding methods being used and compared, we investigate a related, albeit modest, meta question 'for a fixed evaluation function, can we identify properties/issues of objective functions that drive the apparent disparity in performances of embedding methods?'. To our knowledge we are the first to do this.
  - *"Previous methodologies...implicity address [non-identifiability] already, as they place implicit constraints (like requiring U and V to be...symmetric)...[hence] 'so what?'"* It is not clear what implicit constraints the reviewer is referring to in general. But U and V are non-square so cannot be symmetric. Perhaps "requiring U and V to be identical on account of X being symmetric" was intended, as in §4.1.1. But this strategy is limited to situations with X symmetric. As we discuss in §4.1.1, the remedy suggested by the authors of GloVe is ad hoc (much less natural than the alternative we suggest) and leads to embeddings that are not even optimal with respect to the model's own objective!

    Our view is that to even begin to address the reviewer's "meta question" the first step is to understand embedding methods' objective functions and the properties of embeddings that result. We feel it is paramount to understand the impact of the identifiability issue first, instead of adding to the growing list of embedding methods by throwing out new objective functions.
  - Suggestions: (i) *"looking at a carefully selected larger space, [e.g. Artetxe et al 2018]".* The linear transform in their paper (in our notation $V = \Lambda^\alpha Q V^*$) is just identified with a one-dimensional subset of $GL(d)$, and since g is invariant to any $Q \in O(d)$ their approach is covered by our Proposition 1. (ii) *"modifying the training objective".* Addressed previously. (iii) *"embedding methodologies that are theoretically more sound and performs better..."* It is unclear what the reviewer means by 'theoretically more sound'. Perhaps it's related to the suggestion in (ii)?
- **Reviewer 3** makes some helpful suggestions about expanding §4.2, in particular to include *"NON-WORD-tasks (not just word similarity...using [inner product])".* We emphasise it is not just similarity but analogy tasks too for which g involves V only via the inner product of its columns, though we acknowledge the lack of analogy results, which some readers may find more interesting, in §4 and will include some in the revised version. One point of clarification: the reviewer notes that we focus on the *"rather special linear case [of] LSA"*, though we do so only in §2 as a particular example that aids exposition of the identifiability issue; the paper's key results in §3 and §4 are appropriate to any model of the very general form (1), including GloVe and word2vec.