Welcome

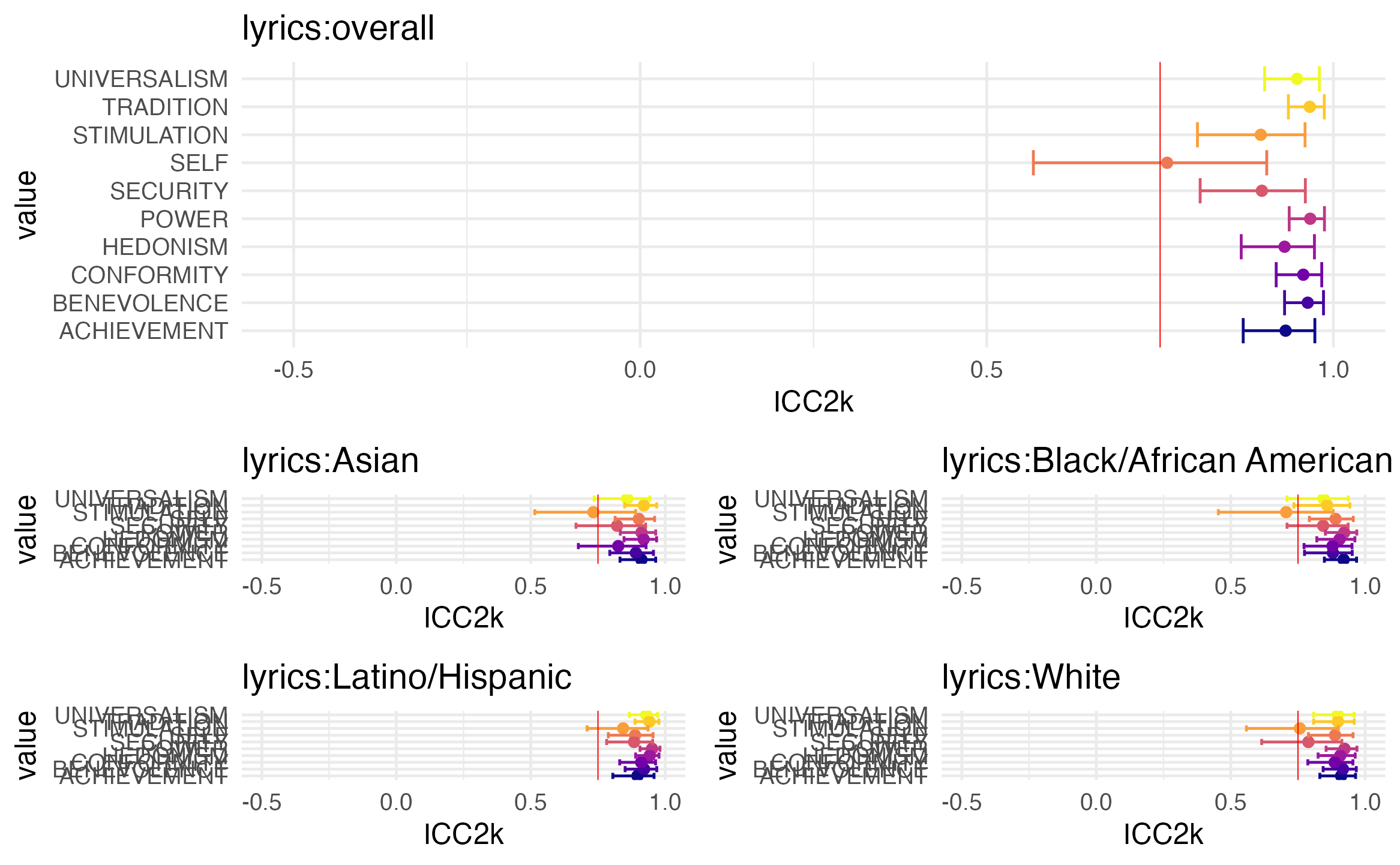
Table of contents

# 1. Results

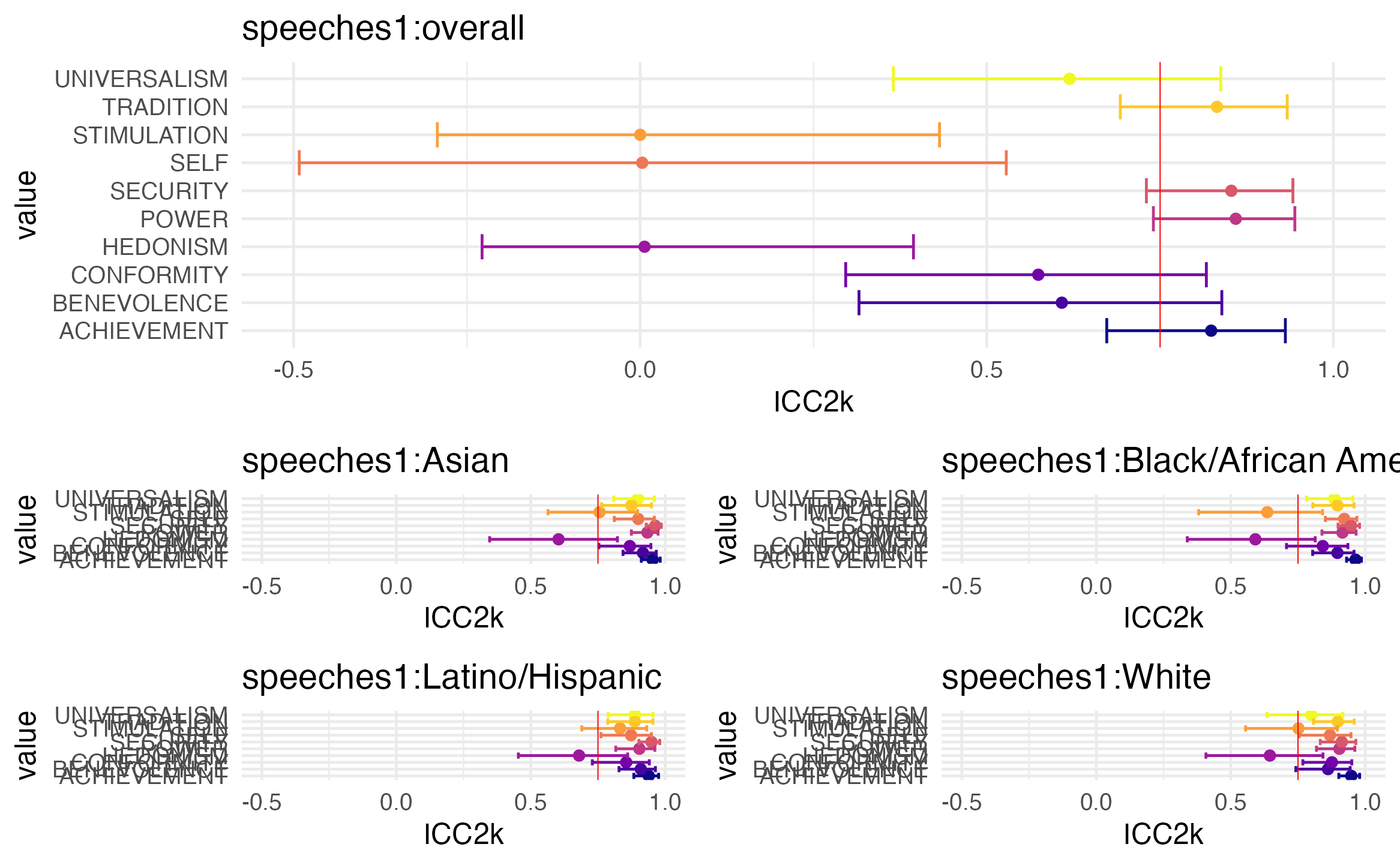
## 1.1 Pilot Study

### 1.1.1 Determining whether annotator sample must be stratified

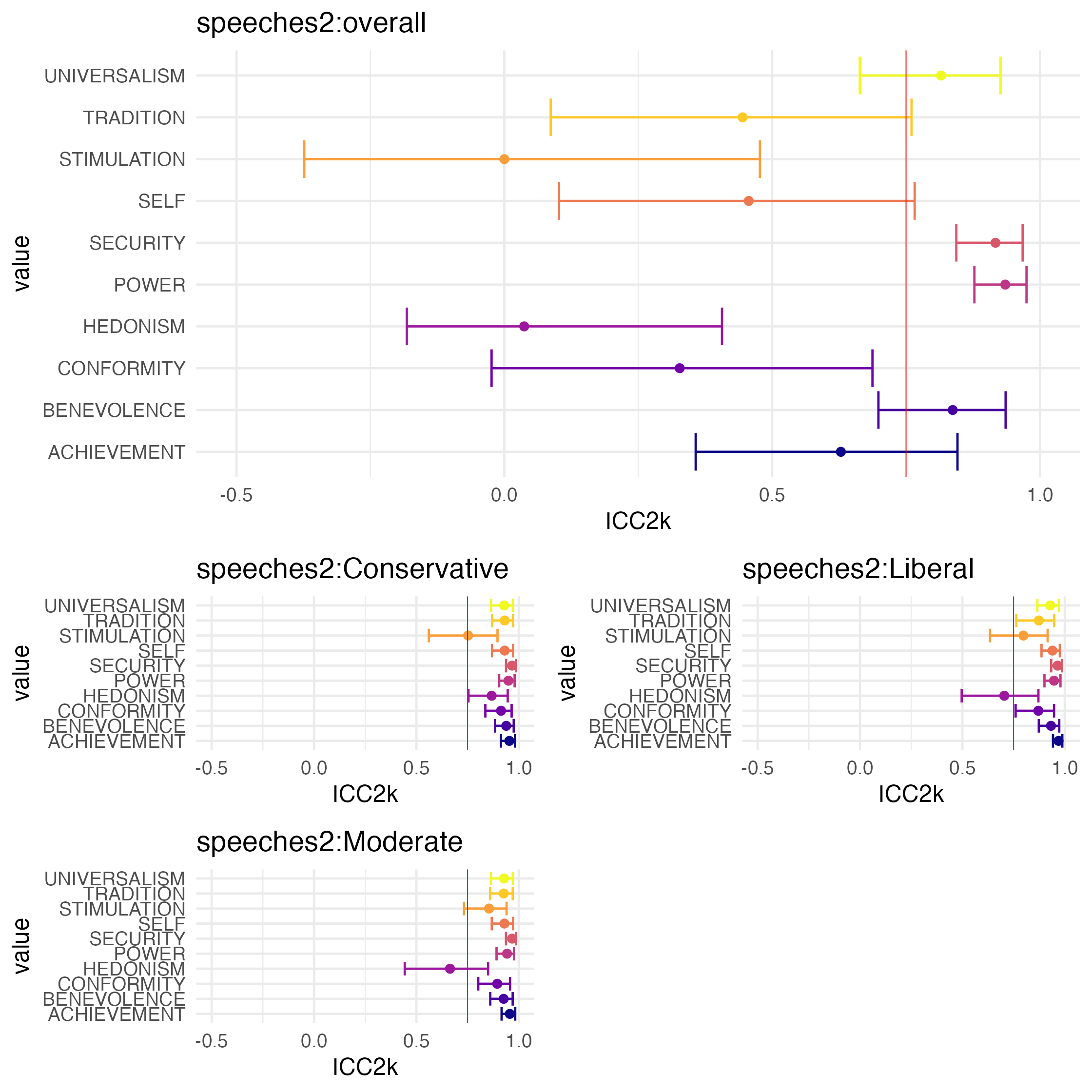
We estimate ICC2k overall, and for subgroups separately. For our lyrics pilot, we observe overall ‘good’ ICC2k estimates, with the exception of SELF-DIRECTION, which shows a large confidence interval that exceeds our threshold. These ICC2k scores overall do not appear to improve when we estimate them separately for each Ethnicity.



This is not the case for our two pilots which used political speech data. Overall, agreement on the speech data was unusably poor, however it appeared to improve when estimated separately for each ethnicity, with most values showing ICC2k scores over the threshold.



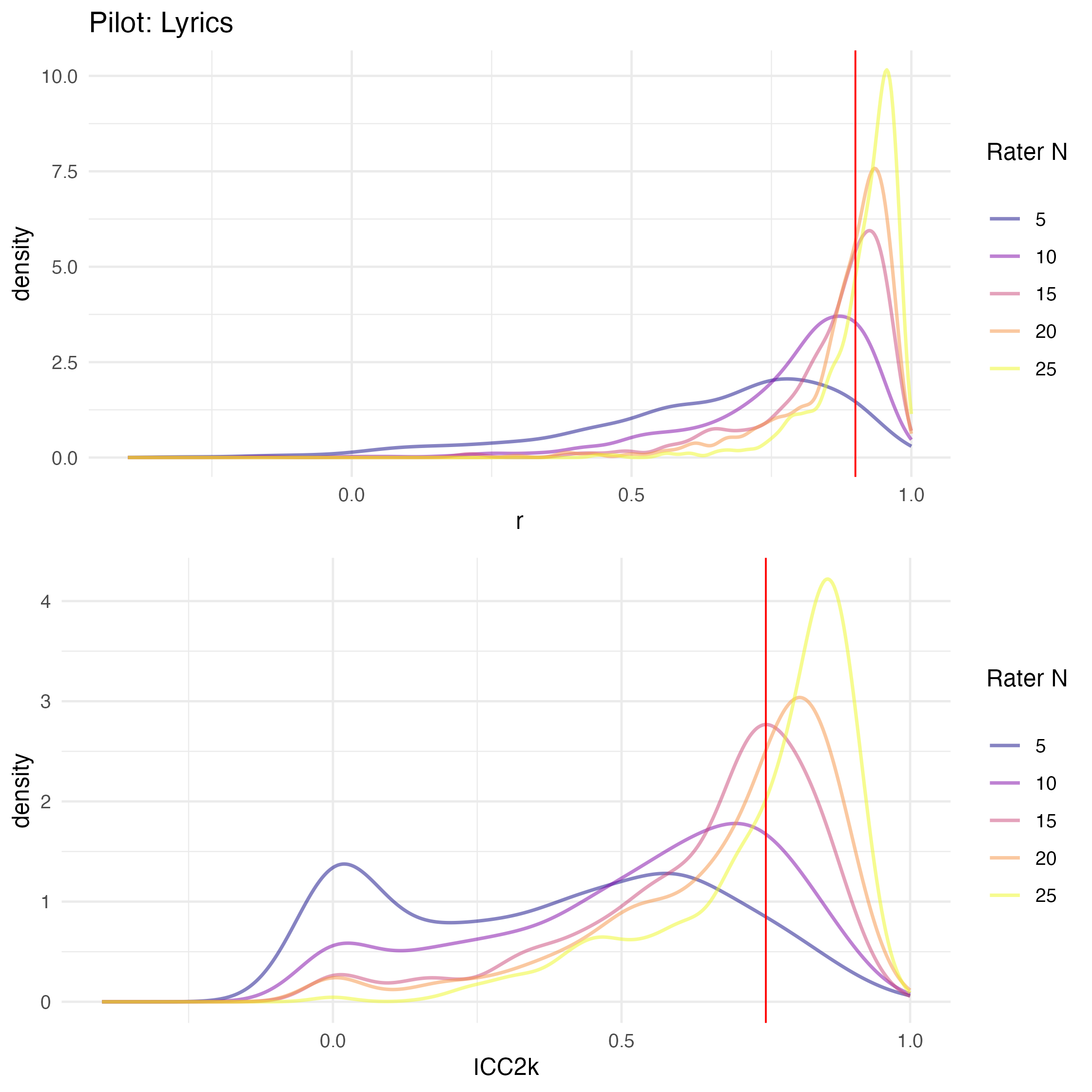
We observe similar results for our second speeches pilot, which show substantially higher ICC2k scores when they are estimated separately for each political spectrum category.



We tentatively conclude that Ethnicity will have a substantially smaller effect on ratings of song lyrics than either Ethnicity or Political Spectrum will have on ratings of speeches. Thus, the sample for our main study data collection will be stratified by Ethnicity, and for political speeches, it will also be stratified by political affiliation.

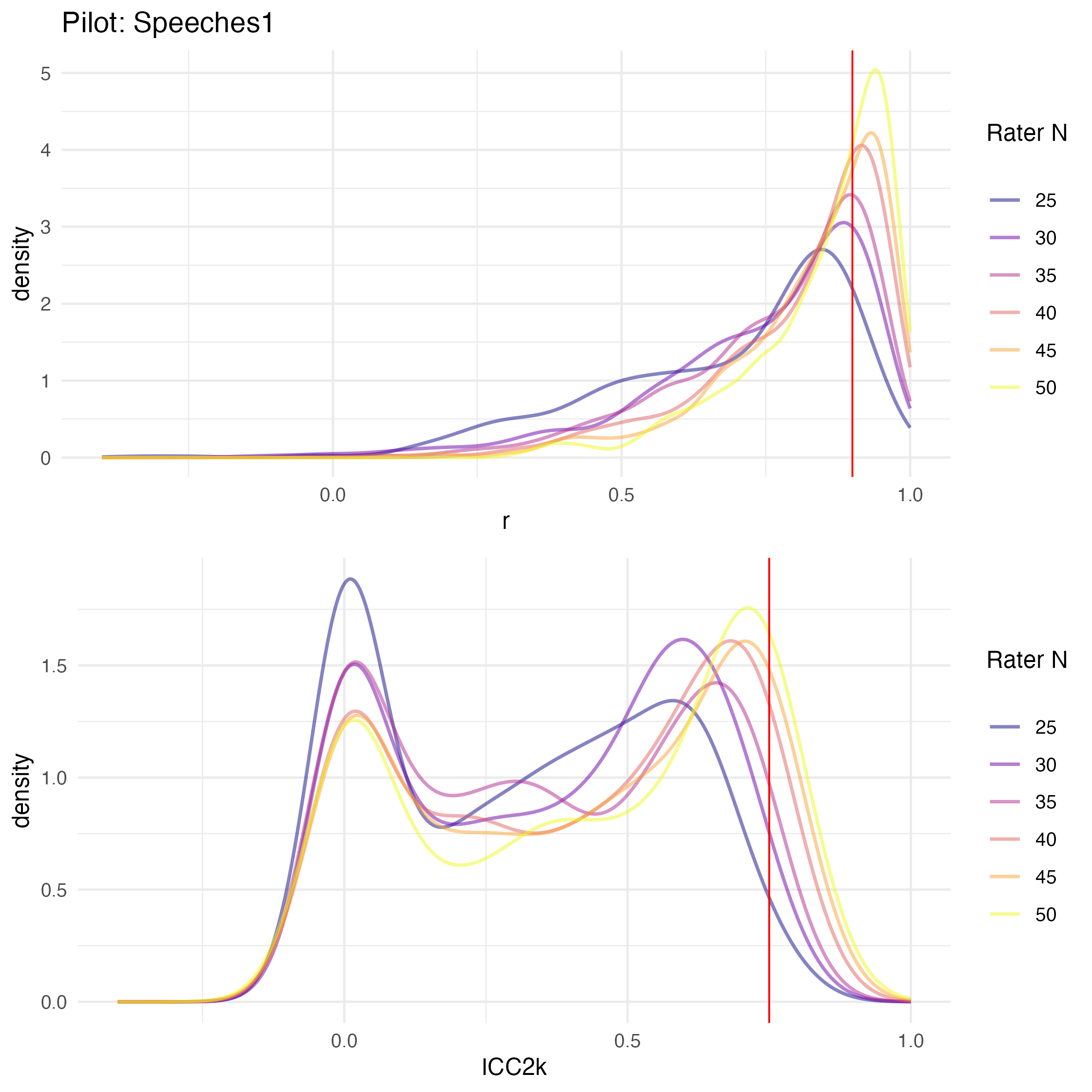
### 1.1.2 Estimating the number of raters

The results from our bootstrapping procedure show adequate ICC2k scores with 15 ratings for lyrics, in both ICC2k scores, and correlations between subsample means and the overall mean.

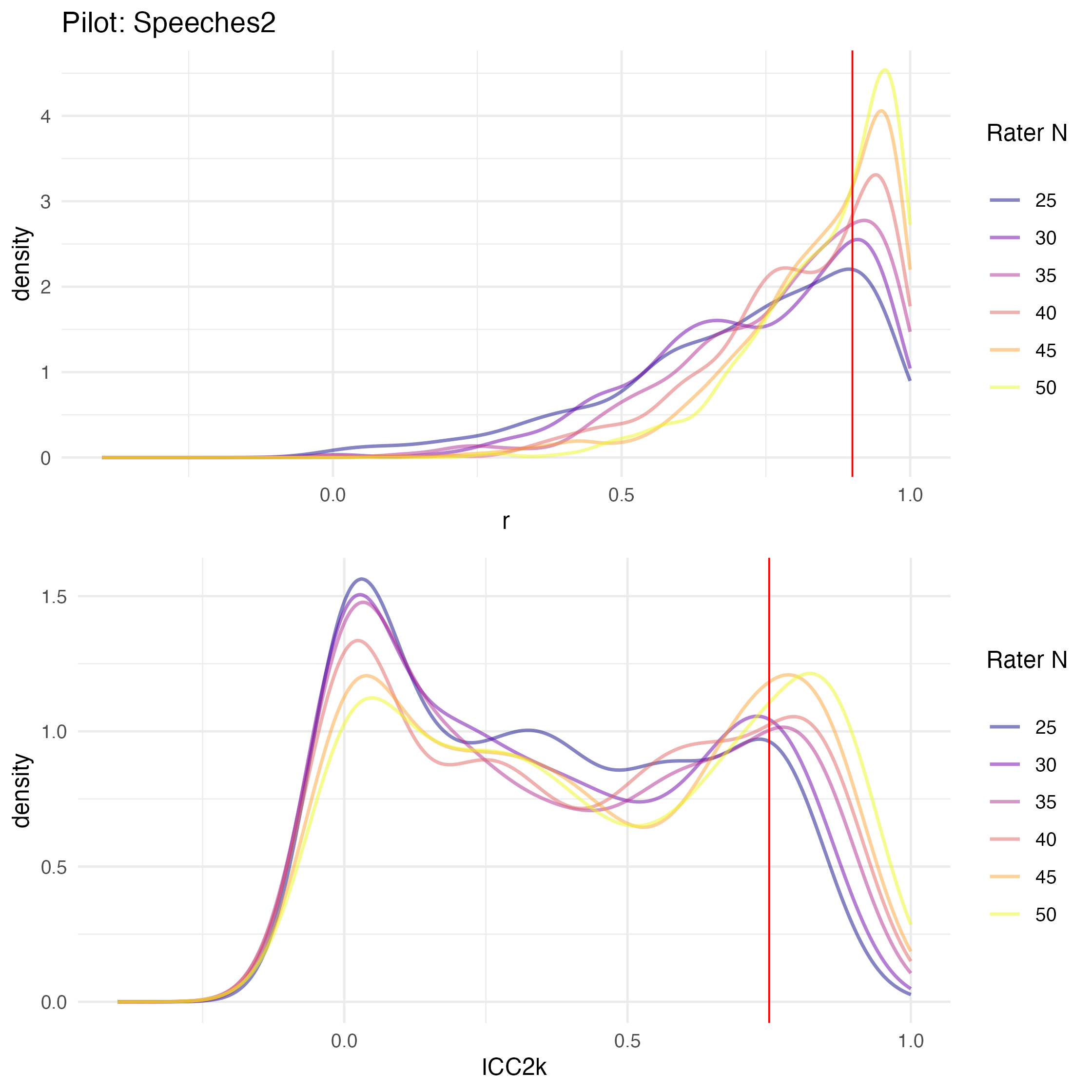


For both samples in our speech pilots, we observe a bimodal distribution in ICC2k scores, with peaks nearing our threshold, but also around 0.

Our sample of annotators in our first speeches pilot were stratified by Ethncitiy. Our bootstrapping analysis shows sufficiently high correlations between subsample means and the overall mean at around 40 ratings, however, ICC2k scores do not reach sufficiently high ICC2k scores even with 50 ratings.



Our sample of annotators in our first speeches pilot were stratified by political affiliation. We observe a similar pattern, where the correlations between subsample means and the overall mean reach our threshold 35 ratings, but ICC2k scores are not sufficiently high even with 50 ratings.



We thus proceed with a target median 15 ratings per song lyric, and median 40 ratings per political speech.